1.03.04 - Ciência da Computação / Sistemas de Computação

#PRACEGOVER: ÁUDIO-DESCRIÇÃO AUTOMÁTICA DE IMAGENS

Gabriel O. dos Santos[1]*, Sandra Avila[2], Esther L. Colombini[3]
1. Estudante do Instituto de Computação da UNICAMP (IC-UNICAMP)
2. Professora do IC-UNICAMP - Departamento de Sistemas de Informação/Orientadora
3. Professora do IC-UNICAMP - Departamento de Sistemas de Informação/Co-Orientadora

Resumo

Pessoas com deficiência visual enfrentam muitos problemas relacionados à acessibilidade da Internet, pois grande parte do conteúdo publicado nesse meio é ainda exclusivamente visual. Assim, gerar descrições do conteúdo de imagens de forma automática é essencial para a inclusão desse público. Essa tarefa é um desafio conhecido na literatura como *image captioning*. A maior parte dos trabalhos sobre esse problema lida com a geração de descrições em inglês devido ao grande volume de dados anotados disponível, enquanto que são raras as bases com dados anotados em outros idiomas. Tendo esse cenário em vista, nós utilizamos os dados produzidos a partir da iniciativa PraCegoVer [6], e criamos a primeira base de dados com descrições em português. Atualmente, a base conta com mais de 550 mil imagens anotadas com as áudio-descrições criadas pelos adeptos desse movimento. Além da base de dados, nós também propusemos um *framework* para a coleta, tratamento e análise de dados do Instagram.

Palavras-chave: Image Captioning; Base de Dados; Descrições em Português

Apoio financeiro: FAPESP

Trabalho selecionado para a JNIC: UNICAMP

Introdução

Descrever o conteúdo de imagens de forma automática utilizando linguagem natural é uma tarefa essencial para a inclusão de pessoas com deficiência visual na Internet. Esse é ainda um grande desafio, conhecido na literatura como *image captioning*, que requer entender a relação semântica entre os objetos, seus atributos e ações. Nos últimos anos, a literatura avançou significativamente graças ao grande volume de dados anotados [1-5]. Entretanto, a maioria das bases de dados disponíveis contém múltiplas descrições de referência escritas em inglês, sendo escassas aquelas com textos em outros idiomas. Além disso, as descrições presentes nesses conjuntos têm, em média, 10 palavras com uma baixa variância em relação a esse tamanho, como é o caso da base de dados mais famosa dessa literatura, a MS COCO [2].

Assim, inspirado pela iniciativa PraCegoVer [6], que visa estimular a publicação de imagens acompanhadas de áudio-descrições nas redes sociais, nós utilizamos o conteúdo publicado no Instagram para criar a base de dados *#PraCegoVer*. Esse é o primeiro conjunto de dados proposto para *image captioning* com descrições em português. A base contém cerca de 550 mil pares de imagem-descrição, sendo que as descrições têm em média 40 palavras, com uma grande variação de tamanho. A falta de múltiplas anotações e o aumento do número médio de palavras bem como da variância tornam a nossa base ainda mais desafiadora que as outras anteriormente propostas. Para validar nossa base, nós treinamos modelos com a arquitetura estado-da-arte AoANet [7] e comparamos com a MS COCO. Assim, demonstramos experimentalmente que os modelos que normalmente são bem sucedidos em outras bases, repetem as mesmas palavras várias vezes dentro de uma sentença, o que ilustra a dificuldade da base de dados *#PraCegoVer*.

Em resumo, nossas principais contribuições são: a criação de um *framework* para a coleta e pré-processamento de publicações no Instagram a partir de uma *hashtag*; a proposta da base de dados *#PraCegoVer*; a demonstração de que algoritmos estado-da-arte ainda não resolvem os problemas introduzidos em nossa base, apresentando resultado inferior quando comparado a MS COCO.

Metodologia

Este trabalho consiste na criação do *dataset #PraCegoVer*. A Figura 1 mostra uma visão geral do procedimento adotado para a construção da nossa base de dados. Inicialmente, nós coletamos as publicações associadas à marcação #PraCegoVer. Para isso, desenvolvemos um *web crawler* que é executado diariamente e visita as páginas do Instagram em busca de publicações que marcaram #PraCegoVer. Em seguida, tratamos os textos dessas publicações a fim de extrair apenas os trechos com as áudio-descrições, eliminando, portanto, *emoticons*, *hashtags*, *links*, marcações de perfis, e quaisquer textos adicionais. Posteriormente, convertemos os textos para minúsculo, removemos palavras irrelevantes (*stopwords*) e os transformamos em vetores TF-IDF (*Term Frequency-Inverse Document Frequency*). Ainda, extraímos vetores de características das imagens utilizando uma rede chamada MobileNetV2 [8], e reduzimos a dimensionalidade dos vetores gerados de 1280 para 900 dimensões com o algoritmo UMAP (*Uniform Manifold Approximation and Projection for Dimension Reduction*) [9] para ocupar menos espaço em memória. Então, usamos o

