CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

REFINAMENTO DE MODELOS DE RESPOSTAS A PERGUNTAS VISUAIS BINÁRIAS

Ramon Ferreira Silva

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador: Eduardo Bezerra da Silva, D.Sc.
Coorientador: Joel André Ferreira dos Santos, D.Sc.

Rio de Janeiro,
Fevereiro 2019

REFINAMENTO DE MODELOS DE RESPOSTAS A PERGUNTAS VISUAIS BINÁRIAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Ramon Ferreira Silva

Banca Examinadora:

Presidente, Eduardo Bezerra da Silva, D.Sc. (Orientador)

Prof. Joel André Ferreira dos Santos, D.Sc. (Coorientador)

Prof. Kele Teixeira Belloze, D.Sc.

Prof. Ronaldo Ribeiro Goldschmidt, D.Sc. (IME)

Rio de Janeiro,

Fevereiro 2019

CEFET/RJ – Sistema de Bibliotecas / Biblioteca Central

S586 Silva, Ramon Ferreira
Refinamento de modelos de respostas a perguntas visuais binárias / Ramon Ferreira Silva.—2019.
78f. : il. (algumas color.) , tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2019.
Bibliografia : f. 71-78
Orientador : Eduardo Bezerra da Silva
Coorientador : Joel André Ferreira dos Santos

1. Processamento de linguagem natural (Computação). 2. Redes neurais. 3. Visão por computador. 4. Ciência da computação. I. Silva, Eduardo Bezerra da (Orient.). II. Santos, Joel André Ferreira dos (Coorient.). III. Título.

CDD 006.35

Elaborada pela bibliotecária Mariana Oliveira CRB-7/5929

DEDICATÓRIA

Dedico esta dissertação à minha família, e em especial, à minha esposa Vanessa pelo apoio incondicional e constante incentivo.

AGRADECIMENTOS

O presente trabalho foi desenvolvido com o apoio direto ou indireto da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Ao Professor Eduardo Bezerra da Silva (CEFET/RJ), pela orientação na minha pesquisa e pelos diálogos sempre frutíferos e incentivadores, também pela educação e boa vontade com que me recebe em sua sala.

Ao Professor Joel André Ferreira dos Santos (CEFET/RJ), pelo incentivo, paciência, confiança e orientação.

À Professora Kele Teixeira Belloze (CEFET/RJ), pelo incentivo à apresentação de meu trabalho e aulas de Metodologia de Pesquisa que tanto me ajudaram.

Ao Professor Ronaldo Goldschmidt (IME), pelo apoio e incentivo.

Aos Professores da Escola de Informática e Computação do CEFET/RJ Diego Barreto Haddad, Diego Nunes Brandão, Eduardo Soares Ogasawara , Gustavo Paiva Guedes, Jorge de Abreu Soares, Raphael Carlos Santos Machado e Pedro Henrique González Silva por compartilhar seus ensinamentos e dar apoio durante todo o mestrado.

Ao Augusto José Fonseca, aluno de Iniciação Científica da Escola de Informática e Computação do CEFET/RJ, pela ajuda de com o desenvolvimento de ferramenta web para curadoria e ajuda na pesquisa.

Aos alunos de Iniciação Científica e Mestrado da Escola de Informática e Computação do CEFET/RJ, Ivair Nobrega Luques, Marcello Serqueira, Gabriel Luz, Ricardo Sant'Ana, Bryan dos Santos, Mateus Pereira, Raphael Correia de Souza Fialho, Nathália Gomes, Rafaela de Castro do Nascimento, Luiz Miguel Viana Barbosa, Leandro Maia e Daniel Favoreto pelo apoio na realização desta pesquisa.

RESUMO

Refinamento de Modelos de Respostas a Perguntas Visuais Binárias

Respostas a Perguntas Visuais (*Visual Question Answering*, RPV) é uma tarefa que une os campos da Visão Computacional e do Processamento de Linguagem Natural (*Natural Language Processing*, PLN). Tomando como entrada uma imagem I e uma pergunta em linguagem natural Q acerca de I, um modelo para RPV deve ser capaz de produzir uma resposta R (também em linguagem natural) para Q de maneira coerente. Um tipo particular de pergunta visual é aquele no qual a pergunta é binária (i.e., uma pergunta cuja resposta pertence ao conjunto $\{$sim, não$\}$). Atualmente, redes neurais profundas representam o estado da arte para o treinamento de modelos de RPV. Apesar de seu sucesso, a aplicação de redes neurais à tarefa de RPV requer uma quantidade muito grande de dados para que se consiga produzir modelos com precisão adequada. Os conjuntos de treinamento atualmente utilizados para criar modelos de RPV são resultantes de processos laboriosos de rotulação manual (i.e., feita por seres humanos). Esse contexto torna relevante o estudo de abordagens automáticas ou semiautomáticas para aumentar esses conjuntos de treinamento durante o treinamento. Esta dissertação propõe duas abordagens para aumentar um dado conjunto de treinamento para RPV. Em particular, propomos uma ferramenta de curadoria para criação semiautomática de novos exemplos, além de um procedimento baseado em destilação de dados para criar exemplos de forma automática.

Palavras-chave: Respostas a Perguntas Visuais; Aumento de Dados, Destilação de Dados

ABSTRACT

Refinement of Models in Binary Visual Question Answering

Visual Question Answering (VQA) is a task that connect the fields of Computer Vision (CV) and Natural Language Processing (NLP). Taking as input an image I and a natural language question Q about I, a model for VQA must be able to produce a coherent answer R (also in natural language) to Q. A particular type of visual question is one in which the question is binary (i.e., a question whose answer belongs to the set $\{$yes, no$\}$). Currently, deep neural networks are the technique that corresponds to the state of the art for training of VQA models. Despite its success, the application of neural networks to the VQA task requires a very large amount of data in order to produce models with adequate precision. The datasets currently used for the training of VQA models are the result of laborious manual labeling processes (i.e., made by humans). This context makes relevant the study of automatic or semi-automatic approaches to increase these datasets during training. This dissertation proposes two approaches to increase a given training dataset for VQA. In particular, we propose (1) a curatorial tool for semi-automatic creation of new examples, and (2) a procedure based on data distillation to create training examples automatically.

Keywords: Visual Question Answering; Data Augmentation; Data Distilation

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AA	Aprendizado Ativo
AD	Aumento De Dados
AM	Aprendizado De Máquina
AP	Aprendizagem Profunda
AS	Aprendizado Supervisionado
DD	Destilação De Dados
FNN	Redes Neurais Artificiais Sem Retroalimentação
GAN	Generative Adversarial Nets
IA	Inteligência Artificial
LSTM	Long-Short Term Memory
NA	Neurônio Artificial
PLN	Processamento De Linguagem Natural
RA	Reforço De Aprendizado De Máquina
RDF	Resource Description Framework
RF	Reconhecimento De Fala
RMP	Redes Multicamadas *Perceptron*
RNA	Redes Neurais Artificiais
RNC	Redes Neurais Convolucionais
RNEC	Redes Neurais Com Empilhamento De Níveis De Atenção
RNMD	Redes Neurais Com Memória Dinâmica
RNR	Redes Neurais Recorrentes
RNS	Redes Neurais Siamesas
RP	Retro-propagação
RPV	Resposta A Perguntas Visuais
VC	Visão Computacional

SUMÁRIO

1- Introdução

1.1- Contextualização

A tarefa de Respostas a Perguntas Visuais (RPV) pode ser definida da seguinte forma: dada como entrada uma coleção de triplas (I, Q, R), cada qual composta de uma imagem I, uma pergunta Q e uma resposta R (ambas em linguagem natural), o objetivo é produzir um modelo preditivo $\mathcal{M}$. Após o treinamento, podemos fornecer um par imagem/pergunta (I_q, Q_q) como entrada para $\mathcal{M}$. O modelo deve então produzir uma resposta R_q adequada para Q_q no contexto de I_q.

RPV é uma tarefa que une os campos da Visão Computacional (VC) e do Processamento de Linguagem Natural (PLN). A VC compreende tarefas que exigem que o modelo de predição interprete dados visuais e apreenda com eles. Desse modo, podemos dizer que as técnicas de VC têm o propósito de ensinar máquinas a enxergar o conteúdo de vídeo ou de imagens. Tarefas comuns de VC compreendem o reconhecimento, classificação, contagem e agrupamento de objetos (ANTOL et al., 2015).

O PLN é a área de pesquisa que aborda como modelos podem compreender e manipular dados na forma de texto e fala em linguagem natural, e os utilizar para realizar tarefas úteis. Tarefas comuns do PLN compreendem tradução de texto, correção ortográfica, transformação de texto em voz e vice-versa, geração automática de sentenças, etc.

Atualmente, as Redes Neurais Artificiais (RNA) são o estado da arte para a tarefa de RPV (X. LIN; PARIKH, 2017). Dado um par (I_q, Q_q), um modelo de RNA para RPV normalmente infere a resposta R_q em duas fases, conforme ilustrado na Figura 1. Na primeira fase, redes neurais apropriadas são usadas para extrair vetores de características tanto de Q_q quanto de I_q (separadamente). Em seguida, esses vetores são unidos em um único vetor, que codifica a informação contida no par (I_q, Q_q). Esse vetor é a entrada da segunda fase, que corresponde a outra rede neural cujo objetivo é produzir uma resposta adequada R_q para o par (I_q, Q_q).

Dá-se o nome de Aumento de Dados (AD) (*data augmentation*) ao processo de

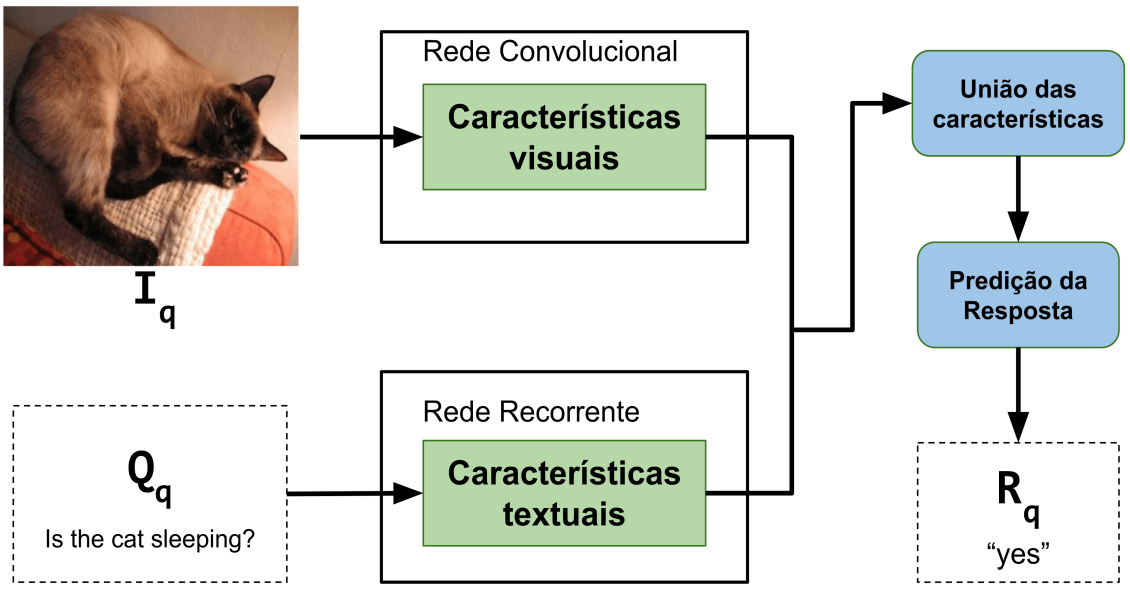


Figura 1 – Dado como entrada um par (I_q, Q_q), uma rede neural recorrente extrai um vetor de características do texto de Q_q, e uma rede neural convolucional extrai um vetor de características de I_q. Esses dois vetores são passados para uma terceira rede neural que infere a resposta.

estender (aumentar a quantidade de exemplos) de um conjunto de dados de treinamento. Técnicas para realizar AD têm o propósito de expandir um conjunto de treinamento original, e consequentemente, gerar maior variação de exemplos para serem utilizados no treinamento de modelos para aprendizado de máquina. No caso da tarefa de RPV, técnicas de AD podem ser aplicadas tanto sobre as imagens quanto sobre as perguntas do conjunto de treinamento original. Em particular, usaremos uma técnica de AD conhecida como Destilação de Dados (DD), que consiste em transferir o conhecimento contido em um conjunto de dados de treinamento para uma conjunto de dados que ainda não foi rotulado.

Além das técnicas de AD, outra abordagem que pode ser usada para tirar maior proveito de um conjunto de treinamento original é o Aprendizado Ativo (AA). Há diversas situações em que existem exemplos não rotulados em abundância. Técnicas de AA permitem consultar alguma fonte de informação externa (que pode ser inclusive um ser humano) para auxiliar na tarefa de rotular novos exemplos. De acordo com SETTLES (2012), a abordagem de AA conhecida como Abordagem Direcionada à Objetivo (ADO) (*goal-driven approach*) consegue um ganho significativo para tarefas de classificação binária (i.e., tarefas de classificação em que cada exemplo de treinamento está associado à uma das duas classes possíveis).

Nesta dissertação, consideramos a existência de (i) um conjunto de treinamento para RPV a ser aumentado e (ii) de uma coleção de imagens $\{I_e\}$ provenientes de uma fonte externa. A partir disso, esta dissertação apresenta duas propostas para melhorar o desempenho de modelos preditivos para RPV, por meio da adição de novos exemplos ao conjunto de treinamento original, utilizando técnicas de AD e AA. Como ponto de partida das duas propostas apresentadas nesta dissertação, pretendemos utilizar as imagens da fonte externa de forma inteligente para criar novos exemplos rotulados que, uma vez criados, podem ser adicionados ao conjunto de treinamento original. Essas propostas se baseiam na ideia básica de que, se uma imagem I_e proveniente da fonte externa for de alguma forma similar à imagem presente no exemplo (I, Q, R) do conjunto de dados original, então é provável que um novo exemplo seja criado utilizando I_e e os itens Q e R.

1.2- Justificativa

Um problema no contexto do treinamento de modelos preditivos para RPV é que as RNA atualmente utilizadas exigem uma quantidade significativa de dados de treinamento para conseguir aprender as características relevantes de Q e I de maneira satisfatória. Para cada ordem de grandeza no tamanho de um conjunto de dados de treinamento, estima-se que o aumento médio na acurácia do modelo treinado está entre 10 e 12 pontos percentuais (X. LIN; PARIKH, 2017). Com isso, a quantidade de dados de treinamento necessária tenha que aumentar exponencialmente para que se consiga um aumento linear na acurácia desses modelos. É importante ressaltar que tais dados de treinamento são normalmente rotulados por seres humanos de forma manual, tarefa na qual cada par composto por uma imagem e uma pergunta é associado a uma resposta por um curador. Conseguir rotular uma quantidade grande de dados para o treinamento das redes neurais envolvidas na tarefa de RPV é um procedimento muito custoso. Nesse contexto, se fazem relevantes técnicas que sejam capazes de aumentar, automática ou semi-automaticamente, a disponibilidade de dados para treinamento.

Melhorar o desempenho de modelos preditivos para RPV consiste não só na melhoria das arquiteturas de RNA, mas também no incremento da quantidade e qualidade dos dados utilizados para treinamento desses modelos. Investigar técnicas que ajudem

a selecionar novos exemplos de modo eficiente torna-se relevante para o aumento de desempenho do modelo a ser gerado, dado um conjunto limitado de dados com anotações relevantes.

A RPV utiliza dois tipos de dados diferentes, texto e imagem. Aumentar esses dados exige criatividade. Quando se trata de imagens, pode-se recortar, girar, alterar as cores (PEREZ; J. WANG, 2017). Quando se trata de texto, pode-se adicionar palavras à sentenças existentes, ou gerar novas sentenças a partir da combinação de palavras presentes no vocabulário. Entretanto, aumentar a quantidade de dados para treinamento será apenas a primeira etapa do processo. Com os novos dados o próximo passo será tirar o melhor proveito deles. Utilizando o AA teremos um melhor aproveitamento na seleção de exemplos entre os exemplos de treinamento disponíveis e também eliminando exemplos redundantes ou pouco significativos.

1.3- Objetivos

Diante dos desafios apresentados na seção 1.2, o objetivo geral desta dissertação é encontrar pontos de melhoria nas abordagens atuais para treinamento de modelos de RPV por meio da aplicação de técnicas de AD e de DD para aumentar o tamanho do conjunto de dado de treinamento original. Como restrição de escopo, consideramos nesta dissertação apenas o caso em que a resposta de cada exemplo de treinamento é binária (i.e., *sim* ou *não*). Os objetivos específicos desta dissertação são os seguintes:

1. Projetar e construir uma ferramenta de curadoria que possa ser usada para apresentar exemplos ainda não rotulados a um oráculo (no nosso caso, um ser humano), com a restrição de minimizar o esforço desse oráculo para enriquecer conjuntos de treinamento para predição de perguntas a respostas binárias.

2. Investigar a aplicação da técnica de AD conhecida como DD, que tem se mostrado especialmente promissora no contexto da tarefa de RPV (RADOSAVOVIC et al., 2018). Essa técnica se baseia em captar o conhecimento contido em um conjunto de treinamento para rotular exemplos previamente não anotados (rotulados).

1.4- Contribuições

As principais contribuições desse trabalho se encontram resumidas nos pontos a seguir:

- Uma ferramenta WEB de curadoria que dá suporte a seus usuários na criação de novos exemplos para estender um conjunto de treinamento (previamente existente) para RPV de forma colaborativa.
- Um procedimento de DD para permitir o aumento automático de um conjunto de treinamento para RPV.

1.5- Metodologia

Como metodologia de trabalho, avaliamos a utilização de técnicas de AD e AA em conjunto de dados de treinamento para RPV. Em particular, propomos duas abordagens diferentes e comparamos se houve ganho (do ponto de vista de qualidade do modelo de predição) com essas abordagens.

A primeira abordagem (indicada por ① na Figura 2) consiste em utilizar modelos de predição baseados em RNA para selecionar exemplos promissores em conjuntos de dados externos que possuem dados não rotulados. Esses modelos buscam por similaridade entre os exemplos do conjunto de treinamento e os exemplos do conjunto externo. Essa seleção ocorre em três etapas. A primeira avalia as características visuais das imagens e determina um grau de similaridade entre os exemplos do conjunto de treinamento e os exemplos do conjunto externo. Na segunda etapa, outro modelo avalia a similaridade entre o texto das perguntas associadas a imagem do conjunto de treinamento e a imagem candidata, visando melhorar a assertividade entre imagem e pergunta. Por fim, esses dados pré-selecionados são submetidos a um oráculo (i.e., curadores humanos) para confirmar ou descartar os dados pré-selecionados. Os dados selecionados nesta abordagem serão adicionados ao conjunto de treinamento original, formando o conjunto de treinamento aumentado.

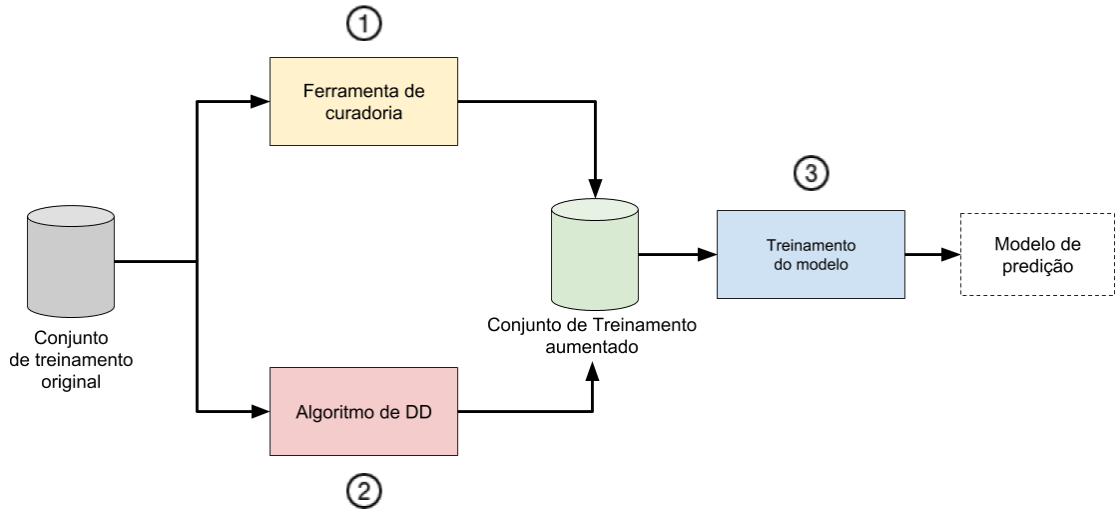


Figura 2 – Estas são as duas diferentes abordagens escolhidas para este trabalho. Em 1 temos a abordagem utilizando AA através da ferramenta de curadoria de dados, enquanto em 2 temos a abordagem de DD. Ambas as abordagens visam gerar novos dados rotulados para incorporação no conjunto de treinamento do modelo de RPV em 3.

A segunda abordagem (indicada por ② na Figura 2), consiste em aplicar uma técnica de AD conhecida como DD. Nesta abordagem, um modelo de predição (i.e., redes neurais) similar à da primeira abordagem, irá selecionar os dados não anotados de um conjunto externo que possuírem o maior grau de similaridade com os dados do conjunto de treinamento original. A diferença desta abordagem em relação à primeira é que, em vez de submeter os dados a um oráculo, esses dados são submetidos a um comitê de modelos preditores, cujos componentes votam em quais exemplos devem ser incorporados. Esse comitê é formado por dois ou mais modelos de predição treinados em subconjuntos diferentes do conjunto de dados de treinamento original, de modo a cada membro do comitê gerar uma predição diferente sobre os dados externos avaliados. Os votos do comitê são bases em média de pontuação e soma absoluta das pontuações dadas pelos membros do comitê. Os dados selecionados recebem uma cópia dos rótulos dos respectivos dados do conjunto de treinamento original. Os novos dados rotulados serão adicionados ao conjunto de treinamento original, formando o conjunto de treinamento aumentado.

Ao final, os conjuntos de dados de treinamento aumentados que foram gerados por cada uma dessas abordagens são utilizados para treinar o modelo de aprendizado para RPV, e seus resultados avaliados, como podemos ver em 3 da Figura 2.

1.6- Organização dos Capítulos

O restante desta dissertação está organizado conforme a seguir. No Capítulo 2 será apresentado o referencial teórico sobre RPV. Nesse capítulo também estão descritos os principais conjuntos de dados que são utilizados para treinamento e validação de modelos de RPV. No Capítulo 3 são descritos os principais trabalhos relacionados com o presente trabalho, divididos entre as abordagens de AA com AD. No capítulo 4 é apresentada a proposta do trabalho. Neste capítulo demonstramos como foram utilizadas as técnicas de AA e AD. No Capítulo 5 são apresentados como foram executados os experimentos e seus resultados são analisados. Por fim, no Capítulo 6, são apresentadas a conclusões obtidas com o trabalho, além da realização de uma análise retrospectiva e a apresentação de possíveis trabalhos futuros.

2- Respostas a Perguntas Visuais

2.1- Introdução

RPV é uma tarefa complexa que une os campos da VC e do PLN. Modelos de aprendizado de RPV precisam ter um entendimento detalhado da imagem e compreender muito bem o texto da pergunta. O nível de compreensão necessário para executar esta tarefa torna a RPV um desafio completo de Inteligência Artificial (IA) (ANTOL et al., 2015).

Ao contrário dos modelos baseados puramente em linguagem natural (texto ou voz) estudados extensivamente na comunidade de PLN, na RPV as redes neurais são projetadas para conseguir inferir uma resposta R automaticamente dado uma pergunta Q, de acordo com o conteúdo de uma imagem I. A maioria dos modelos de RPV recentes são propostos com base em RNA. Responder à uma pergunta sobre uma imagem requer que a rede neural desenvolva uma cadeia complexa de raciocínio sobre as características presentes no texto e na Figura (YANG et al., 2016).

Este capítulo está organizado da seguinte forma: na Seção 2.2 abordaremos os principais conceitos sobre as RNAs. Na Seção 2.3 serão descritas as principais arquiteturas de modelos de aprendizado para a tarefa de RPV. Na Seção 2.4 serão descritos os principais conjuntos de dados com foco em tarefas de RPV. Na Seção 2.5 abordaremos os conceitos e técnicas de AD. Na Seção 2.6 apresentaremos alguns conceitos sobre a AA. E, na Seção 2.7 será exposta uma técnica em particular de AD conhecida como DD (*Data Distilation*).

2.2- Redes Neurais e Aprendizado Profundo

O Aprendizado de Máquina (AM) é a base para as abordagens atuais de IA. Dentro do AM temos o subcampo de Aprendizagem Profunda (AP) que se utiliza de métodos analíticos para representar conceitos. A AP organiza esses conceitos em múltiplas

camadas que seguem uma hierarquia de conhecimento, onde camadas de níveis mais baixos se combinam para compor representações de aprendizagem em camadas de níveis mais altos de abstração (LECUN; YOSHUA BENGIO; GEOFFREY E. HINTON, 2015). As RNAs que possuem mais de uma camada intermediária (oculta), são chamadas de RNAs profundas, daí vêm o nome da AP.

Mesmo o conceito de AP ter sua origem na década de 80 (WILLIAMS; G E HINTON, 1986), a sua utilização se tornou mais difundida a partir dos anos 2000, devido ao crescimento na quantidade de dados e de poder computacional necessários para executar as tarefas de AP de modo satisfatório (LECUN; YOSHUA BENGIO; GEOFFREY E. HINTON, 2015).

O AP tem obtido ótimos resultados na resolução de problemas ligados às áreas de Visão Computarional (KRIZHEVSKY; SUTSKEVER; GEOFFREY E HINTON, 2012), Reconhecimento de Fala (RF) (GRAVES; MOHAMED; GEOFFREY E. HINTON, 2013), Aprendizado Supervisionado (AS) (LECUN; YOSHUA BENGIO; GEOFFREY E. HINTON, 2015), PLN (Y. BENGIO; COURVILLE; VINCENT, 2013) e Reforço de Aprendizado de Máquina (RA) (SCHMIDHUBER, 2015; MINSKY, 1961). Para os desafios de VC e PLN , a AP é o estado da arte, obtendo resultados muito superiores em relação às abordagens tradicionais (LECUN; YOSHUA BENGIO; GEOFFREY E. HINTON, 2015).

Para a criação de modelos de AP com boas representações internas de aspectos do mundo real, as RNAs foram inspiradas em modelos computacionais biológicos, sendo capazes de realizar tarefas de AM tais como percepção visual, auditiva e compreensão de linguagem natural. As RNAs são apresentadas como sistemas de Neurônio Artificial (NA) interconectados que podem computar valores (LECUN; KAVUKCUOGLU; FARABET et al., 2010).

Na Figura 3 é apresentado o processo de ativação de um NA iniciado através de um sinal de estímulo externo, como a entrada de *pixels* de uma imagem. Os NAs quando ativados ponderam, transformam e repassam o sinal para o NA da próxima camada da RNA, repetindo esse processo até a camada de saída (LECUN; KAVUKCUOGLU; FARABET et al., 2010).

Os NAs possuem uma etapa de pré-ativação feita através de uma função linear, porém funções lineares são insuficientes para dar expressividade ao NA e, por isso, é adicionada uma função de ativação não-linear na saída de cada NA, aumentando a sua expressividade (SUTSKEVER; MARTENS; GEOFFREY E HINTON, 2011). Na Figura 4

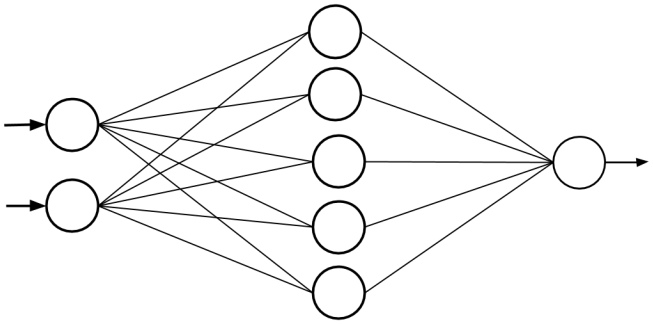


Figura 3 – Arquitetura de uma Rede Neural Artificial. Os NAs de entrada recebem o estímulo externo e repassam para a próxima camada da RNA, essa por sua vez repete o processo até que o sinal alcance a camada de saída (WILLIAMS; G E HINTON, 1986)

.

está representado a anatomia de um NA.

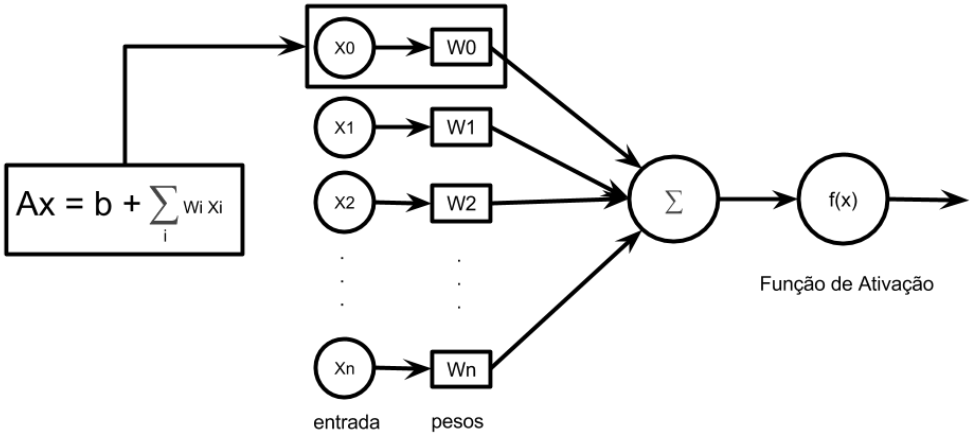


Figura 4 – Neurônio Artificial. A função linear $Ax = b$ somada aos pesos atribuídos ao NA, funciona como uma função de pré-ativação. A função não-linear $f(x)$ é aplicada para aumentar a expressividade do neurônio (WILLIAMS; G E HINTON, 1986)

.

O processo de aprendizagem se dá através de algoritmos de Retro-propagação (RP) (*back-propagation*) em que o sinal de entrada é propagado através de todas as camadas da RNA até a sua saída. Em seguida o resultado obtido é comparado com o resultado esperado, então calcula-se o vetor de erros e atualizam-se os pesos de cada neurônio em cada camada da rede neural, buscando aproximar os resultado obtido do resultado esperado (WILLIAMS; G E HINTON, 1986).

Abaixo são descritas as arquiteturas de RNAs utilizadas neste trabalho.

- **Redes Neurais Artificias sem Retroalimentação (*Feed-Forward Neural Network*)**. As Redes Neurais Artificiais sem Retroalimentação (FNN)s têm seus NAs agrupados em camadas. O sinal percorre a RNA em uma única direção, da entrada para a

saída. NAs da mesma camada não se conectam (SCHMIDHUBER, 2015).

- **Redes Neurais Convolucionais (Convolutional Neural Network)**. As Redes Neurais Convolucionais (RNC)s são uma FNN onde os NAs são organizados de modo a representar as sobreposições do campo visual. Esse tipo de RNA são variações de Redes Multicamadas *Perceptron* (RMP), que são desenhadas para ter o mínimo de pré-processamento (Y. BENGIO; COURVILLE; VINCENT, 2013). As RNCs são amplamente utilizadas em tarefas de VC.

- **Redes Neurais Recorrentes (*Recurrent Neural Network*)**. Nas Redes Neurais Recorrentes (RNR), a saída de alguns NAs alimentam a entrada outros NAs situados na mesma camada ou em camadas anteriores. O Sinal percorre as RNAs em duas direções possibilitando a capacidade de memória dinâmica de curto prazo e capacidade de representação de estados mais complexos em sistemas dinâmicos (SCHMIDHUBER, 2015).

- **Redes Recorrentes *Long-Short Term Memory***. As Long-Short Term Memory (LSTM) são um tipo especial de RNR capazes de aprender dependências de longo prazo através de um mecanismo de memória. Essas redes foram introduzidas por (HOCHREITER; SCHMIDHUBER, 1997). As LSTM funcionam muito bem em uma grande variedade de problemas e são amplamente utilizadas em PLN, assim como em outros problemas relacionados à dados sequenciais.

- **Redes Neurais Siamesas (*Siamese Neural Network*)**. Nas Redes Neurais Siamesas (RNS), duas RNAs chamadas de cabeças, extraem vetores de características de dois objetos (i.e. imagens, fragmentos de texto, etc) que foram dados como entradas para para as RNAs da cabeça e, então os vetores de características desses objeto são enviados à saída da RNS, chamada de cauda e que possui algoritmos para calcular a similaridade entre os vetores de características dos objetos da entrada. As RNS foram originalmente propostas por (BROMLEY et al., 1993) e geralmente são aplicadas a problemas de reconhecimento de objetos similares (i.e. verificação de assinaturas, reconhecimento facial, etc).

2.3- Arquiteturas de Redes Neurais para RPV

Atualmente, existem quatro tipos de arquitetura mais comuns para RPV. A seguir descrevermos cada uma dessas arquiteturas: A Seção 2.3.1 aborda a arquitetura de incorporação conjunta; a Seção 2.3.2 aborda a arquitetura baseada em modelos de composição modular; a Seção 2.3.3 aborda a arquitetura de modelos com mecanismos de atenção, e a Seção 2.3.4 aborda a arquitetura de modelos com aprimoramento de conhecimento.

2.3.1 Arquiteturas de Incorporação Conjunta

O conceito de incorporação conjunta foi explorado nos trabalhos de (MAO et al., 2015; DONAHUE et al., 2015; VINYALS et al., 2015; XU et al., 2015). Estes trabalhos foram motivados pelo avanço de técnicas de VC e PLN e permitiram aprender representações extraídas de texto e imagem num mesmo espaço de processamento.

(VINYALS et al., 2015) propõem um modelo neural e probabilístico, que pode ser treinado para receber uma imagem I como entrada e produzir uma sentença R como saída. Em R, cada palavra é selecionada a partir do vocabulário conhecido pela rede tentando maximizar a probabilidade de obter uma descrição adequada para a imagem. Trabalhos mais recentes como o Multimodal Compact Bilinear (JIANG et al., 2014) e Redes Residuais Multimodais (KIM et al., 2016) trouxeram melhorias significativas para esta abordagem.

Por se tratar da forma mais direta, a *Arquitetura de Incorporação Conjunta* é a base das demais abordagens de RPV. O modelo de aprendizado que utilizaremos nesse trabalho, segue essa Arquitetura. A Figura 1, ilustra o funcionamento dessa Arquitetura.

2.3.2 Arquiteturas de Composição Modular

Visando dar maior flexibilidade às arquiteturas monolíticas de RNRs e RNCs, uma nova proposta de arquitetura baseada em composição de módulos foi sugerida por (ANDREAS et al., 2016) e depois aprimorada por (JOHNSON et al., 2017b).

Essa arquitetura apresenta uma abordagem que realiza a composição de módulos de redes neurais de maneira dinâmica, de modo que qualquer composição de módulos possa gerar uma rede neural válida. Potenciais ganhos, estão na melhor capacidade de supervisão do aprendizado, melhor transferência de aprendizado e a possibilidade de treinar diferentes composições de redes para cada tarefa (ANDREAS et al., 2016). A Figura 5 ilustra um exemplo de rede neural de composição modular.

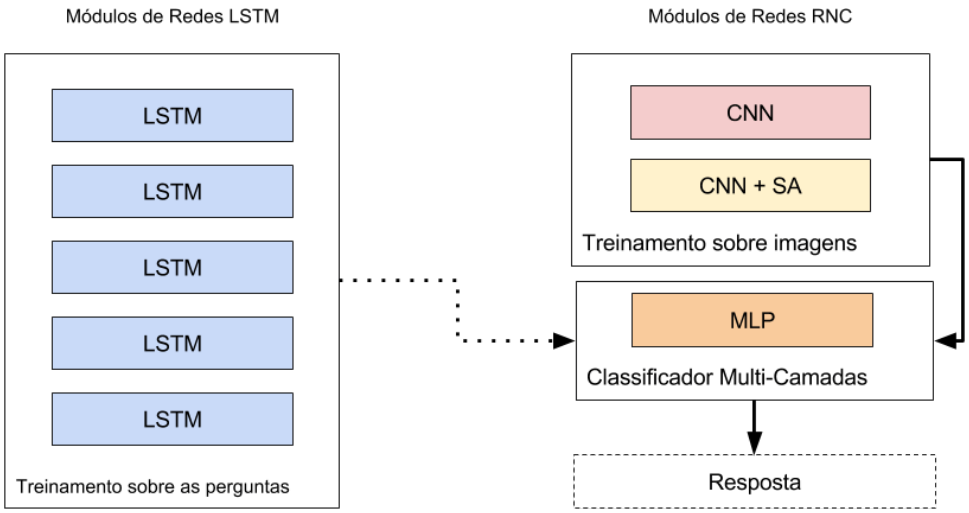


Figura 5 – Arquitetura de Redes Modulares criada de modo que qualquer composição entre os modulos de redes LSTM resultem em uma RNRs válida. As redes RNCs o classificador RMP são compostos por apenas um módulo cada.

Com essas características, é possível explorar a estrutura linguística compositiva das perguntas, como aquelas que possuem uma grande variedade e demandam várias etapas de processamento, como reconhecimento de conceitos, contagem, classificação, etc. As entradas e saídas de cada módulo nessa arquitetura, podem ser imagens, regiões (de atenção) ou etiquetas de classificação (texto).

A análise textual é um passo crucial nesse método onde basicamente identifica-se a relação gramatical entre as partes da sentença. (ANDREAS et al., 2016) utilizam regras

prescritas para transformar as árvores de análise em estruturas de busca, na forma de composição de módulos, como poderemos observar na Figura 6.