algoritmo HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) [10] para agrupar as publicações, de modo que aquelas cujas imagens têm conteúdo similar ficassem dentro do mesmo grupo. Como é comum re-publicações, onde apenas alguns detalhes da imagem ou texto são alterados como destacado na Figura 2, e esses exemplos podem causar o *overfitting* dos modelos, construímos um algoritmo que detecta e elimina essas duplicações. Esse algoritmo tem como entrada os vetores TF-IDF dos textos e os vetores de características das imagens com dimensão reduzida. A partir desses dados, duas matrizes de similaridade são constituídas, sendo uma para os textos e outra para as imagens. Um grafo de similaridade, que mantém em uma mesma componente conexa publicações duplicadas, é então extraído dessas matrizes. Portanto, dentro de cada grupo de publicação encontrado anteriormente, executamos esse algoritmo para remover as duplicações presentes nele. Tendo limpado os dados, separamos o conjunto em 60% para treino, 20% para validação e 20% para teste. Por fim, para validar nossa base, treinamos modelos com a arquitetura AoANet [7]. A descrição completa da criação da base de dados pode ser encontrada em [12].

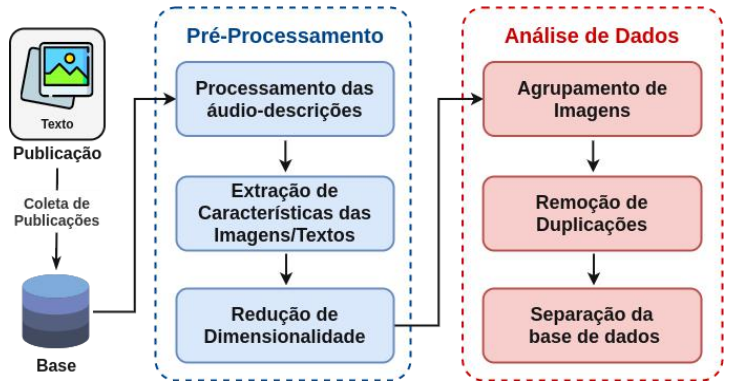

Figura 1: Procedimento adotado para a construção da base de dados *#PraCegoVer*. Primeiro, nós coletamos as publicações a partir do Instagram. Em seguida, limpamos as áudio-descrições, extraímos as características de imagens, reduzimos a dimensão dos vetores de características, agrupamos as imagens, removemos publicações duplicadas e separamos o conjunto de dados limpo.

Figura 2: Amostra de um grupo de publicações com imagens relacionadas a perfumes. Destacamos as imagens duplicadas.

Resultados e Discussão

Como resultado, temos a primeira grande base de dados construída para o problema de *image captioning* com textos em português. Esse conjunto conta atualmente com 550 mil pares de imagem-descrição, coletados de 15 mil diferentes perfis do Instagram. Devido à popularização do uso da marcação #PraCegoVer, nosso conjunto tem ainda grande potencial de crescimento. A Figura 3 mostra a evolução do número de publicações relacionadas ao #PraCegoVer e do tamanho da nossa base ao longo do tempo. Podemos notar que existe ainda muito espaço para crescimento da base, pois atualmente ela contém apenas um terço do volume total de publicações disponíveis. Portanto, continuaremos coletando mais dados ao longo do tempo. Durante a análise dos dados, realizamos o agrupamento das imagens a partir de características comuns. Nesse processo, encontramos mais de 600 classes distintas de imagens, representando uma grande diversidade de fotografias de animais, carros, aviões, flores, imagens publicitárias, objetos, *etc*. A título de comparação, o MS COCO, o conjunto mais utilizado na literatura de *image captioning*, contém cerca de 300 mil imagens distribuídas em 80 categorias de imagens com apenas um elemento principal, e 91 outras categorias mais gerais. O conjunto #PraCegoVer tem descrições com 40 palavras em média, enquanto que o tamanho médio das sentenças do MS COCO é 10 palavras. Podemos observar ainda, a partir da Figura 4, que o tamanho das sentenças é mais distribuído em nossa base que no MS COCO, *i.e.,* a variância é maior em nosso conjunto. A média e variância maiores introduzem dificuldades adicionais ao problema de gerar descrições, pois a métrica otimizada durante o treinamento, a CIDEr-D [11], penaliza significativamente o modelo por gerar sentenças cujo tamanho difere do tamanho da descrição de referência em poucas palavras.