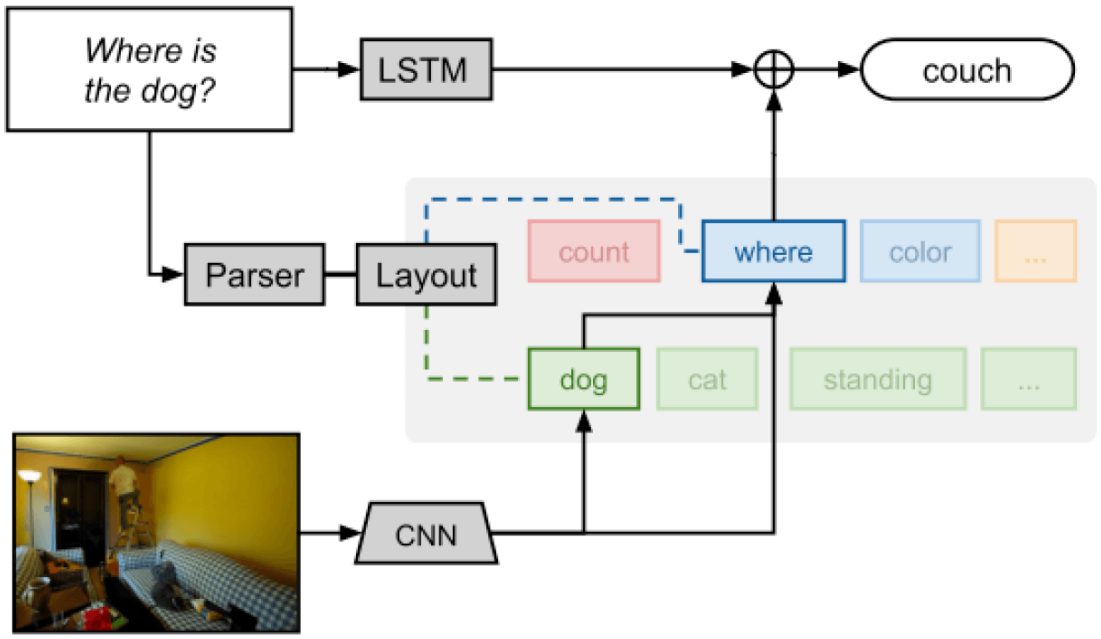


Figura 6 – Nessa arquitetura (ANDREAS et al., 2016) utilizam regras pré-escritas para transformar as árvores de análise em estruturas de busca. Cada módulo é especializado em um tipo de característica. Imagem retirada de (ANDREAS et al., 2016).

Arquiteturas de Redes de Memória Dinâmica

As Redes Neurais com Memória Dinâmica (RNMD) (KUMAR et al., 2016) são um tipo particular de arquitetura modular. A arquitetura da RNMD é composta por quatro módulos principais, que podem ser implementados de maneira independente.

- **Módulo de Entrada**: codifica as entradas de texto bruto em vetores de características. Esse módulo se concentra em problemas de linguagem natural.
- **Módulo de Perguntas**: este módulo codifica a pergunta em uma representação vetorial e a envia para a memória episódica.
- **Módulo de Memória Episódica**: Dada uma coleção de representações, o módulo de memória episódica escolhe quais partes do vetor de características ele deve

concentrar seu mecanismo de atenção. Depois ele produz uma memória vetorial, contendo a questão atual e a questão anterior e mantendo informações recentes relevantes a cada iteração.

- **Módulo de Resposta**: O módulo de resposta gera uma resposta a partir do vetor de memória final do módulo de memória episódica.

Os quatro módulos que compõe a RNMD podem ser vistos na Figura 7.

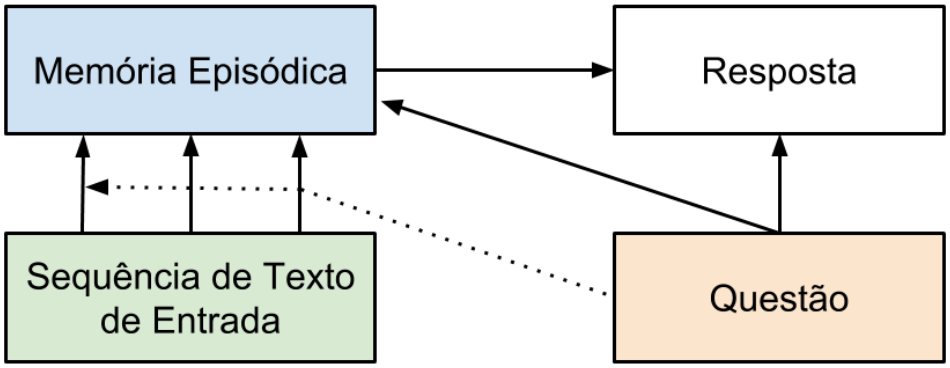


Figura 7 – Redes Neurais de Memória Dinâmica. A comunicação entre cada módulo é indicada por setas e usa representações vetoriais. Perguntas ativam os portões que permitem que vetores para determinadas entradas sejam dados ao módulo de memória episódica. O estado final da memória episódica é a entrada para o módulo de resposta. Imagem retirada de (KUMAR et al., 2016).

A chave deste método é permitir que o módulo de memória episódica seja executado várias vezes sobre os fatos, permitindo um raciocínio transitivo. A novidade principal é o uso de uma função de custo em cada uma dessas passagens, ao invés de usá-la uma única vez no final. Após o treinamento, a inferência é realizada usando apenas uma dessas passagens.

Em comparação com as Redes Neurais Modulares, as RNMD possuem um desempenho semelhante em perguntas binárias, porém são um pouco inferiores em consultas numéricas e notavelmente melhores em outros tipos de perguntas. (XU et al., 2015) propuseram um novo modelo de RNMD com mecanismo de atenção, e a chamou de DMN+. Como esta implementação foi possível demonstrar resultados notáveis sem a ajuda de treinamento supervisionado.

2.3.3 Arquiteturas com Mecanismo de Atenção

(YANG et al., 2016) apresentam as Redes Neurais com Empilhamento de Níveis de Atenção (RNEC) que utilizam a representação semântica de uma pergunta Q para buscar uma resposta R em regiões relacionadas de uma imagem I.

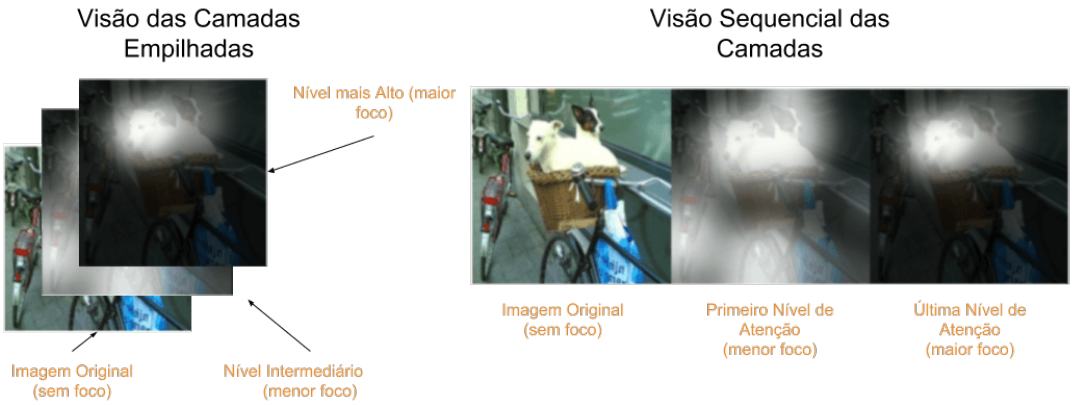


Figura 8 – Rede Neural com Sobreposição de Níveis de Atenção. Imagem adaptada de (XU et al., 2015)

Em modelos com mecanismo de sobreposição de níveis de atenção, o nível alto dá foco em regiões de I que são mais relevantes para a obtenção de R. Combinando as características extraídas de I com as regiões de atenção, a RNEC é capaz de predizer R dado Q.

Comparada com modelos que simplesmente combinam o vetor características de Q com o vetor de características de I, os modelos com mecanismo de atenção constroem um acesso mais informativo, uma vez que regiões visuais que são mais relevantes para R recebem um peso maior. Os mecanismos de atenção permitem que os modelos de aprendizado utilizem Q para selecionar quais partes da imagem devem receber atenção.

2.3.4 Arquiteturas com Aprimoramento de Conhecimento

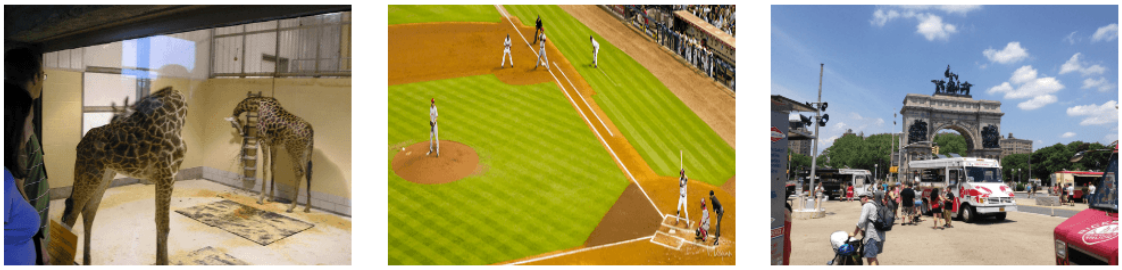
As tarefas de RPV envolvem compreensão do conteúdo das imagens, mas, requer informações prévias não visuais, que podem variar de “senso comum” ao conhecimento

específico ou mesmo enciclopédico.

Esta observação permite localizar dois principais pontos fracos das abordagens de incorporação conjunta. Primeiro, eles só podem capturar o conhecimento que está presente no conjunto de treinamento, e é óbvio que os esforços na ampliação de conjuntos de dados nunca alcançarão uma cobertura completa do mundo real. Em segundo lugar, as redes neurais treinadas em tais abordagens têm uma capacidade limitada, o que também é inevitável, considerando a quantidade de informação que se deseja aprender. Uma alternativa é desacoplar o raciocínio do armazenamento.

Uma grande quantidade de trabalhos se propõem a estruturar representações do conhecimento em forma de fatos, que podem ser consumidos por humanos e computadores, como DBPedia (AUER et al., 2007), Freebase (BOLLACKER et al., 2008), YAGO (SUCHANEK; KASNECI; WEIKUM, 2007), OpenIE (BANKO et al., 2007), NELL (CARLSON et al., 2010), WebChild (TANDON et al., 2014) e ConceptNet (H. LIU; SINGH, 2004).

Nessas bases de conhecimento cada fato F, que representa um pedaço de conhecimento que é tipicamente representado na forma de uma tripla $(arg1, rel, arg2)$, onde $arg1$ e $arg2$ representam dois conceitos e rel a relação entre esses conceitos. Desse modo, os elementos de F formam um grafo interligado, seguindo uma especificação modelada em Resource Description Framework (RDF) (PAN, 2009). Na Figura 9, podemos ver que as perguntas são respondidas utilizando um conhecimento de senso comum, que não está presente nas imagens.



Q: Tell me the common property of the animal in this image and elephant.
A: mammal, animals in Africa

Q: List all equipment I might use to play this sport.
A: baseball bat, baseball, baseball glove, baseball field

Q: Is the image related to tourism ?
A: yes

Figura 9 – Modelos aprimorados com conhecimento externo. O modelo é capaz de responder as perguntas utilizando um conhecimento que não está explícito nem no texto da pergunta nem na imagem. Imagem retirada de (P. WANG et al., 2015).

2.4- Conjunto de Dados para RPV

Existe um grande número de conjuntos de dados utilizados para avaliar o desempenho de modelos em tarefas de RPV. Cada um desses conjuntos de dados possui no mínimo, uma imagem I, uma pergunta Q e uma resposta R. Sendo que alguns desses conjuntos possuem dados adicionais, como rótulos de regiões de atenção, respostas incorretas e outros rótulos adicionais nas imagens.

2.4.1 Principais Conjuntos de dados

DAQUAR Foi o primeiro conjunto de dados projetado especificamente para a tarefa de RPV. O conjunto apresenta mais de 12.000 pares de perguntas-respostas produzidas por seres humanos sobre imagens coloridas, o conjunto foi apresentado como uma abordagem moderna para testes de Turing visuais (MALINOWSKI; FRITZ, 2014).

COCO-QA Esse conjunto de dados representa um esforço para aumentar a escala dos conjuntos de dados para tarefas de RPV. Nesse conjunto de dados as perguntas e respostas foram geradas a partir das legendas das imagens do conjunto de dados COCO-QA (REN; KIROS; ZEMEL, 2015).

FM-IQA Esse conjunto de dados utiliza as imagens do conjunto COCO-QA, porém as anotações para as imagens foram feitas de maneira colaborativa por várias pessoas e de forma livre. Por ser um conjunto de dados com uma diversidade muito grande, ele exige que o algoritmo tenha um grande nível de compreensão visual e textual. Originalmente esse conjunto de dados foi criado em chinês, e depois foi traduzido para o inglês (GAO et al., 2015).

VQA-Real Um dos conjuntos de dados mais amplamente utilizados para RPV. Este conjunto foi concebido pela equipe da *Universidade Virginia Tech*, e, normalmente, nos referimos a ele apenas como *VQA*. Ele está dividido em duas partes: a primeira contendo imagens do mundo real, e, a segunda contendo ilustrações, sendo chamada

de *VQA-Abstract* (ANTOL et al., 2015).

VQA-Abstract Essa parte do conjunto está separada do conjunto de imagens do mundo real, e tem seu próprio conjunto de perguntas e respostas. O objetivo do conjunto de dados é de avaliar o aprendizado de conceitos de alto nível (ANTOL et al., 2015).

VQA-Balanced É uma versão do conjunto de dados VQA-abstract que contém cenas complementares e similares as já existentes, apenas com algumas mudanças sutis que ocasionam a alteração na resposta. O objetivo deste conjunto é reduzir o viés estatístico presente no conjunto de dados original (P. ZHANG et al., 2016).

Visual GENOME Maior conjunto de dados até o momento para RPV, com cerca de 1 milhão e 700 mil perguntas em mais de 100 mil imagens com anotações de objetos, atributos e relacionamentos dentro de cada imagem para aprender esses modelos. Cada imagem tem uma média de 21 objetos, 18 atributos e 18 relações em pares entre objetos (KRISHNA et al., 2017).

Visual 7W É um subconjunto do Visual GENOME, que possui rótulos adicionais. As perguntas são configuradas sob a forma de múltipla escolha, onde cada pergunta possui quatro respostas, sendo apenas uma a correta. 7W significa *who, what, where, when, why, how, e which* (Y. ZHU et al., 2016).

Visual Madlibs Esse conjunto de dados foi projetado para avaliar tarefas do tipo *"preencha as lacunas"*. O conjunto foi criado através da geração automática seguindo um modelo, e reúne descrições específicas sobre pessoas e objetos, suas aparências, atividades e interações, bem como inferências sobre a cena geral ou seu contexto mais amplo (YU et al., 2015).

KB-VQA Conjunto de dados criado para avaliar técnicas que empregam o uso de bases de conhecimento externa. As questões no conjunto de dados são geradas por seres humanos com base em vários modelos pré-definidos. As perguntas recebem um dos três rótulos que refletem a informação necessária para respondê-los: "Visual", "Senso comum" e "Enciclopédico" (P. WANG et al., 2015).

FVQA Esse conjunto de dados contém apenas perguntas com questionamentos não visuais. Foi projetado para conter rótulos adicionais que facilitam a supervisão de treinamento de métodos que usam bases externas de conhecimento (WANG et al., 2016), como a DBPedia (AUER et al., 2007).

CLEVR É um conjunto de dados de diagnóstico que testa uma variedade de habilidades de raciocínio visual. Contém um viés estatístico mínimo e apresenta rótulos detalhados que descrevem o tipo de raciocínio que cada pergunta requer (JOHNSON et al., 2017a).

2.4.2 Conjuntos de Dados Balanceados

P. ZHANG et al. (2016) demonstraram que os conjuntos de dados para RPV apresentam um forte viés estatístico, como por exemplo, consultas iniciada com as palavras "*What sport is*" possui a resposta "*Tennis*" em 41% dos casos. Quando se trata perguntas binárias, 69% das perguntas podem ser respondidas com a palavra "SIM". Por conta desse viés estatístico a rede neural pode aprender a responder as perguntas apenas usando seu conhecimento textual e estatístico, sem precisar realmente aprender sobre as características da imagem. Os autores tentam contrapor esse viés estatístico aumentando estes conjuntos de dados com novos exemplos de imagens que possuem respostas opostas para uma mesma pergunta. O aumento dos dados também envolve adicionar rótulos às imagens que podem ser acessados pelo modelo no momento do treinamento.

Figura 10 – Exemplo de Imagens retiradas de um conjunto de dados expandido contendo cenas complementares para uma mesma pergunta. No conjunto há duas imagens bem parecidas, sendo uma com resposta positiva e outra com resposta negativa. Imagem retirada de P. ZHANG et al. (2016)

A Tabela 1 resume os principais conjuntos de dados utilizados para treinamento de modelos para RPV. Nesta tabela constam informações sobre a origem e quantidade

Tabela 1 – Principais Conjuntos de Dados para RPV. Dados retirados de Q. WU et al. (2016b)

Conjunto de Dados	Origem das Imagens	Quantidade de Imagens	Quantidade de Perguntas	Perguntas Imagens
DAQUAR	NYU-Depth V2	1449	12.468	8,6
COCO-QA	COCO	117684	117.684	1
FM-IQA	COCO	120360	-	-
VQA-Real	COCO	204721	614.163	3
VQA-Abstract	ClipArt	50000	150.000	3
VQA-Balanced	ClipArt	15623	33.379	2,13
Visual Genome	COCO	108.249	1.700.000	13,3
Visual 7W	COCO	47300	300.327	6,34
Visual Madlibs	COCO	10738	360.001	33,5
KB-VQA	COCO	700	2.402	3,43
FRPV	ImageNet	1906	4.608	2,41
CLEVR	CLEVR	100.000	999.968	9,99

de imagens e de perguntas presentes e cada um desses conjuntos de dados.

2.5- Aumento de Dados

Os modelos baseados em redes neurais profundas RNC e RNR melhoram a medida que aumenta a quantidade de dados disponíveis para o seu treinamento. Mesmo quando os dados são de qualidade inferior, esses algoritmos conseguem melhorar seu desempenho, desde que dados úteis possam ser extraídos pelo modelo a partir desse conjunto de dados.

Uma forma de aumentar o conjunto de dados para RPV é realizar transformações sobre o componente I de cada exemplo de treinamento. Por exemplo, pode-se recortar, girar, alterar as cores, etc. Com relação ao componente Q, pode-se substituir palavras em seu conteúdo ou mesmo gerar novas perguntas a partir da combinação de palavras existentes no vocabulário. Uma vantagem dessas abordagens é que novos exemplos de treinamento podem ser produzidos rapidamente e sem intervenção humana. Por outro lado, não há garantias de que os exemplos resultantes dessas transformações produzam algum efeito positivo sobre a acurácia dos modelos.

De acordo com VAN DYK e MENG (2001), “o termo aumento de dados refere-se a métodos para a construção de algoritmos iterativos de otimização ou amostragem através da introdução de dados não observados ou variáveis latentes” .

No caso particular do aumento de dados aplicado a imagens, são normalmente realizadas transformações sobre o conjunto de imagens originais para produzir novas

imagens. Transformações comumente realizadas sobre as imagens consistem em recortes, aplicação de filtros, rotação, etc. Cada uma das imagens transformadas é adicionada ao conjunto original. Desse modo, é possível aumentar várias vezes o tamanho dos conjuntos de dados. Na Figura 11 podemos observar como algumas transformações são aplicadas sobre uma imagem para gerar novos exemplos de treinamento.

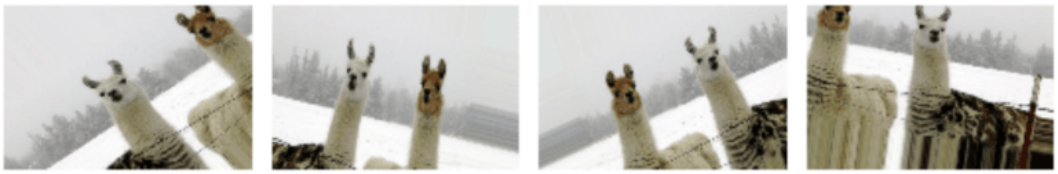


Figura 11 – Exemplo de Aumento de Dados, no qual uma imagem passou por várias transformações para gerar novos exemplos de treinamento. Imagem retirada de P. WANG et al. (2015).

R. WU et al. (2015) afirmam que a AD é fundamental para a melhoria do desempenho e generalização de modelos de baseados em redes neurais, porém os autores advertem que técnicas de AD fazem com que a quantidade de dados disponíveis exploda em quantidade, o que demanda um poder de processamento muito maior para treinar os modelos.

2.6- Aprendizado Ativo

Um problema no contexto do treinamento de modelos preditivos para RPV é que as redes neurais atualmente utilizadas exigem uma quantidade significativa de dados para conseguir aprender as características de Q e I de maneira satisfatória. De fato, é provável que esses novos exemplos sejam redundantes com relação aos já existentes. O efeito disso é que teríamos um conjunto de treinamento maior (e que, portanto, demanda mais tempo de treinamento), mas sem retorno significativo na acurácia do modelo. O ideal é que os exemplos produzidos para estender (aumentar) o conjunto de treinamento original sejam selecionados de tal forma a serem minimamente redundantes, com o propósito de manter a diversidade nos exemplos do conjunto de dados resultante.

Coletar grandes quantidades de dados é uma tarefa muito dispendiosa e mesmo

que se obtenha uma quantidade suficiente de dados para o treinamento, ainda sim, faltará exemplos para os conceitos mais raros. Na Figura 12 o AA ajuda a abordar esse problema, selecionando os dados com quantidade de aprendizado mais significativo dentro do conjunto. Na Figura 12 podemos ver o ciclo de AA onde o modelo primeiramente é treinado em um subconjunto de dados retirado do conjunto original, após isso o aprendizado é expandido através da seleção de exemplos potencialmente mais informativos. O AA geralmente é utilizada sobre conjuntos de treinamento para tarefas isoladas como VC ou PLN (X. ZHU; LAFFERTY; GHAHRAMANI, 2003; HOULSBY et al., 2011; GORDON; VAN DURME, 2013; SENER; SAVARESE, 2017), porém os conjuntos de treinamento para tarefas de RPV contemplam tanto dados visuais (i.e., imagens) quanto textuais (i.e., perguntas e respostas), o que adiciona mais complexidade para utilizar o AA. X. LIN e PARIKH (2017), entretanto, abordam o AA para modelos de VC e PLN em conjunto para utilizar em tarefas de RPV. O ponto chave do AA é que o modelo é capaz de selecionar quais exemplos são mais significativos para o treinamento.

Existem três abordagens típicas de AA:

Síntese de consultas membros possui utilização razoável para muitos problemas, porém rotular as instâncias arbitrárias pode ser inviável se as anotações forem realizadas por uma pessoa.

Seleção de exemplos sequenciais muito utilizada em dados sequencias, como texto e voz, essa abordagem reduz o esforço de anotação, mas limita o tamanho do conjunto de dados utilizado no aprendizado.

Seleção de exemplos a partir de um subconjunto uma abordagem amplamente utilizada em tarefas gerais de aprendizado de máquina em geral, sendo comprovadamente eficiente em classificação de texto (HOI; JIN; LYU, 2006), de imagens (C. ZHANG; T. CHEN, 2002), de vídeo (HAUPTMANN et al., 2006), voz e fala (TUR; HAKKANI-TÜR; SCHAPIRE, 2005), extração de informação (SETTLES; CRAVEN, 2008) e diagnóstico de doenças (Y. LIU, 2004).

Esta dissertação considera o terceiro cenário apresentado. Ainda, existem algumas estratégias para a selecionar quais dados devem ser rotulados. Para este trabalho, escolhemos a estratégia de *Redução Esperada do Erro*, que consiste em selecionar para rotular exemplos de dados que trariam a maior redução de erro de generalização do modelo. Essa abordagem de decisão visa medir não o quanto o modelo provavelmente

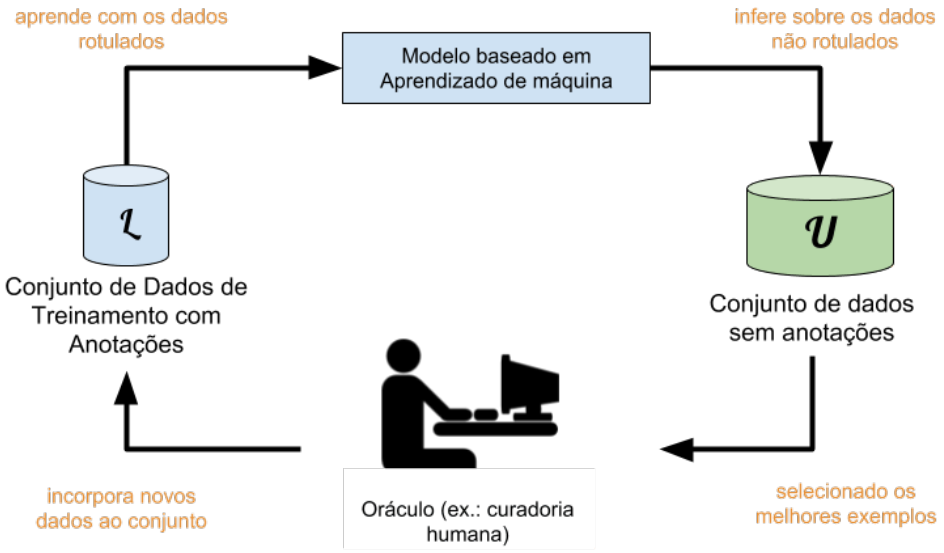


Figura 12 – No AA um subconjunto é retirado do conjunto de dados original e anotado, após essa fase mais dados são selecionados do subconjunto restante. O critério para a seleção desses dados geralmente leva em conta o grau de informação nova presente no exemplo a ser escolhido, esses novos dados então são adicionados ao conjunto de treinamento e o ciclo recomeça, até que não haja mais exemplos significativos no conjunto de dados restantes, que serão descartados. Imagem adaptada de SETTLES (2012)

mudará, mas quanto seu erro de generalização provavelmente será reduzido. A ideia é estimar o erro futuro esperado de um modelo treinado usando as instâncias não rotuladas restantes. O objetivo aqui é reduzir o número total de previsões incorretas.

2.7- Destilação de Dados

A DD é um método de aprendizado supervisionado que visa transferir o conhecimento contido em dados previamente anotados para dados que ainda não possuem anotações, por meio de um conjunto de modelos de aprendizado treinados para reconhecer dados similares em outros conjuntos. Esses modelos juntos formam um comitê que escolhem por voto quais dados podem receber as anotações. A agregação de múltiplos modelos é um forma eficiente de melhorar a precisão de um método (KROGH; KROGH; VEDELSBY, 1995).

Este método é um regime especial de aprendizado semi-supervisionado (RADOSAVOVIC et al., 2018), que ao contrário de BUCILUA, CARUANA e NICULESCU-MIZIL

(2006) e G. HINTON, VINYALS e DEAN (2015) que buscam destilar o conhecimento a partir do modelo, a DD busca a destilação a partir dos dados propriamente ditos.

A DD possui quatro passos para ser executada:

1. **Treinamento de modelos usando dados já rotulados**, utilizando o aprendizado supervisionado tradicional, com o objetivo de treinar os modelos que servirão ao comitê de predições fornecendo suas predições;

2. **Transformações nos dados de entrada** com o objetivo de ter entradas diferentes para treinar instâncias diferentes do modelo. O único requerimento para as transformações é que elas gerem modelos suficientemente diferentes para formar o comitê. As transformações podem melhorar o modelo por uma boa margem (RADOSAVOVIC et al., 2018);

3. **Gerar anotações para os dados não rotulados**. Neste passo, agrega-se as predições das múltiplas instâncias do modelo, de modo a obter uma predição de qualidade superior em relação a predição de uma única instância do modelo de aprendizado;

4. **Destilação do Conhecimento**. Nesta etapa utiliza-se o conhecimento adquirido a partir dos dados não rotulados para gerar uma nova instância do modelo de aprendizado, chamada de modelo estudante. Essa instância do modelo será treinada usando a união dos dados originais com a adição dos dados candidatos que foram anotados automaticamente.

A Figura 13 demostra o fluxo de funcionamento do método de destilação de dados. Onde temos como entrada uma imagem rotulada que passa por N transformações (i.e., recorte, rotação, alterações na cor) para gerar N variações da imagem. Cada variação da imagem será utilizada para treinar um de N modelos de predição. Essas instâncias treinadas dos modelos formam um comitê de predições, que, através de uma função de agregação, realiza uma predição única para cada imagem original. Essas predições são utilizadas para treinar um novo modelo de aprendizado, chamado de modelo estudante.

Desse modo, os dados contidos em um conjunto candidato podem ser incorporados ao conjunto de treinamento original, sem que haja distorções. Os novos dados são copiados de maneira arbitrária, sem que haja a necessidade de intervenção (curadoria). Casos de falsos positivos ou falsos negativos são aceitos nesse momento, pois o modelo

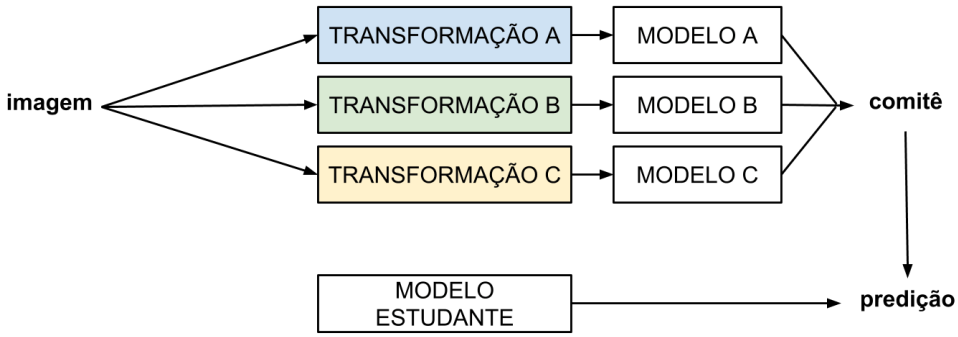


Figura 13 – Funcionamento da Destilação de Dados

tende a selecionar uma maior quantidade de dados bons que, que no final, trarão ganhos no desempenho dos modelos de aprendizado (RADOSAVOVIC et al., 2018).

3- Trabalhos Relacionados

Neste capítulo, são analisados alguns trabalhos relacionados com essa pesquisa. Esses trabalhos estão divididos de acordo com os aspectos explorados nessa dissertação. Na seção 3.1 descrevemos e comparamos trabalhos relacionados à Respostas a Perguntas Visuais, na Seção 3.2 descrevemos e comparamos trabalhos relacionados ao Aumento de Dados, na Seção 3.3 descrevemos e comparamos trabalhos relacionados ao Aprendizado Ativo e por fim na Seção 3.4 descreveremos trabalhos relacionados à Destilação de Dados.

3.1- Respostas a Perguntas Visuais

Existem muitos trabalhos que se propõem a criar conjuntos de dados e métodos para RPV. Dentre os trabalhos que propõem conjuntos de dados, podemos destacar o DAQUAR (MALINOWSKI; FRITZ, 2014) primeiro conjunto de dados projetado para RPV, porém esse conjunto de dados possui uma escala reduzida. Sendo o VQA (ANTOL et al., 2015) o primeiro conjunto de dados em grande escala para tarefas de RPV. Há também outros trabalhos com escopo mais específico, como o KB-VQA (P. WANG et al., 2015) projetado para utilizar bases de conhecimento externas, o FVQA (WANG et al., 2016) projetado para explorar características não visuais dos dados, o CLEVR (JOHNSON et al., 2017b) projetado para testar habilidades de raciocínio mais complexos dos modelos de aprendizado, o Visual Genome (KRISHNA et al., 2017) que se propõe a ser o maior conjunto de dados para tarefas de RPV. Ainda temos os conjuntos de dados balanceados que foram projetados para resolver os problemas de viés estatístico presente nos conjuntos de dados (P. ZHANG et al., 2016; GOYAL et al., 2017). Os principais conjuntos de dados criados para tarefas de RPV foram mostrados na Seção 2.4.

Em relação aos arquiteturas de métodos, podemos destacar JOHNSON et al. (2017b) com a abordagem de Redes Modulares e Mecanismo de Atenção. O trabalho de ANTOL et al. (2015), da *Universidade Virginia Tech*, e é um dos mais importante trabalhos

na área, pois além de propor um conjunto de dados, propôs uma arquitetura de RNAs que possuem um bom desempenho e serviu de base para outros trabalhos importantes como LU et al. (2016), YANG et al. (2016), P. ZHANG et al. (2016), Q. WU et al. (2016a) e Y. ZHU et al. (2016) e entre outros.

Ainda no escopo de RPV, podemos dividir em subproblemas como: perguntas de contagem numéricas, identificação de cores, localização de objetos e perguntas binárias. Uma pergunta Q é considerada binária quando sua resposta R pertence ao conjunto $\{$Sim, Não$\}$. Perguntas binárias foram abordadas nos trabalhos de P. ZHANG et al. (2016), onde os autores afirmam: ”Nós visualizamos a tarefa de responder perguntas binárias como uma tarefa de verificação visual”. GEMAN et al. (2015) utilizam abordagens de AD para gerar perguntas binárias sobre as imagens.

Nesta dissertação serão utilizadas em conjunto abordagens de AD, para gerar perguntas binárias adicionais para o conjunto de dados, e AA para selecionar os exemplos dentre os dados gerados de modo a obter melhor proveito das perguntas.

3.2- Aumento de Dados

O Aumento de Dados (AD) consiste em expandir os conjuntos de dados existentes através da manipulação dos dados existentes, para que mais exemplos de treinamento possam ser vistos e aprendidos pelo modelo. Em GEMAN et al. (2015), os autores criaram um sistema de AD capaz de gerar uma sequência de perguntas binárias que formam uma história sobre a cena presente na imagem. Os autores justificam que a composição de perguntas binária é capaz de capturar toda a informação presente na cena. Afirmam também que que gerar perguntas binárias é mais simples que as perguntas com respostas livres. Essa abordagem será utilizada nesta dissertação. Nesta abordagem, ao utilizar apenas um recorte da imagem, se o objeto alvo estiver encoberto ou parcialmente encoberto, pode gerar dificuldades interpretativas para o curador. Ao utilizar imagens inteiras não temos esse problema. Além de que utilizar novas imagens ao invés de reaproveitar recortes das mesmas já existentes no conjunto aumenta a variabilidade de exemplos de treinamento.

P. WANG et al. (2015) exploram o AD utilizando uma arquitetura de RNA conhecida

como Generative Adversarial Nets (GAN) (GOODFELLOW et al., 2014). O problema de utilizar diretamente os dados gerados desta forma está na ausência de rótulos aos exemplos gerados e na qualidade incerta desses exemplos. Em nosso trabalho, optamos por selecionar dados de outro conjunto, garantindo imagens de boa qualidade e com algum nível de rotulagem, para facilitar na escolha de melhores exemplos.

R. WU et al. (2015) demostra em seu trabalho que um aumento agressivo nos dados de treinamento podem reduzir drasticamente o sobre ajuste do modelo aos dados presentes no conjunto, porém o autor salienta que o poder computacional para processar o conjunto aumentado também se eleva consideravelmente. O autor contavam com um agrupamento de super computadores interligados por redes de dados de altíssima velocidade, esses recursos estão disponíveis apenas em grandes empresas de tecnologia e não são uma solução escalável em termos de custo de processamento. O autor selecionou os algorítimos de AD sem se importar com questões como custo e poder computacional necessário.

MCLAUGHLIN, RINCON e P. MILLER (2015) apresentam um novo método para AD baseado na troca de cena de fundo da imagem, como resultado os autores afirmam que há melhoria na capacidade de generalização do modelo de aprendizado. Porém este métodos possui complicações, uma vez que os autores utilizam imagens com apenas uma objeto em cena (da classes pessoa). Não há como saber como o método se comportará em casos de cenas mais complexas, contendo vários objetos.

Em relação a dados em formato de texto em linguagem natural, não é razoável gerar AD simplesmente embaralhando letras ou palavras, pois em um texto a ordem dos elementos importa. A idealmente o AD deveria vir da escrita de novas sentenças por humanos, mas isso é inviável devido a enorme quantidade de amostras. X. ZHANG e LECUN (2015) sugerem que as palavras do texto sejam substituídas por sinônimos. Essa abordagem geraria uma grande quantidade de novas perguntas, porém a variabilidade das perguntas seria baixa, pois semanticamente ainda seria o mesmo questionamento.

Enquanto DONG et al. (2017) apresentam um novo método baseado em GAN para gerar texto a partir das imagens, e o batizam de como *Image-Text-Image*. Esta abordagem possui o mesmo problema da proposta de (P. WANG et al., 2015), na qual a qualidade dos dados gerados pode ser duvidosa.

Por fim, VAN DYK e MENG (2001) discutem os principais métodos de AD em seu artigo de discussão *The Art of Data Augmentation*. Este trabalho é apenas um estudo

sobre o tema, embora muito completo, não apresenta nada de novo.

Neste trabalho, a tarefa de aumento de texto toma uma dimensão especial, dado que o texto das perguntas e respostas devem fazer sentindo em relação a imagem. Por isso, optamos por simplesmente copiar as perguntas e adaptar as repostas às novas imagens selecionadas. Enquanto os trabalhos anteriormente citados focam apenas no aumento de dados em texto ou em imagens isoladamente, mas nunca em conjunto muito menos no contexto de RPV.