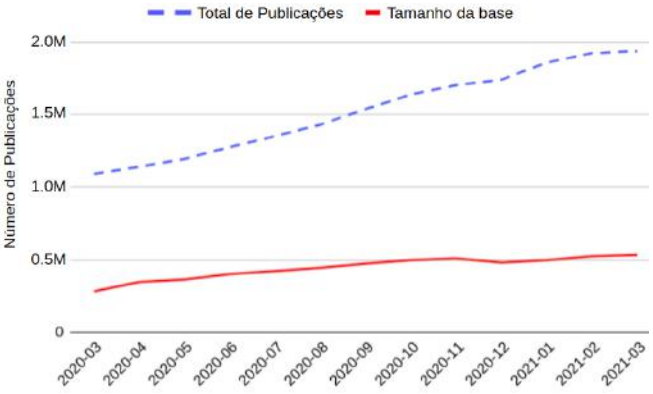

Figura 3: Gráfico de crescimento do número de publicações relacionadas ao #PraCegoVer (linha pontilhada) e tamanho da nossa base (linha contínua) ao longo do tempo.

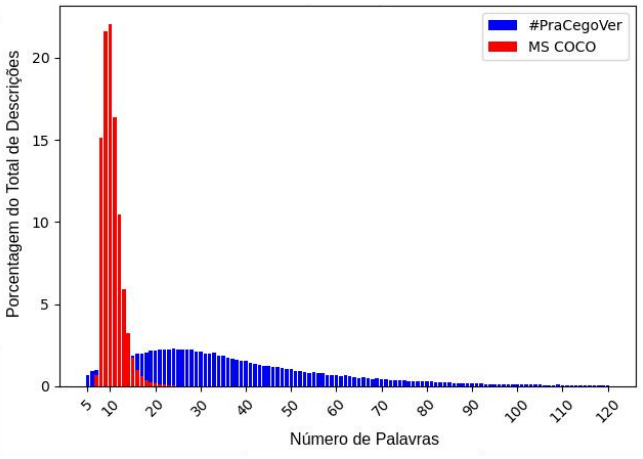

Figura 4: Amostra de um grupo de publicações com imagens relacionadas a pássaros.

(a) **Original**: Na foto, Thalita Gelenske e Thaís Silva estão abraçadas com Luana Génot na livraria Travessa. Ao fundo, diversos livros coloridos estão na prateleira. Nas laterais da foto, existem 2 banners: um deles vermelho, com o logo da, e o outro com a divulgação do livro da Luana.'

Gerada: Foto de uma mulher segurando um livro com livros. Ao fundo, uma estante com livros. Texto: “A sua. É sua festa. É sua!”.

(b) **Original**: Fotografia aérea sobre o pedágio da Terceira Ponte. A foto contém alguns prédios, um pedaço da Terceira Ponte e o fluxo de carros.'

Gerada: foto aérea aérea aérea da cidade de Florianópolis mostrando casas casas, mostrando casas casas. Ao fundo, algumas casas e casas.'

(c) **Original**: Quadrado laranja. Ao centro o texto em cor branca: 13ª Semana pela paz em casa. Em ambos os lados pequenas barras na cor azul com os respectivos dados: 2.333 processos movimentados. 610 sentenças. 552 despachos. 348 medidas protetivas. 239 audiências.'

Gerada: Imagem com fundo amarelo. Texto: “: você tem o que você ?”. Texto: “: você tem direito: você tem direito: você tem direito: você tem direito: R \$ 10 \%

(d) **Original**: Imagem de uma placa sinalizadora em ambiente urbano, com o fundo vermelho, letras pretas e o conteúdo: "Atenção - Mantenha-se à esquerda".

Gerada: Imagem de fundo de um quadro negro com a frase `` UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK

Figura 5: Exemplos de imagens seguidas da descrição originalmente escrita pelo autor da publicação, e do texto gerado automaticamente pelo modelo treinado na base #PraCegoVer.