3.3- Aprendizado Ativo

(KROGH; KROGH; VEDELSBY, 1995) propõem um esquema de AA utilizando comitês de predição. O método é basicamente uma generalização de métodos de seleção de dados. A ideia dos autores foi demonstrar o ganho ao utilizar dados não rotulados selecionadas através de um comitê. Os autores demonstram que há um ganho expressivo na capacidade de predição dos comitês em tarefas de classificação, enquanto o nosso trabalho foca na seleção de melhores exemplos para a tarefas de RPV.

C. ZHANG e T. CHEN (2002) apresenta um método onde o algoritmo mantém uma lista de atributos armazenada e a cada iteração é apresentado a um curador humano a imagem e os possíveis rótulos, para então o curador confirmar ou rejeitar estes rótulos. Em comparação com este método, nós mantemos uma lista com prováveis perguntas para uma nova imagem, e o curador escolhe assimilar aquela pergunta a nova imagem ou não.

Para HOI, JIN e LYU (2006) a categorização de texto em larga escala é um importante tópico de pesquisa, sobretudo para mineração de dados da Web. Um dos desafios de categorização de texto em grande escala é como reduzir os esforços humanos em rotular documentos de texto para construir uma classificação confiável modelos. Os autores apresentam um algoritmo de AA que seleciona lotes de documentos de texto para serem rotulados manualmente em cada iteração. Nós selecionamos lotes de imagens para serem anotadas com perguntas e respostas.

E SETTLES e CRAVEN (2008) conduzem um estudo empírico sobre as principais técnicas de AA para extração de informação e anotação de documentos de textos em

grande escala. Os autores criticam as estratégias para seleção de dados e propõem novas estratégias para lidar com as limitações desses métodos. Os autores utilizam estratégias similares as nossas, como por exemplo, seleção de exemplos por comitê. O trabalho dos autores foca na extração de informação contida em grandes quantidades de texto, enquanto nosso trabalho foca na seleção de melhores exemplos de imagens que encaixem com o texto da pergunta.

Segundo SETTLES (2012), o AA é um subcampo da AM, em que a hipótese principal é de que se o modelo puder escolher dados a partir dos quais ele aprende, seu desempenho será melhor, mesmo utilizando uma quantidade menor de dados. Este trabalho é uma síntese de estudos relativos a AA e não propõe nenhuma nova solução para o problema.

X. LIN e PARIKH (2017) apresentam um estudo empírico sobre AA para tarefas de RPV, na qual um modelo de aprendizado seleciona pares de perguntas e imagens informativos de um repositório dados e apresenta a um oráculo para melhorar o desempenho do modelo utilizando uma quantidade limitada de dados. Os autores demonstram que as três abordagens para AA (ver Seção 2.6) demonstram ganhos de desempenho nos modelos de aprendizado em relação aos modelos de aprendizado que foram treinados de maneira tradicional. Os autores também focam em perguntas binárias, em que afirmam ter conseguindo ganhos em acurácia. Em comparação com o nosso trabalho, X. LIN e PARIKH (2017) utilizam apenas técnicas de AA e tentam obter ganho utilizando apenas a parte do conjunto de dados que contribui efetivamente com o aprendizado do modelo, enquanto em nosso trabalho buscamos utilizar todo o conjunto de treinamento e ainda aumentá-lo com mais dados selecionados de outros conjuntos.

3.4- Destilação de dados

BUCILUA, CARUANA e NICULESCU-MIZIL (2006) apresentam um método para "comprimir" grandes conjuntos de modelos (comitês) em modelos menores e mais rápidos, geralmente sem perda significativa de desempenho. Essa compressão se dá pelo treinamento de RNAs pequenas que aprendem a imitar as funções do comitê. No nosso trabalho não chegamos a utilizar o método de compressão de modelos, embora reconhe-

cemos que há ganhos em velocidade e consumo de recursos com essa abordagem. Na Seção 6.2, descrevemos como pretendemos utilizar este método.

Seguindo uma linha investigativa similar à de BUCILUA, CARUANA e NICULESCU-MIZIL (2006), G. HINTON, VINYALS e DEAN (2015) sugerem uma variação da técnica com modelos comprimidos, nessa variação os autores utilizam modelos comprimidos em conjunto com um ou mais modelos completos, que são chamados de modelos especialistas. Esses autores experimentaram essa variação em conjuntos de imagens e de áudio (reconhecimento de voz e fala) obtendo ganhos em todos esses conjuntos. A utilização de mais modelos especialistas é encorajada, pois quanto maior a quantidade desses modelos, maior o ganho em contrapartida a necessidade de mais poder computacional para processar muitos modelos grandes. Assim como no método anterior, pretendemos estudar mais profundamente como utilizar estes métodos de compressão para ter otimização de processamento.

RADOSAVOVIC et al. (2018) comparam a destilação de modelos com destilação de dados, sendo esta primeira citada nos trabalhos acima, onde é formado um comitê de modelos de predição que juntos conseguem realizar uma predição melhor que um modelo sozinho não seria capaz de realizar. Já na destilação de dados, o comitê funciona como em um AA, selecionando dados candidatos e transferindo as anotações existentes para os novos dados candidatos. O comitê vota em quais dados devem receber as anotações existentes no conjunto de treinamento.

Nesta dissertação, utilizamos a DD para transferir as anotações contidas nos conjuntos de treinamento para os dados selecionados no conjunto candidato. Deste modo, é possível aproveitar todos os dados anotados de um conjunto para tirar proveito de uma quantidade muito grande de dados não anotados. Para os dados anotados utilizamos as imagens com perguntas e respostas do conjunto VQA, e para dados não anotados utilizamos as imagens do conjunto Imagenet (RUSSAKOVSKY et al., 2015).

4- Abordagens para Refinamento de Modelos de Respostas a Perguntas Binárias

A tarefa de RPV consiste em treinar um modelo preditivo que, dados como entrada uma imagem I e uma pergunta Q sobre I, possa produzir de maneira automática, uma resposta R coerente com Q. Para isso, o modelo precisa ser treinado para compreender os conceitos presentes na imagem e na pergunta. O estado da arte para esta tarefa consiste na utilização de redes neurais profundas, que são modelos de aprendizado de máquina que demandam muitos dados para seu treinamento. Coletar grandes quantidades de dados para treinamento dessas redes neurais é uma tarefa laboriosa. Adicionalmente, mesmo que haja dados disponíveis em grande quantidade, ainda é grande a probabilidade de haver escassez de exemplos para os conceitos mais raros no domínio de aplicação em questão. Aumentar a quantidade de dados é uma tarefa endereçada pelas técnicas de AD (Seção 3.2), enquanto que selecionar melhores exemplos para aumentar o conjunto de treinamento original é uma tarefa endereçada por técnicas de AA (Seção 2.6).

Nesta dissertação, partimos do pressuposto de que existe um conjunto de treinamento original $\mathcal{D}$ para treinamento de modelos RPV. Também supomos que está disponível uma coleção de imagens externa $\{I_e\}$. A partir disso, propomos duas abordagens que usam técnicas de AD e de AA para o realizar o aumento de $\mathcal{D}$ (usando elementos da coleção $\{I_e\}$) visando melhorar o desempenho do treinamento. Na primeira abordagem, desenvolvemos uma ferramenta de curadoria para direcionar a criação de novos exemplos a serem adicionados ao conjunto de treinamento originalmente existente. Na segunda abordagem, baseada em destilação de dados, propomos o uso de uma rede neural siamesa para produzir um comitê de modelos, com o propósito de selecionar de forma automática quais exemplos devem ser utilizados para aumentar o conjunto de treinamento $\mathcal{D}$.

Como restrição de escopo, as perguntas com respostas binárias foram escolhidas para avaliação de desempenho nesta dissertação. A razão desta restrição é que a avaliação de modelos para perguntas binárias é mais direta quando comparada às respostas de perguntas de escopo aberto que podem ser sentenças complexas.

Este capítulo está organizado conforme a seguir. Na Seção 4.1, apresentamos a ferramenta de curadoria criada para auxiliar na criação de novos exemplos de treinamento. Na Seção 4.2 apresentamos nossa proposta baseada em destilação de dados para anotação (rotulação) automática de novos exemplos.

4.1- Curadoria de Dados

Neste trabalho, foi desenvolvida uma ferramenta de software (na forma de uma aplicação WEB) para dar suporte aos seus usuários (que chamamos de curadores) na criação de novos exemplos para aumentar um conjunto de dados para RPV de forma colaborativa. Nessa ferramenta, os curadores são apresentados a uma imagem candidata I_c (que não faz parte do conjunto de treinamento original) identificada previamente como sendo similar a alguma imagem contida no conjunto de dados original. A ferramenta também seleciona uma pergunta candidata Q_c e solicita ao curador que indique se Q_c é adequada para formar uma nova tripla com a imagem I_c. Uma pergunta Q_c é considerada adequada se está relacionada com o conteúdo da imagem candidata I_c. Caso seja, a ferramenta solicita ao curador que indique sua resposta R_c.

Os processos de (i) identificação de imagens similares às já existentes no conjunto de dados original e de (ii) seleção das perguntas a apresentar: são realizados pela ferramenta de forma inteligente com o propósito de aproveitar ao máximo a participação do curador na construção do conjunto de dados estendido. Em particular, os módulos dessa ferramenta foram implementados com o uso de diferentes modelos de redes neurais para processamento dos dados visuais (imagens) e textuais (perguntas e respostas), com o propósito de selecionar as melhores consultas (i.e., um par I_c, Q_c) a serem apresentadas aos curadores. O resultado desse processo é uma nova coleção de triplas (Q_c, I_c, R_c) que são utilizadas para aumentar o conjunto de treinamento de um modelo RPV. O código da ferramenta é fornecido em domínio público[1].

A ferramenta é composta de quatro módulos, com as seguintes atribuições: filtragem de imagens, extração de termos, filtragem de perguntas e apresentação de consultas. A Figura 14 ilustra a arquitetura da ferramenta, com seus módulos e conjuntos de dados.

[1] Disponível em https://github.com/MLRG-CEFET-RJ/redes-siamesas-curadoria

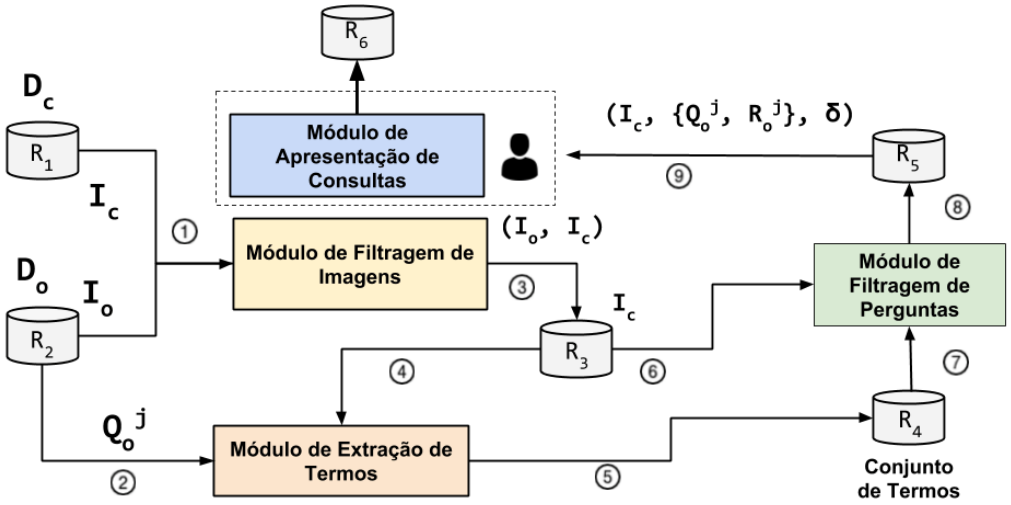


Figura 14 – Arquitetura da ferramenta de curadoria.

Módulo de filtragem de imagens

Esse módulo tem como entrada o conjunto de dados original $\mathcal{D}_o$ e o conjunto de imagens candidatas $\mathcal{D}_c$ (fluxo de dados 1). A responsabilidade desse módulo é identificar quais imagens em $\mathcal{D}_c$ (repositório R_1) são potencialmente úteis para compor novos exemplos. Inicialmente, são computadas as similaridades entre cada par de imagens em $\mathcal{D}_o \times \mathcal{D}_c$. Para isso, esse módulo possui internamente uma rede neural siamesa (BROMLEY et al., 1993), que recebe como entrada duas imagens e produz uma medida de similaridade σ entre as duas imagens, tal que $0 \leq \sigma \leq 1$. Para implementação desse módulo, essa rede siamesa foi treinada usando pares de imagens retirados de $\mathcal{D}_o \times \mathcal{D}_o$ (repositório R_2). O treinamento dessa rede foi realizado de tal forma que, quanto maior o valor de σ, mais similar é o par de imagens.

Os pares de imagens consideradas similares são armazenados (fluxo de dados 3) em um repositório (R_3) como um conjunto de tuplas $\{(I_o, I_c)\}$, em que $I_o \in \mathcal{D}_o$, $I_c \in \mathcal{D}_c$. Foi selecionado o limiar $\sigma_{\text{MIN}} = 0{,}5$ acima do qual duas imagens são consideradas similares. Desse modo, são selecionadas as entradas de $\mathcal{D}_c$ cuja similaridade com ao menos uma imagem de $\mathcal{D}_o$ seja maior do que σ_{MIN}. O valor $0{,}5$ foi escolhido de tal modo a (i) filtrar pares de imagens muito dissimilares, e (ii) permitir que imagens não tão similares (i.e., com σ próximo de $0{,}5$) pudessem passar para os próximos módulos. Os módulos seguintes da ferramenta de curadoria realizam computações adicionais para

confirmar quais imagens candidatas no repositório R_3 são relevantes para apresentação a um curador.

Módulo de extração de termos

Esse módulo realiza duas atividades. Na primeira delas, cada imagem candidata em R_3 (fluxo de dados 4) é submetida a um modelo pré-treinado de detecção de objetos[2], que utiliza uma rede neural de detecção denominada *Single Shot Detection* (W. LIU et al., 2015). Esse modelo rotula cada imagem candidata I_c com os termos correspondentes aos objetos detectados em I_c. A rede pode identificar 20 classes diferentes de objetos, contidas no conjunto VOC (RUSSAKOVSKY et al., 2015). Vamos denotar por W_c o conjunto de termos correspondentes à imagem candidata I_c.

Em sua segunda atividade, esse módulo toma como entrada os conteúdos textuais das perguntas associadas a cada imagem $I_o \in \mathcal{D}_o$ considerada similar à imagem candidata (fluxo de dados 2). Para cada imagem I_o, cada pergunta Q_o^j associada é submetida a um POS Tagger[3], que identifica as funções gramaticais de cada palavra componente de Q_o^j. Para cada pergunta Q_o^j, esse módulo armazena em um repositório (R_4) o conjunto de termos correspondentes aos substantivos encontrados em Q_o^j (fluxo de dados 5). Os termos correspondentes às demais classes gramaticais são ignorados. Vamos denotar por W_o^j o conjunto de termos correspondentes à j-ésima pergunta associada à imagem I_o.

A Figura 15 apresenta uma entrada do ImageNet na qual estão demarcados alguns objetos, que foram previamente detectados. Nessa imagem, foram detectados objetos das classes *Bike,Car e Dog*.

[2]Utilizamos o modelo pré-treinado fornecido em `https://github.com/balancap/SSD-Tensorflow`.
[3]Utilizamos o POS tagger fornecido pela ferramenta NLTK (`https://www.nltk.org/`).

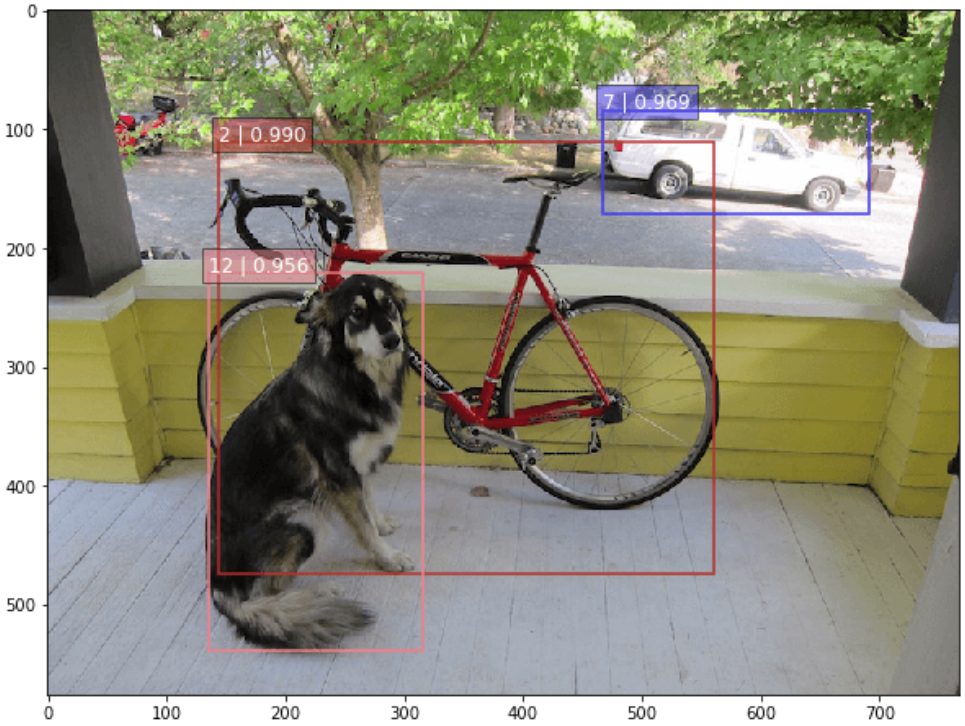


Figura 15 – Imagem do ImageNet na qual estão destacados alguns objetos detectados por uma rede *Single Shot Detection*.

Módulo de filtragem de perguntas

Nesse módulo, são tomados como entrada as imagens candidatas (fluxo de dados 6) e os conjuntos de termos W_c e W_o^j previamente computados e armazenados em R_4 (fluxo de dados 7). Em seguida, cada termo de W_c e W_o^j é fornecido a uma rede neural pré-treinada[4] de mapeamento de *word embeddings*. Dado um termo como entrada, essa rede neural produz a sua representação vetorial semântica correspondente. Esse modelo foi treinado com o algoritmo Word2Vec (MIKOLOV et al., 2013) para extração do vetor de cada termo.

Após o mapeamento de termos para vetores semânticos, esse módulo computa δ, a distância euclidiana média entre os vetores de W_c e os vetores de W_o^j. Se essa distância média for menor do que um limiar pré-estabelecido (que é um parâmetro de configuração da ferramenta), então a j-ésima pergunta é descartada do conjunto de potenciais perguntas a serem apresentadas ao curador. Esse processo é repetido para cada

[4] Utilizamos o modelo pré-treinado fornecido em `https://code.google.com/archive/p/word2vec/`

pergunta associada a cada imagem I_o em R_3. As perguntas remanescentes (juntamente com suas respostas correspondentes e as imagens candidatas) são armazenadas no repositório R_5 da ferramenta (fluxo de dados 8).

Módulo de apresentação de consultas

Temos como entrada desse módulo (fluxos de dados ⑨) um conjunto de tuplas $(I_c, \{Q_o^j, R_o^j\}, \delta)$ do repositório R_5, em que I_c é uma imagem candidata que foi considerada similar à imagem $I_o \in \mathcal{D}_o$, $\{Q_o^j, r_o^j\}$ é o subconjunto de pares pergunta/resposta associados a I_o remanescentes do procedimento de filtragem de perguntas (Seção 4.1), e δ é a distância média entre W_c e W_o^j.

Em uma consulta ao curador, esse módulo apresenta uma imagem I_c e cada elemento do conjunto $\{Q_o^j, R_o^j\}$. O curador deve então assinalar se a pergunta Q_o^j se aplica ou não à imagem I_c. No caso de se aplicar, o curador assinala uma resposta R_c^j, que pode ser ou não igual à resposta original R_o^j. Nesse momento, um novo exemplo da forma (I_c, Q_o^j, R_c^j) foi criado. Esse novo exemplo é armazenado para posteriormente ser incorporado ao conjunto de treinamento aumentado. As consultas são apresentadas ao curador na ordem crescente do valor de δ. A Figura 16 apresenta a tela principal de curadoria da ferramenta. A tela apresenta uma imagem candidata I_c, com uma pergunta selecionada pelo módulo de filtragem de perguntas Q_o. As triplas (I_c, Q_o, E_c) são armazenadas no conjunto de dados de treinamento aumentado R_6.

4.2- Procedimento de Destilação de Dados

Uma estratégia comum para melhorar a precisão de um modelo de reconhecimento visual é a aplicação de transformações sobre dados de entrada e depois agregar os resultados das variações geradas. Exemplos podem ser vistos no trabalho de R. WU et al. (2015). Com base nessa ideia, outra abordagem que propomos para aumentar o conjunto de treinamento original, desta vez de modo automático, é a definição de um algoritmo de

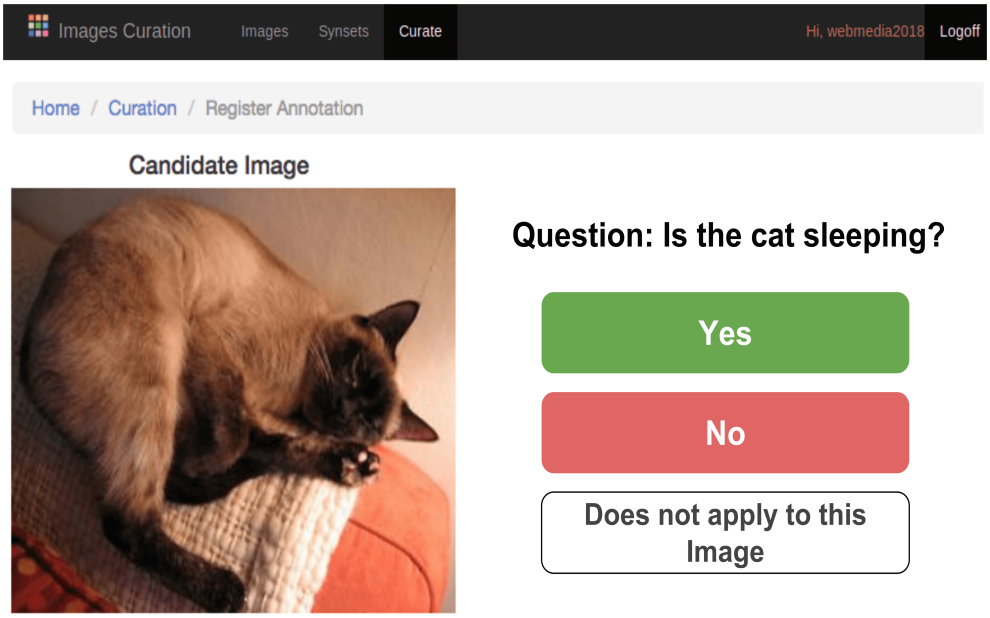


Figura 16 – É apresentada ao curador uma imagem do conjunto candidato e uma pergunta do conjunto original, então o curador pode assinalar as repostas ”Sim”ou ”Não”, ou mesmo descartar a imagem assinalando ”Não se aplica”.

destilação de dados (ver Seção 2.7).

Um procedimento de destilação de dados envolve quatro etapas: (1) treinar um modelo em dados anotados manualmente; (2) aplicar o modelo treinado a múltiplas variações de dados não anotados; (3) submeter as predições a um comitê de predição; e, (4) converter as predições sobre os dados não anotados em novas anotações (RADOSAVOVIC et al., 2018). Nas próximas seções, descrevemos de que forma nosso algoritmo de destilação de dados instancia essas etapas. Na Figura 17 é apresentado o esquema de destilação de dados utilizado nesse trabalho.

Treinamento de modelos usando dados já rotulados

Na nossa abordagem, em vez de aplicar transformações sobre o conjunto de treinamento inteiro, decidimos extrair n subconjuntos diferentes do conjunto de dados original $\mathcal{D}_o$. Cada subconjunto é formado por imagens selecionadas de maneira estratificada, de modo a manter as proporções iguais para exemplos positivos e negativos, e as mesmas proporções do número de objetos por classe.

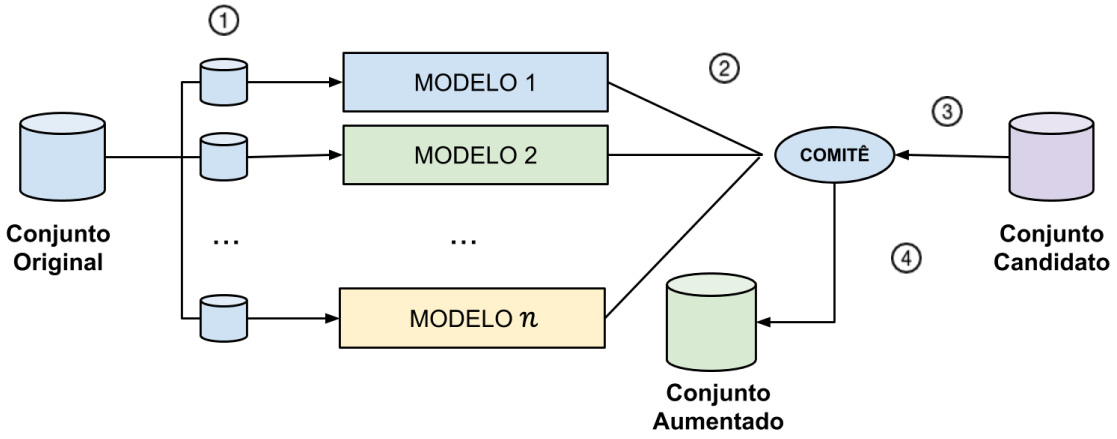


Figura 17 – Esquema de destilação de dados utilizado. Em ① são selecionados subconjuntos de dados do conjunto original para treinar modelos de aprendizado; em ② as predições do modelo são submetidas a um comitê de seleção que selecionará quais imagens do conjunto candidato possuem maior similaridade com imagens do conjunto original obtidas em ③; e em ④ as perguntas e respostas da imagem original são copiadas para as imagens candidatas e adicionadas ao conjunto aumentado.

Os n subconjuntos de tamanho igual são então utilizados para treinar n instâncias de modelos componentes baseados em RNS. O conjunto de treinamento necessário para treinar uma rede siamesa deve ser composto de exemplos que são triplas. Em cada tripla, há duas imagens e um bit que indica se essas imagens são similares ou não.

Para formar cada exemplo (tripla) dos n conjuntos para treinamento das redes siamesas, realizamos o procedimento a seguir. Inicialmente, selecionamos aleatoriamente duas imagens a partir de $\mathcal{D}_o$. Se as duas imagens selecionadas forem pertencentes às mesmas classes, esse bit é definido como igual a 1 (indicador de similaridade). Em caso contrário, é definido como igual a 0. Uma vez criados os n conjuntos de treinamento, n redes siamesas são então treinadas. Os n modelos resultantes são usados como componentes de um comitê para predição, conforme descrito na próxima seção.

Transformações nos dados de entrada

No nosso caso, o passo equivalente as transformações nos modelo de se dá pela seleção de subconjuntos de dados retirados do conjunto de dados original. Cada

subconjunto é utilizado para treinar um componente diferente do comitê de predição. Os comitês são a agregação das predições de múltiplos modelos componentes, onde é possível obter uma predição superior a qualquer uma das predições dos modelos individuais. Uma observação é que a predição agregada gera novo conhecimento e, em princípio, é possível usar esta informação para gerar novas anotações para os dados (RADOSAVOVIC et al., 2018). No nosso caso, usamos a informação gerada pelo comitê para criar novos exemplos rotulados para serem posteriormente adicionados ao conjunto $\mathcal{D}_o$.

Gerar anotações para os dados não rotulados

Utilizamos o comitê para determinar qual imagem I_c (proveniente de $\mathcal{D}_c$) possui a maior similaridade em relação à uma dada imagem I_o (proveniente de $\mathcal{D}_o$).

Repare que cada um dos n componentes do comitê é uma rede siamesa que, dado um par de imagens como entrada, produz um número (i.e., uma pontuação) entre 0 e 1 indicando a similaridade entre as duas imagens fornecidas. Cada par (I_c, I_o) é fornecido para os n componentes dos comitês, o que resulta em n valores (entre 0 e 1) para cada par. Nesse ponto, precisamos definir alguma forma de agregar a informação produzida pelos componentes do comitê, com o propósito de determinar qual é a imagem em $\mathcal{D}_c$ mais similar a uma dada imagem de $\mathcal{D}_o$. Dessa forma, propomos duas alternativas para agregação das pontuações produzidas pelos diversos componentes do comitê, a saber: computar a soma ou o valor máximo das pontuações produzidas pelos componentes. Na descrição fornecida a seguir para essas duas alternativas, considere que $s_i(I_o, I_c)$ é a pontuação (similaridade) computada para o par de imagens (I_o, I_c) pelo i-ésimo componente do comitê.

Na abordagem de soma, as pontuações s_i ($1 \leq i \leq n$) obtidas para um determinado par (I_o, I_c) a partir do i-ésimo componente do comitê são usadas para computar uma média , conforme a Equação 1.

$$M(I_o, I_c) = \frac{\sum_{i=1}^{n} s_i(I_o, I_c)}{n} \tag{1}$$

Já na abordagem de cômputo do valor máximo, o valor máximo das pontuações

dos n componente é computado, conforme a Equação 2.

$$M(I_o, I_c) = \max_i s_i(I_o, I_c) \tag{2}$$

Com a utilização dessas duas estratégias, conseguimos dois tipos diferentes de AD sobre os conjuntos de treinamento. Existem outras estratégias mais complexas para agregação de comitês que não foram abordadas neste trabalho. Ver Seção 6.2.

Gerar anotações para os dados não rotulados

Considere que o comitê determinou que a imagem I_c possui o maior valor de similaridade em relação a uma dada imagem I_o da tripla $(I_o, Q_o, R_o) \in \mathcal{D}_o$. Sendo assim, a imagem I_c é usada para é usada para forma uma nova tripla (I_c, Q_o, R_o), que será incorporado ao conjunto de treinamento aumentado automaticamente.

5- Experimentos

Neste capítulo, descrevemos o planejamento e a realização dos experimentos. Na Seção5.1 são descritos os conjuntos de dados escolhidos para os experimentos e quais as motivações para a escolha desses conjuntos. Na Seção 5.2 são apresentados os experimentos realizados. Por fim, na Seção 5.3 apresentamos os resultados e as respectivas análises.

5.1- Conjuntos de Dados

Para a realização dos experimentos, utilizamos o conjunto de treinamento *VQA-Real* (ANTOL et al., 2015) como conjunto de dados original D_o, pois é nesse conjunto de dados que os modelos de RPV escolhidos são treinados. O conjunto *VQA-Real* em sua versão V1, utilizada neste trabalho, possui 82.783 imagens de treinamento, distribuídas em 90 classes diferentes. Todas as imagens e classes são provenientes do conjunto de dados para VC conhecido como COCO (*Common Objects in Context*) (T.-Y. LIN et al., 2014). O conjunto *VQA-Real* ainda conta com 248.349 perguntas e respostas (rótulos) para treinamento. Cada imagem do conjunto *VQA-Real* possui exatamente três perguntas e suas respostas. Todas as perguntas do conjunto possuem duas versões, a primeira versão contém a pergunta e nenhuma resposta, e chamamos esta de perguntas com *resposta livre*, pois o modelo gera a reposta como bem entender; e a segunda, chamada de perguntas de *múltipla escolha*, o modelo escolhe de um conjunto de possíveis respostas qual é a resposta mais apropriada para a pergunta e imagem. O modelo é treinado para responder perguntas das duas versões.

Como conjunto de dados candidato D_c, escolhemos o *Imagenet* (RUSSAKOVSKY et al., 2015). O conjunto de dados *Imagenet* possui $14.197.122$ de imagens distribuídas em 21.841 classes (ou subclasses) que seguem a hierarquia do *WordNet* (G. A. MILLER, 1995). Somente as classes mais populares do *Imagenet* possuem algum tipo de rótulo, quase sempre etiquetas textuais sobre quais objetos estão presentes em cena.

Para a experimentação selecionamos 68.568 imagens do *Imagenet* para instanciar $\mathcal{D}_c$ (66.302 da hierarquia de classes *Dog* e 2.266 da hierarquia de *Cat*). Além disso, selecionamos 4.973 imagens (juntamente com os pares pergunta/resposta associados) para instanciar $\mathcal{D}_o$ (2.761 da classe *Dog* e 2.212 da classe *Cat*)

Apenas essas classes de objetos foram selecionadas para a experimentação, pois instanciar todo o conjunto D_o e D_c, pois haverá bilhões de pares (Io, Ic) a serem processados, o que torna inviável à execução dos experimentos com os recursos disponíveis para esta pesquisa.

5.2- Fase de Experimentação

Infraestrutura

Os experimentos foram realizados utilizando códigos escritos na linguagem *Python* e utilizando os frameworks para construções de redes neurais Keras (CHOLLET et al., 2015), pyTorch e Tensorflow. O experimentos foram executados em servidores disponibilizados pelo CEFET/RJ para trabalhos de pesquisa feitos por professores e alunos. Cada servidor conta cona processadores *Intel™ Core i7*, 32 gigabytes de memória RAM, discos SSD de 256 gigabytes, discos magnéticos de 2 terabytes e um par de placas gráficas *NVidia™ GTX 1080 TI* com 12 gigabytes de memória.

Curadoria de dados

Abaixo apresentamos alguns resultados do funcionamento de nossa ferramenta no contexto dessa instanciação:

- Quantidade de pares (I_o, I_c) fornecidos na entrada do módulo de filtragem de imagens (Seção 4.1): $340.988.664$;

- Quantidade de pares (I_o, I_c) remanescentes do processo de filtragem de imagens (Seção 4.1): $11.695.923$;
- Quantidade de termos (substantivos) extraídos (Seção 4.1): 356.004; média de substantivos por pergunta: $1,7126$; média de termos por imagem: $2,5686$;
- Quantidade de pares pergunta/imagem resultantes da filtragem de perguntas (Seção 4.1): 259.716;
- Quantidade de curadores 16;
- Quantidade de triplas (I_c, Q_o, R_c) curadas: 18.047, média de 1.127 triplas por curador;
- 10.070 triplas (I_c, Q_o, R_c) válidos, sendo 6.737 perguntas assinaladas com a resposta "não", 3.333 com a resposta "sim";
- 7.977 perguntas assinaladas como "Não se aplica"(descartadas);
- 884 questões únicas foram adicionadas ao conjunto de treinamento;
- 883 novas imagens I_c selecionadas do conjunto externo;
- Tempo médio que um curador leva para responder a uma pergunta: 12s, resultando em 60h9m de curadoria, média de 3h46m por curador;

Repare que, com a instanciação realizada, são fornecidos aproximadamente 341 milhões de pares de imagens para o primeiro módulo. Após o processamento realizado pelos demais módulos, a lista de pares imagem/pergunta contém 259.716 itens, o que corresponde a uma redução de aproximadamente 1.300 mil vezes na quantidade de consultas a serem apresentadas na curadoria.

O esforço para a realização da curadoria é muito grande, e mesmo contando com 12 curadores voluntários, e não haveria tempo para terminar de curar as imagens e perguntas antes da conclusão deste trabalho. Como alternativas para esta situação poderíamos aumentar a quantidade de curadores utilizando a ferramenta de curadoria, ou reduzir a quantidade de pares a serem curados para o experimento. Seguimos pelo caminho de reduzir a quantidade de pares a serem curados, selecionando um subconjunto estratificado de pares para a curadoria; esse subconjunto contêm cerca de 26 mil pares selecionados de modo a ter a maior variedade de pares e classes de objetos.

Nessa estratificação, selecionamos imagens do subconjunto de treinamento, mantendo a proporção original entre as classes de objetos presentes na imagens.

Como curadores utilizamos alunos de iniciação científica ou dos cursos de pós-graduação em ciência da computação do CEFET/RJ. Todos os curadores foram voluntários do trabalho deste projeto.

Destilação de dados

Para o algoritmo de *Destilação de Dados* optamos por criar um comitê de predição contendo três componentes, esses componentes uma vez treinados em D_o, de maneira estratificada, onde subconjuntos de dados foram retirados do conjunto original, respeitando a proporção entre as classes de objetos presentes nas imagens. A sobreposição dos exemplos entre os subconjuntos é nula para o caso de três componentes, mas ela existirá e irá aumentar caso mais componentes forem treinados.

Cada componente (modelo treinado) do comitê levou cerca de 28 horas para ser treinado e a predição sobre os dados não rotulados levou cerca de 12 dias para cada componente.

Como estratégia de agregação de predições dadas pelos componente do comitê, optamos por duas abordagens simples, a primeira visa selecionar quais dados do conjunto candidato obtiveram a melhor média de pontuação entre as predições dadas pelos componentes do comitê. Com essa abordagem foram selecionadas 2.459 novas imagens para serem adicionadas ao conjunto de treinamento para RPV, sendo que cada imagem recebeu as três perguntas da imagem do conjunto de treinamento original a qual ela foi considerada similar.

Como segunda abordagem, selecionamos quais dados do conjunto candidato obtiveram a maior pontuação entre as predições dadas pelos componentes do comitê. Com essa abordagem também foram selecionadas 2.459 novas imagens para serem adicionadas ao conjunto de treinamento para RPV, sendo que cada imagem recebeu as três perguntas da imagem do conjunto de treinamento original a qual ela foi considerada similar.

Com a realização da DD buscamos aumentar em algumas vezes o tamanho do

conjunto de dados de treinamento, porém a quantidade de pares necessários que precisariam ser avaliados tornou isto proibitivo, dado os recursos computacionais disponíveis. Por isso tivemos que instanciar uma quantidade menor de pares para esta tarefa, conseguindo um aumento de apenas $2,97\%$ no conjunto de dados. Abaixo apresentamos algumas estatísticas do algoritmo de *destilação de dados*.