Para ilustrar os desafios adicionais que nossa base traz para o problema de *image captioning*, nós treinamos o AoANet, um dos modelos estado-da-arte para o MS COCO, no nosso conjunto. Na Figura 5 apresentamos quatro exemplos de imagens acompanhadas das descrições originais e daquelas geradas pelo modelo treinado em nossa base. De maneira geral, podemos notar que o modelo tende a repetir palavras a fim de gerar sentenças de tamanho maior. Apesar dessas repetições, podemos ver que as descrições geradas nos exemplos da Figura 5(a) e Figura 5(b) de certa forma ainda conseguem descrever suas respectivas imagens. Especialmente com relação à Figura 5(b), vale destacar a presença da palavra "Florianópolis", que sugere de forma errada que a fotografia em questão é de uma paisagem da cidade de Florianópolis. Isso ocorre devido à presença de substantivos próprios nas descrições anotadas em nossa base, portanto substituir essas palavras por outras mais genéricas deve ajudar os modelos a gerar melhores descrições. Finalmente, as Figuras 5(c) e 5(d) ilustram um comportamento recorrente do modelo com relação às imagens publicitárias. Veja que são gerados símbolos totalmente sem sentido como em *"R \$ 10 \%"*, ou tokens "*UNK"* indicando palavras desconhecidas ou fora do vocabulário do modelo, e a descrição como um todo não reflete em nada o conteúdo textual presente na imagem. Todavia, esse é um problema muito complexo e dificilmente um modelo treinado para a tarefa de *image captioning* será capaz de capturar elementos textuais. Portanto, acreditamos que um *framework* combinando modelos especializados em certas tarefas, como por exemplo para *image captioning* e OCR (*Optical Character Recognition*), será capaz de gerar melhores descrições.

Conclusões

Neste trabalho, introduzimos o primeiro grande conjunto de dados construído para o problema de *image captioning* com descrições em português, *#PraCegoVer*. Essa base de dados traz, além da questão do idioma, desafios adicionais tais como maior média e variância do tamanho das descrições, e uma única referência para cada imagem. Demonstramos ainda que os modelos estado-da-arte ainda têm uma performance bastante ruim para esse cenário. Além do conjunto de dados, também propusemos um *framework* para a coleta e análise de dados do Instagram a partir de uma *hashtag*. Adicionalmente, desenvolvemos um algoritmo para a eliminação de publicações duplicadas com base no conteúdo visual e textual.

Como trabalho futuro, continuaremos coletando dados vinculados à marcação #PraCegoVer de forma a aumentar o volume de dados no *#PraCegoVer*. Além disso, faremos limpezas adicionais nas descrições anotadas, a fim de substituir substantivos próprios por palavras genéricas, o que reduzirá o tamanho do vocabulário do conjunto e, possivelmente, melhorará o treinamento de modelos sobre nossa base. Tendo em vista os problemas da métrica de avaliação CIDEr-D, em relação ao contexto de descrição única como referência, sentenças longas e com variância de tamanho considerável, exploraremos maneiras de flexibilizar a penalização por diferença de tamanhos entre a sentença gerada e a de referência, e também incluir penalizações ao modelo caso ele gere descrições com palavras repetidas várias vezes. Dessa forma, esperamos que treinando os modelos considerando essas alterações na CIDEr-D ele seja capaz de gerar textos mais descritivos quanto ao conteúdo das imagens. Por fim, investigaremos se a combinação de modelos de OCR com os de *image captioning* produz melhores descrições para o caso de imagens publicitárias.

Referências bibliográficas

[1] Agrawal, et al. "nocaps: novel object captioning at scale.", em ICCV, 2019.

[2] Chen, et al. "Microsoft coco captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325 (2015).

[3] Hodosh, et al.. "Framing image description as a ranking task: Data, models and evaluation metrics.", Journal of Artificial Intelligence Research, 2013.

[4] Plummer, et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.", em ICCV, 2015.

[5] Gurari, et al. "Captioning images taken by people who are blind.", em ECCV, 2020.

[6] Criadora do projeto #PraCegoVer incentiva a descrição de imagens na web. Web Para Todos. Disponível em http://mwpt.com.br/criadora-do-projeto-pracegover-incentiva-descricao-de-imagens-na-web, Acesso em 03 de Abril de 2021.

[7] Huang, et al. "Attention on attention for image captioning.", em ICCV, 2019.

[8] Sandler, et al. "MobileNetv2: Inverted residuals and linear bottlenecks.", em CVPR, 2018.

[9] McInnes, et al. "UMAP: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).

[10] McInnes, et al. "HDBSCAN: Hierarchical density based clustering." Journal of Open Source Software, 2017.

[11] Vedantam, et al. "CIDERr: Consensus-based image description evaluation.", em CVPR, 2015.

[12] Santos, et al. "#PraCegoVer: A Large Dataset for Image Captioning in Portuguese." arXiv preprint arXiv:2103.11474 (2021).