Para selecionar quais pares devem ser avaliados pelo algoritmo de DD, utilizamos os mesmos módulos de filtragem de imagens e termos utilizados na ferramenta de curadoria. Assim conseguimos otimizar o processamento do algorítimo, descartando antecipadamente pares com baixa similaridade.

- Quantidade de predições sobre pares (I_o, I_c) geradas por cada componente do comitê: $11.695.923$, totalizando $35.087.769$ de predições;
- quantidade de imagens I_c adicionadas ao conjunto de testes: 2.459;
- quantidade de novas triplas (I_c, Q_o, R_o) adicionadas ao conjunto de treinamento: 7.377
- Quantidade de pares (I_o, I_c) utilizados para treinar cada componente do comitê: $1.560.490$
- Tempo médio para treinar cada componente do comitê: 28 horas
- Tempo médio para inferência de cada componente do comitê: 12 dias (279h30m)

5.3- Resultados

Nesta seção, descrevemos os resultados obtidos para os experimentos utilizando aumento de dados através da ferramenta de curadoria e do algoritmo de *destilação de dados*.

Curadoria de dados

Após a curadoria, cada par de imagens (I_o, I_c), torna-se uma tripla (I_c, Q_o, R_c), composta, por uma imagem candidata I_c, uma pergunta Q_o obtida do conjunto original, e um resposta R_c assinalada pelo curador. Realizamos os experimentos adicionando 2, 4, 6, 8 e 10 mil triplas ao conjunto de dados original D_o para treinamento do modelo de aprendizado e verificar se houve melhorias nas predições em cada faixa de aumento. Após o treinamento das redes neurais utilizando os novos de conjuntos dados aumentados, não obtivemos ganhos na acurácia da rede, como podemos observar na Tabela 2. Nesta tabela temos duas seções. A primeira seção contêm os resultados para os experimentos onde a resposta é livre, ou seja, a rede neural tem a liberdade de dar a resposta em texto livre, sem nenhum tipo de restrição. A segunda seção contêm os resultados para os experimentos utilizando as mesmas perguntas, porém é apresentado ao modelo de predição uma lista com múltiplas respostas, e o modelo deve selecionar a resposta mais adequada. Cada linha da tabela (outros, numérica, binária), representa o acerto do modelo perguntas com determinado tipo de respostas, enquanto acerto geral diz qual o percentual de acerto do modelo levando em consideração todas as perguntas do conjunto. Nas colunas (original, 2, 4, 6, 8 e 10 mil) temos quais conjuntos de dados foram utilizados para treinar o modelo, sendo o conjunto original e os conjuntos com aumento de dados em até 10 mil perguntas, seguido intervalos de aumento de 2 mil. Em negrito, está destacado o melhor resultado para cada tipo de resposta e para o acerto geral.

Tabela 2 – Resultados obtidos em tarefas de RPV após o aumento de dados através da ferra- menta de curadoria.

	Original	2 Mil	4 Mil	6 Mil	8 Mil	10 Mil
Resposta Livre						
Outros	40,30	40,20	40,21	**40,38**	40,35	40,33
Numérica	32,89	33,19	33,00	**33,20**	33,19	32,97
Binária	79,88	79,77	**79,91**	79,56	79,73	79,65
Acerto Geral	54,18	54,13	54,16	54,15	**54,19**	54,12
Resposta Múltipla Escolha						
Outros	50,05	50,01	49,97	**50,20**	50,06	50,04
Numérica	34,08	34,34	34,29	34,40	**34,48**	34.19
Binária	79,92	79,80	**79,95**	79,60	79.76	79.68
Acerto Geral	59,22	59,19	**59,23**	59,22	59.22	59.14

Podemos observar que não houve melhoria preditiva do ponto de vista estatístico para nenhum dos conjuntos de dados aumentados pela ferramenta de curadoria, e nem ao menos houve melhoria na acurácia das perguntas binárias. De um conjunto de dados de treinamento aumentado para o outro houve apenas pequenas oscilações, nem sempre positivas.

O aumento de 10 mil triplas (I_c, Q_o, R_c) representa uma quantidade muito pequena do total de perguntas do conjunto de treinamento original que possui 248.349 triplas (I_o, Q_o, R_o). Por isso não conseguimos tirar conclusões sobre os números apresentados.

Como todas as triplas adicionadas ao conjunto de dados de treinamento D_o, foram todas pertencentes as classes *"cão"* e *"gato"*, fizemos então uma avaliação dos resultados levando em consideração apenas essas duas classes. Como podemos ver na Tabela 3. Nesta tabela, temos os resultados de acerto geral (acurácia sobre todas as perguntas) do modelo levando em consideração perguntas com respostas livres, e com respostas múltipla escolha, quando o modelo selecionar a resposta mais adequada dada uma lista de opções.

Tabela 3 – Resultados obtidos pelo aumento de dados utilizando a ferramenta de curadoria para as tarefas de RPV considerando apenas as imagens, perguntas e respostas das classes de *"cão"* e *"gato"*.

Resultados para as classes de *"cão"* e *"gato"*						
	Original	2 Mil	4 Mil	6 Mil	8 Mil	10 Mil
Resposta Livre	**54,02**	53,98	52,64	53,75	53,67	53,64
Resposta Múltipla Escolha	**60,58**	60,39	59,32	60,57	60,18	60,19

Como podemos observar na Tabela 3, houve perda de acurácia em todos os conjuntos de dados de treinamento aumentados. As perdas são muito pequenas e não podemos tirar conclusões ainda. Porém acreditamos que isso tenha ocorrido por conta de um enviesamento que possa ter sido inserido no conjunto, por conta de aumentar a proporção da quantidade de triplas dessas classes. No conjunto de treinamento original D_o, existem 2.459 imagens das classes *cão e gato*, que formam e 7.377 triplas (imagem, pergunta e resposta). Após o aumento de dados no conjunto através ferramenta de curadoria, o número de imagens das classes *cão e gato* aumentou para um total de 3.432 imagens e 17.377 triplas (imagens, perguntas e respostas).

Destilação de Dados

Como resultado da execução o algoritmo de destilação de dados, para cada par de imagens (I_o, I_c), foram geradas três triplas (I_c, Q_o, R_o), compostas por uma imagem candidata I_c, uma pergunta Q_o obtida do conjunto original, e um resposta R_o também obtida do conjunto original. Cada imagem I_o possui três perguntas Q_o e três respostas R_o associadas. Essas perguntas e respostas são copiadas para a imagem IC e então as novas triplas são incorporadas ao conjunto de dados de treinamento aumentado. Como consequência, temos uma nova imagem Ic para cada imagem I_o. Como entrada do algoritmo temos 4.259 imagens do conjunto de treinamento original D_o, então, obtivemos 4.259 novas imagens. Após o treinamento das redes neurais utilizando os novos conjuntos dados aumentados, não obtivemos ganhos na acurácia da rede, como podemos observar na Tabela 4.

Tabela 4 – Resultados obtidos em tarefas de RPV após o aumento de dados através do algoritmo de Destilação de Dados

	Original	Média	Máximo
Resposta Livre			
Outros	**40,30**	40,13	40,10
Numérica	32,89	**33,36**	33,09
Binária	**79,88**	79,85	79,58
Acerto Geral	**54,18**	54,15	54,00
Resposta Múltipla Escolha			
Outros	**50,05**	49,78	49,80
Numérica	34,08	**34,63**	34,32
Binária	**79,92**	79,88	79,61
Acerto Geral	**59,22**	59,14	59,02

Observarmos nestes resultados que não houve nenhuma melhoria do ponto de vista estatístico para nenhum dos conjuntos de dados aumentados pelo algoritmo, também não houve melhoria na acurácia das perguntas binárias.

Apesar de não poder tirar conclusões com os resultados obtidos pelos experimentos, observamos ainda, que os conjuntos de dados obtiveram um desempenho pior que o conjunto original, uma causa possível para esse quadro de aparente piora no desempenho pode se dar ao fato de o algoritmo de DD selecionar pares de maneira automática, sem intervenção de um curador, podendo ocasionar a seleção de pares de qualidade inferior aos da ferramenta de curadoria, e introduzir mais erros ao conjunto de

treinamento original.

Como todas as triplas adicionadas ao conjunto de dados de treinamento D_o, foram todas pertencentes as classes *"cão"* e *"gato"*, fizemos então uma avaliação dos resultados levando em consideração apenas essas duas classes. Como podemos ver na Tabela 5.

Tabela 5 – Resultados obtidos pelo aumento de dados utilizando o algoritmo de *destilação de dados* para as tarefas de RPV considerando apenas as imagens, perguntas e respostas das classes de *"cão"* e *"gato"*.

Resultados para as classes de *"cão"* e *"gato"*			
	Original	Média	Máximo
Resposta Livre	**54,02**	52,60	52,48
Resposta Múltipla Escolha	**60,58**	59,53	59,44

Como podemos observar na Tabela 5, houve perda de acurácia em todos os conjuntos de dados de treinamento aumentados. As perdas são muito pequenas e não podemos tirar conclusões, porém, podemos imaginar que a piora se deve ao fato de que o algorítimo não selecionou pares de qualidade muito boa.

Por fim, na tabela 6 colocamos todos os resultados lado a lado para se ter uma visão geral dos experimentos.

Tabela 6 – Todos os resultados com conjuntos aumentados pela ferramenta de curadoria e algoritmo de destilação de dados

	Ferramenta de Curadoria						**Destilação de Dados**	
	Original	2 Mil	4 Mil	6 Mil	8 Mil	10 Mil	Média	Máximo
Resposta Livre								
Outros	40,30	40,20	40,21	40,38	**40,35**	40,33	40,13	40,10
Numérica	32,89	33,19	33,00	**33,20**	33,19	32,97	33,36	33,09
Binária	79,88	79,77	**79,91**	79,56	79,73	79,65	79,85	79,58
Acerto Geral	54,18	54,13	54,16	54,15	**54,19**	54,12	54,15	54,00
Resposta Múltipla Escolha								
Outros	50,05	50,01	49,97	**50,20**	50,06	50,04	49,78	49,80
Numérica	34,08	34,34	34,29	34,40	**34.48**	34.19	34,63	34,32
Binária	79,92	79,80	**79,95**	79,60	79.76	79.68	79,88	79,61
Acerto Geral	59,22	59,19	**59,23**	59,22	59.22	59.14	59,14	59,02

6- Conclusões

6.1- Análise Retrospectiva

Ainda não há uma quantidade significativa de trabalhos na literatura que exploram a AA e o AD em conjunto, sobretudo no contexto da RPV. Nesta dissertação, procuramos investigar a utilização dessas técnicas em conjunto esperando obter ganhos nos modelos de predição para RPV.

Uma das propostas apresentadas nesta dissertação foi uma ferramenta de curadoria de dados que forma utilizados no treinamento de modelos preditivos para tarefas de RPV. A ferramenta compõe uma solução completa de curadoria, compreendendo desde a etapa de aquisição de imagens candidatas, por meio de uma coleção externa de imagens, extração de características relevantes tanto das imagens quanto do conteúdo textual de perguntas, filtragem de entradas não relevantes, além da própria interface gráfica para interação com o curador. Por fim, o código da ferramenta proposta é fornecido em domínio público, e pode ser usado para adaptação a outras fontes de dados.

Inicialmente, acreditávamos que somente o aumento de dados por meio do sistema de curadoria seria o suficiente para obter os a melhora desejada. Após a implementação da primeira versão do sistema de curadoria, com a arquitetura mostrada na Figura 18, que contava apenas com os módulos de *filtragem de imagens* e *apresentação de consultas*, percebemos que ainda havia muitos ajustes a serem realizados.

Existia uma grande quantidade de dados a serem processados. Levando em conta o conjunto de treinamento original com 82.783 imagens multiplicado pelo conjunto de dados externo que possui $14.197.122$, haveria mais de 1 trilhão e 175 bilhões de pares a serem avaliados pelo módulo de filtragem antes de serem apresentados aos curadores.

Visando reduzir a quantidade de dados processados, a fim de conseguir ter tempo hábil para realizar os experimentos, optamos por processar um subconjunto de dados que utiliza apenas imagens pertencentes à determinadas classes dentro desses conjuntos. Escolhemos duas classes, a saber: *“cão”* e *“gato”*. Essas classes existem nos dois conjuntos de dados que usamos em nosso estudo (i.e., tanto em D_o quanto em D_c).

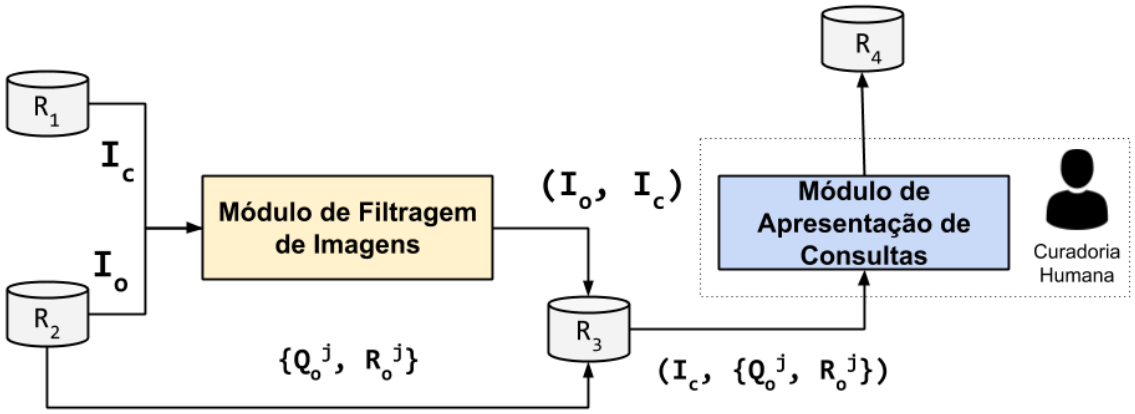


Figura 18 – Na primeira versão da ferramenta de curadoria havia apenas o módulo de filtragem de imagens, e todos os pares considerados similares por esse módulo, eram apresentados aos curadores.

Mesmo após a decisão por filtrar os dados apenas pelas duas classes mencionadas acima, ainda existia uma quantidade significativa de pares a serem curados. Considerando 4.973 imagens do conjunto de treinamento e 68.568 imagens do conjunto externo, isso resulta em mais de 340 milhões de pares a serem avaliados. Após passarem pelo módulo de filtragem de imagens, restaram 11 milhões e 695 mil pares para serem curados.

Iniciamos o processo de curadoria e logo foi percebido que a maioria dos exemplos apresentados aos curadores eram descartados (i.e., marcados como *"não se aplica"*), pois as perguntas não tinham a ver com as imagens, apesar da similaridade encontrada pela rede neural entre as duas imagens, do conjunto de treinamento D_o e do conjunto externo D_c. Como exemplo, podemos citar as imagens abaixo 19, onde a pergunta faz menção à palavra *'Cat'*, mas não há nenhum gato na imagem da direita (proveniente do conjunto externo), embora ela seja similar a imagem da esquerda (proveniente do conjunto de treinamento original).

Para resolver o problema, adicionamos novos módulos de filtragem à ferramenta de curadoria, com o objetivo de melhorar a assertividade dos pares exibidos aos curadores e reduzir o descarte. Então acrescentamos os módulos de *extração de termos* e o *módulo de filtragem de perguntas*. Esses novos módulos tinham o objetivo de analisar a imagem e extrair termos textuais delas (i.e. rótulos) e comparar esses termos com o texto da pergunta. Após a aplicação da filtragem pelos dois novos módulos, restaram 259.713 pares imagem/pergunta para curar.

Figura 19 – Apesar das imagens possuírem alguma semelhança, como por exemplo a cor dos animais e o seu posicionamento, a pergunta em questão não faz sentido para a imagem candidata, pois o texto da pergunta faz menção a um objeto da classe “gato” e na imagem candidata temos um objeto da classe “cachorro”.

A curadoria ainda dependia de esforço humano. E com o passar do tempo não haviam pares curados o suficiente para realização dos experimentos computacionais: apenas 18 mil pares foram curados ao longo dos meses por cerca de 16 curadores diferentes. Decidimos então investigar uma abordagem mais automática para acelerar o aumento de dados. Foi então que chegamos a ideia de utilizar a técnica de destilação de dados.

Com essa nova abordagem, novos pares poderiam ser adicionados automaticamente ao conjunto de treinamento original, sem intervenção humana. Criamos e treinamos um comitê de predição de similaridade, formado por três componentes. A ideia era que, juntos, esses componentes formassem um comitê de votação, no qual a imagens do conjunto externo mais similares a alguma imagem de conjunto original seriam usadas para criar novos exemplos.

Mais uma vez, nos deparamos com um enorme volume de dados não rotulados. Devido às limitações de recursos computacionais de que dispúnhamos, optamos mais uma vez por utilizar apenas duas classes de objetos do conjunto. Ainda assim foram muitos pares a serem analisados, com o agravante de que cada par deveria ser submetido aos três componentes diferentes do comitê. Ao final, conseguimos adicionar apenas 4.793 imagens novas ao conjunto, cada uma contendo três anotações diferentes, copiadas do conjunto original.

Ainda em uma tentativa de acelerar o processo de análise das triplas pelo algoritmo de destilação de dados, fizemos a substituição das redes neurais responsáveis por

analisar as imagens. Anteriormente, utilizamos as RNAs com a arquitetura ResNet50 (HE et al., 2016), e passamos a utilizar uma rede muito mais leve e mais rápida, a MobileNet (HOWARD et al., 2017). Com a mudança, a inferência ficou cerca de 10 vezes mais rápida, indo de meses para dias, porém não foi suficiente para processar todos os dados que gostaríamos.

Diante de todo o trabalho que tivemos ao longo dessa dissertação, a quantidade de exemplos criados (sejam por meio da ferramenta de curadoria, seja por meio do procedimento de destilação de dados) foi bem abaixo do que foi inicialmente planejado e esperado. Porém, com as limitações de tempo e de recursos computacionais, não foi possível explorar mais abordagens para aumentar escalabilidade e agilidade do processo.

Não conseguimos observar nenhuma melhoria no desempenho dos modelos treinados utilizando os conjuntos aumentados em relação ao conjunto original. Acreditamos que a causa para isso é a quantidade muito pequena de exemplos novos criados em ambas as abordagens propostas nesta dissertação, uma vez que, para conseguir melhorias visíveis, seria necessário aumentar o conjunto em pelo menos uma ordem de grandeza (X. LIN; PARIKH, 2017). Todavia, acreditamos que, mantendo os modelos ainda mais tempo processando, será possível adicionar uma quantidade de exemplos suficientes para surtir efeito positivo no modelo RPV resultante. Apesar dos resultados, o trabalho abre caminho para que se consiga melhorias na acurácia de modelos de predição, apenas precisando de mais tempo para se obter melhores resultados. Durante o trabalho tivemos o aprendizado do que funciona e do que pode melhorar, tendo uma base que serve para trabalhos futuros promissores.

Vale salientar que o trabalho *A Crowdsourcing Tool for Data Augmentation in Visual Question Answering Tasks* (SILVA et al., 2018) foi publicado no Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)[1], realizado em Outubro de 2018, na cidade de Salvador, Bahia. Esse artigo descreve a proposta relativa à ferramenta de curadoria.

[1] https://webmedia.org.br/2018/

6.2- Trabalhos Futuros

Há várias extensões planejadas para continuidade do desenvolvimento dessa ferramenta de curadoria. Nesse trabalho, consideramos apenas perguntas binárias (i.e., aquelas cujas as respostas podem ser ($\{$*sim*, *não*$\}$). Pretendemos estender a ferramenta de curadoria para tirar proveito de todas as perguntas do conjunto de treinamento. Além de melhorias planejadas para reduzir o tempo médio que cada curador precisa para curar um par.

Outro ponto de melhoria está no módulo de *extração de termos* da ferramenta de curadoria, onde podemos optar pela utilização de uma modelo geração automática de legendas para imagens, criando uma base de comparação mais rica entre a cena da imagem candidata e o texto da pergunta. Outra possibilidade é a utilização de bancos de *captchas*.

A abordagem de destilação de dados exige um alto poder computacional para ser realizada, pois boas predições dependem de um conjunto grande de modelos treinados. Com o intuito de reduzir o tempo de execução necessário para inferir todos os resultados do comitê de predição, pretendemos utilizar técnicas para otimizar esses modelos, como a técnica compressão de modelos, onde o de conhecimento adquirido pelo treinamento de múltiplos modelos pode ser transferido para um único modelo de tamanho reduzido (BUCILUA; CARUANA; NICULESCU-MIZIL, 2006; G. HINTON; VINYALS; DEAN, 2015).

Além disso, com uma maior disponibilidade de tempo e de recursos computacionais para a realização dos experimentos, pretendemos estender as avaliações para todo o conjunto de dados de treinamento e usar uma parte maior do conjunto de dados externo, bem como testar a utilização de outros conjuntos externos, como alguns dos conjuntos citados na seção 2.4.1. Com isso, pretendemos obter um aumento ainda maior do conjunto de treinamento original.

Objetivamos ainda utilizar as mais de 8 mil triplas sinalizadas como “não se aplica” para retreinar os modelos de predição de similaridade com o objetivo de melhorar a qualidade dos pares exibidos aos curadores da ferramenta de curadoria.

Objetivamos também analisar e testar novas estratégias para agregação das predições do componentes do comitê, como *Boosting* (MELVILLE; MOONEY, 2003),

Random Forests (BREIMAN, 2001), *Bayesian averaging* (DOMINGOS, 2000), *Stacking* (WOLPERT, 1992) entre outros. Também queremos aumentar a quantidade de modelos componentes do comitê de predição.

Por último, uma crescente área da Inteligência Artificial, denominada *Explainable AI* (ADADI; BERRADA, 2018), busca por uma metodologia que explicite o procedimento de escolha realizado por algoritmos de aprendizado de máquina, para fazer com que os resultados produzidos sejam compreensíveis por seres humanos. Os modelos de aprendizado produzidos por RNAs que foram usadas amplamente nesta dissertação, muitas vezes, são chamados de *modelos caixas pretas*, pois não fornecem explicação para a tomada de decisão que realizam. Seria desejável que os modelos RPV fornecessem a explicação para suas escolhas das respostas que produzem. Sendo assim, outra via de continuidade deste trabalho é investigar formas de explicar os resultados produzidos pelos modelos de RPV.

Bibliografia

ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). **IEEE Access**, v. 6, p. 52138–52160, 2018. ISSN 2169-3536. DOI: 10.1109/ACCESS.2018.2870052.

ANDREAS, JACOB et al. Neural Module Networks. In: CVPR. Las Vegas, NV, USA: IEEE Computer Society, 2016. p. 39–48.

ANTOL, STANISLAW et al. VQA: Visual Question Answering. In: abs/1505.00468.

AUER, SÖREN et al. DBpedia: A Nucleus for a Web of Open Data. In: THE Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings. Busan, Korea: Springer, Berlin, Heidelberg, 2007. p. 722–735. DOI: 10.1007/978-3-540-76298-0_52.

BANKO, MICHELE et al. Open Information Extraction from the Web. In: IJCAI. Hyderabad, India: CEUR-WS.org, 2007. v. 7, p. 2670–2676.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation Learning: A Review and New Perspectives. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 8, p. 1798–1828, ago. 2013. ISSN 0162-8828. DOI: 10.1109/TPAMI.2013.50.

BOLLACKER, KURT et al. Freebase. In: PROCEEDINGS of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08. New York, New York, USA: ACM Press, 2008. p. 1247. ISBN 9781605581026. DOI: 10.1145/1376616.1376746.

BREIMAN, LEO. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BROMLEY, JANE et al. **Signature Verification Using a "Siamese"Time Delay Neural Network**. Denver, Colorado: Morgan Kaufmann Publishers Inc., 1993. p. 737–744. (NIPS'93).

BUCILUA, CRISTIAN; CARUANA, RICH; NICULESCU-MIZIL, ALEXANDRU. Model compression. In: ACM. PROCEEDINGS of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Chicago, Illinois: ACM, 2006. p. 535–541.

CARLSON, ANDREW et al. Toward an Architecture for Never-Ending Language Learning. In: AAAI. Atlanta, Georgia, USA: AAAI Press, 2010. v. 5, p. 3.

CHOLLET, FRANCOIS et al. **Keras**. [S.l.]: GitHub, 2015. https://github.com/fchollet/keras.

DOMINGOS, PEDRO. Bayesian averaging of classifiers and the overfitting problem. In: ICML. Haifa, Israel: Omnipress, 2000. v. 2000, p. 223–230.

DONAHUE, JEFFREY et al. Long-term recurrent convolutional networks for visual recognition and description. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE Computer Society, 2015. p. 2625–2634.

DONG, HAO et al. I2T2I: Learning Text to Image Synthesis with Textual Data Augmentation. **CoRR**, abs/1703.06676, 2017. arXiv: 1703.06676.

GAO, HAOYUAN et al. Are you talking to a machine? dataset and methods for multilingual image question. In: ADVANCES in Neural Information Processing Systems. Montreal, Quebec, Canada: CEUR-WS.org, 2015. p. 2296–2304.

GEMAN, DONALD et al. Visual Turing test for computer vision systems. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 112, n. 12, p. 3618–23, mar. 2015. ISSN 1091-6490. DOI: 10.1073/pnas.1422953112.

GOODFELLOW, IAN et al. Generative Adversarial Nets. In: GHAHRAMANI, Z. et al. (Ed.). **Advances in Neural Information Processing Systems 27**. Montreal, Canada: Curran Associates, Inc., 2014. p. 2672–2680.

GORDON, JONATHAN; VAN DURME, BENJAMIN. Reporting bias and knowledge acquisition. In: PROCEEDINGS of the 2013 workshop on Automated knowledge base construction - AKBC '13. New York, New York, USA: ACM Press, 2013. p. 25–30. ISBN 9781450324113. DOI: 10.1145/2509558.2509563.

GOYAL, YASH et al. Making the {V} in {VQA} Matter: Elevating the Role of Image Understanding in {V}isual {Q}uestion {A}nswering. In: CONFERENCE on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE Computer Society, 2017.

GRAVES, ALEX; MOHAMED, ABDEL-RAHMAN; HINTON, GEOFFREY E. Speech Recognition with Deep Recurrent Neural Networks. **CoRR**, abs/1303.5778, 2013. arXiv: 1303.5778.

HAUPTMANN, ALEXANDER G et al. Extreme video retrieval: joint maximization of human and computer performance. In: ACM. PROCEEDINGS of the 14th ACM international conference on Multimedia. Santa Barbara, CA, USA: ACM, 2006. p. 385–394. 433061. ISBN 1-59593-447-2.

HE, KAIMING et al. Deep residual learning for image recognition. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE Computer Society, 2016. p. 770–778.

HINTON, GEOFFREY; VINYALS, ORIOL; DEAN, JEFF. Distilling the knowledge in a neural network. **arXiv preprint arXiv:1503.02531**, 2015.

HOCHREITER, SEPP; SCHMIDHUBER, JÜRGEN. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HOI, STEVEN CH; JIN, RONG; LYU, MICHAEL R. Large-scale text categorization by batch mode active learning. In: ACM. PROCEEDINGS of the 15th international conference on World Wide Web. Edinburgh, Scotland: ACM, 2006. p. 633–642.

HOULSBY, NEIL et al. Bayesian Active Learning for Classification and Preference Learning. **arXiv preprint arXiv:1112.5745**, dez. 2011. arXiv: 1112.5745.

HOWARD, ANDREW G. et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. **CoRR**, abs/1704.04861, 2017.

JIANG, MIN et al. Improving machine vision via incorporating expectation-maximization into Deep Spatio-Temporal learning. In: 2014 International Joint Conference on Neural Networks (IJCNN). Beijing, China: IEEE, jul. 2014. p. 1804–1811. ISBN 978-1-4799-1484-5. DOI: 10.1109/IJCNN.2014.6889723.

JOHNSON, JUSTIN et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In: CVPR. Honolulu, HI, USA: IEEE Computer Society, 2017. ISBN 978-1-5386-0457-1.

JOHNSON, JUSTIN et al. Inferring and Executing Programs for Visual Reasoning. In: ICCV. Venice, Italy: IEEE Computer Society, 2017.

KIM, JIN-HWA et al. Multimodal residual learning for visual qa. In: ADVANCES in Neural Information Processing Systems. Barcelona, Spain: CEUR-WS.org, 2016. p. 361–369.

KRISHNA, RANJAY et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **International Journal of Computer Vision**, Springer, v. 123, n. 1, p. 32–73, 2017.

KRIZHEVSKY, ALEX; SUTSKEVER, ILYA; HINTON, GEOFFREY E. Imagenet classification with deep convolutional neural networks. In: ADVANCES in neural information processing systems. Lake Tahoe, Nevada, USA: Curran Associates Inc., 2012. p. 1097–1105.

KROGH, ANDERS; KROGH, ANDERS; VEDELSBY, JESPER. Neural Network Ensembles, Cross Validation, and Active Learning. **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS**, v. 7, p. 231–238, 1995.

KUMAR, ANKIT et al. Ask me anything: Dynamic memory networks for natural language processing. In: INTERNATIONAL Conference on Machine Learning. New York City, NY, USA: JMLR.org, 2016. p. 1378–1387.

LECUN, YANN; BENGIO, YOSHUA; HINTON, GEOFFREY E. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015. DOI: 10.1038/nature14539.

LECUN, YANN; KAVUKCUOGLU, KORAY; FARABET, CLÉMENT et al. Convolutional networks and applications in vision. In: ISCAS. Paris, France: IEEE, 2010. p. 253–256.

LIN, TSUNG-YI et al. Microsoft {COCO:} Common Objects in Context. **CoRR**, abs/1405.0, 2014.

LIN, XIAO; PARIKH, DEVI. Active Learning for Visual Question Answering: An Empirical Study. **arXiv preprint arXiv:1711.01732**, nov. 2017. arXiv: 1711.01732.

LIU, H; SINGH, P. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. **BT Technology Journal**, Kluwer Academic Publishers, v. 22, n. 4, p. 211–226, out. 2004. ISSN 1358-3948. DOI: 10.1023/B:BTTJ.0000047600.45421.6d.

LIU, WEI et al. SSD: Single Shot MultiBox Detector. **CoRR**, abs/1512.02325, 2015. arXiv: 1512.02325.

LIU, YING. Active learning with support vector machine applied to gene expression data for cancer classification. **Journal of chemical information and computer sciences**, ACS Publications, v. 44, n. 6, p. 1936–1941, 2004.

LU, JIASEN et al. **Hierarchical Question-image Co-attention for Visual Question Answering**. Barcelona, Spain: Curran Associates Inc., 2016. p. 289–297. (NIPS'16). ISBN 978-1-5108-3881-9.

MALINOWSKI, MATEUSZ; FRITZ, MARIO. A multi-world approach to question answering about real-world scenes based on uncertain input. In: ADVANCES in Neural Information Processing Systems 27. Long Beach, CA: Curran Associates, Inc., 2014. p. 1682–1690.

MAO, JUNHUA et al. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). San Diego, CA, USA, abs/1412.6632, 2015.

MCLAUGHLIN, N.; RINCON, J. M. DEL; MILLER, P. Data-augmentation for reducing dataset bias in person re-identification. In: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Karlsruhe, Germany: IEEE Computer Society, ago. 2015. p. 1–6. DOI: 10.1109/AVSS.2015.7301739.

MELVILLE, PREM; MOONEY, RAYMOND J. Constructing diverse classifier ensembles using artificial training examples. In: PROCEEDINGS of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico: Morgan Kaufmann Publishers Inc., 2003. v. 3, p. 505–510.

MIKOLOV, TOMAS et al. Efficient Estimation of Word Representations in Vector Space. **CoRR**, abs/1301.3781, 2013.

MILLER, GEORGE A. WordNet: A Lexical Database for English. **Commun. ACM**, ACM, New York, NY, USA, v. 38, n. 11, p. 39–41, nov. 1995. ISSN 0001-0782. DOI: 10.1145/219717.219748.

MINSKY, MARVIN. Steps toward Artificial Intelligence. **Proceedings of the IRE**, v. 49, n. 1, p. 8–30, jan. 1961. ISSN 0096-8390. DOI: 10.1109/JRPROC.1961.287775.

PAN, JEFF Z. Resource description framework. In: HANDBOOK on ontologies. [S.l.]: Springer, 2009. p. 71–90.

PEREZ, LUIS; WANG, JASON. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. **CoRR**, abs/1712.04621, 2017. arXiv: 1712.04621.

RADOSAVOVIC, ILIJA et al. Data distillation: Towards omni-supervised learning. In: IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE Computer Society, 2018. p. 4119–4128.

REN, MENGYE; KIROS, RYAN; ZEMEL, RICHARD. Image question answering: A visual semantic embedding model and a new dataset. **Proc. Advances in Neural Inf. Process. Syst**, v. 1, n. 2, p. 5, 2015.

RUSSAKOVSKY, OLGA et al. **ImageNet Large Scale Visual Recognition Challenge**. v. 115. [S.l.]: Springer US, 2015. p. 211–252. DOI: 10.1007/s11263-015-0816-y.

SCHMIDHUBER, JÜRGEN. Deep learning in neural networks: An overview. **Neural Networks**, v. 61, p. 85–117, 2015. ISSN 08936080. DOI: 10.1016/j.neunet.2014.09.003.

SENER, OZAN; SAVARESE, SILVIO. ACTIVE LEARNING FOR CONVOLUTIONAL NEURAL NETWORKS: ACORE-SET APPROACH. **stat**, v. 1050, p. 27, 2017.

SETTLES, BURR. Active learning. **Synthesis Lectures on Artificial Intelligence and Machine Learning**, Morgan & Claypool Publishers, v. 6, n. 1, p. 1–114, 2012.

SETTLES, BURR; CRAVEN, MARK. An analysis of active learning strategies for sequence labeling tasks. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the conference on empirical methods in natural language processing. Honolulu, Hawaii: Association for Computational Linguistics, 2008. p. 1070–1079.

SILVA, RAMON et al. A Crowdsourcing Tool for Data Augmentation in Visual Question Answering Tasks. In: DOI: 10.1145/3243082.

SUCHANEK, FABIAN M; KASNECI, GJERGJI; WEIKUM, GERHARD. Yago: A Core of Semantic Knowledge. In: 16TH International Conference on the World Wide Web. Banff, Alberta, Canada: ACM, 2007. p. 697–706.

SUTSKEVER, ILYA; MARTENS, JAMES; HINTON, GEOFFREY E. Generating text with recurrent neural networks. In: PROCEEDINGS of the 28th International Conference on Machine Learning (ICML-11). Bellevue, Washington, USA: Omnipress, 2011. p. 1017–1024.

TANDON, NIKET et al. WebChild. In: PROCEEDINGS of the 7th ACM international conference on Web search and data mining - WSDM '14. New York, New York, USA: ACM Press, 2014. p. 523–532. ISBN 9781450323512. DOI: 10.1145/2556195.2556245.

TUR, GOKHAN; HAKKANI-TÜR, DILEK; SCHAPIRE, ROBERT E. Combining active and semi-supervised learning for spoken language understanding. **Speech Communication**, Elsevier, v. 45, n. 2, p. 171–186, 2005.

VAN DYK, DAVID A; MENG, XIAO-LI. The Art of Data Augmentation. **Journal of Computational and Graphical Statistics**, v. 10, n. 1, p. 1–50, 2001. DOI: 10.1198/10618600152418584.

VINYALS, ORIOL et al. Show and tell: A neural image caption generator. In: PROCEEDINGS of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE Computer Society, 2015. 07-12-June. ISBN 9781467369640. DOI: 10.1109/CVPR.2015.7298935. eprint: 1411.4555.

WANG, PENG et al. Explicit Knowledge-based Reasoning for Visual Question Answering. **CoRR**, abs/1511.02570, 2015. arXiv: 1511.02570.

WANG, Peng et al. FVQA: Fact-based Visual Question Answering. **CoRR**, abs/1606.05433, 2016. arXiv: 1606.05433.

WILLIAMS, DRGHR; HINTON, G E. Learning representations by back-propagating errors. **Nature**, v. 323, p. 533–536, 1986.

WOLPERT, DAVID H. Stacked generalization. **Neural networks**, Elsevier, v. 5, n. 2, p. 241–259, 1992.

WU, QI et al. Ask me anything: Free-form visual question answering based on knowledge from external sources. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE Computer Society, 2016. p. 4622–4630.

WU, QI et al. Visual Question Answering: A Survey of Methods and Datasets. **CoRR**, abs/1607.05910, 2016. eprint: 1607.05910.

WU, REN et al. Deep Image: Scaling up Image Recognition. **CoRR**, abs/1501.02876, 2015. arXiv: 1501.02876.

XU, KELVIN et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: PROCEEDINGS of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015. v. 37. (Proceedings of Machine Learning Research), p. 2048–2057.

YANG, ZICHAO et al. Stacked attention networks for image question answering. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE Computer Society, 2016. p. 21–29.

YU, LICHENG et al. Visual madlibs: Fill in the blank description generation and question answering. In: PROCEEDINGS of the 2015 IEEE International Conference on Computer Vision (ICCV). Washington, DC, USA: IEEE Computer Society, 2015. (ICCV '15), p. 2461–2469. ISBN 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.283.

ZHANG, CHA; CHEN, TSUHAN. An active learning framework for content-based information retrieval. **IEEE transactions on multimedia**, IEEE, v. 4, n. 2, p. 260–268, 2002.

ZHANG, P. et al. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. In: ISBN 978-1-4673-8851-1.

ZHANG, XIANG; LECUN, YANN. Text Understanding from Scratch. **CoRR**, abs/1502.01710, 2015. arXiv: 1502.01710.

ZHU, XIAOJIN; LAFFERTY, JOHN; GHAHRAMANI, ZOUBIN. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining. Washington, DC, USA: AAAI Press, 2003. v. 3.

ZHU, YUKE et al. Visual7w: Grounded question answering in images. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE Computer Society, 2016. p. 4995–5004.