Utilização de uma Rede Neural Pré-Treinada para Descrever Imagens

Jordana Reis
Departamento de Informática da UFES
Brasil
jordana.reis@edu.ufes.br

***Resumo*—Este documento apresenta uma abordagem para descrever fotografias em forma textual. A implementação foi feita utilizando uma rede neural pré-treinada disponível na plataforma TensorFlow, através das APIs Keras.**

Palavras-chave—imagem, descrição de fotografia, transcrição de imagem, rede neural, TensorFlow, Keras.

I. INTRODUÇÃO

Tem sido amplamente aceito pela comunidade relacionada à ciência da cognição, que as crianças aprendem suas línguas nativas não através do aprendizado das palavras, mas pela assimilação das correlações entre o sinal da fala e a informação visual [1]. De acordo com [2], é interessante explorar uma máquina que possa traduzir palavras faladas em imagens, ou seja, traduzir a voz diretamente em imagem. Neste trabalho, a proposta é analisar uma imagem, mais especificamente uma fotografia, e descrevê-la de forma textual, com palavras da língua portuguesa.

Uma das áreas mais importantes da área de visão computacional é o reconhecimento de objetos, ou o processo de identificar um objeto em uma imagem ou vídeo. No mesmo contexto, identificar os objetos na imagem e descrever a cena representada por ela é igualmente importante para o avanço dos estudos neste campo de pesquisa. O propósito deste trabalho é transcrever a cena representada em uma dada imagem em texto na língua portuguesa. Prover a descrição textual de uma imagem requer a implementação de métodos de computação visual para entender o conteúdo da imagem e um modelo de linguagem do campo do processamento de linguagem natural para associação das palavras corretas em ordem adequada para descrever a imagem.

Nos últimos anos, métodos de aprendizagem profunda têm demonstrado bons resultados em problemas de descrição textual de imagens, principalmente com o advento de plataformas de aprendizado de máquina, como o TensorFlow e API de redes neurais que facilitam a experimentação de modelos disponíveis nestas plataformas, como o Keras.

A plataforma TensorFlow disponibiliza uma biblioteca de código aberto para aprendizado de máquina, aplicável a uma ampla variedade de tarefas. Trata-se de um sistema para criação e treinamento de redes neurais para detectar e decifrar padrões e correlações. Construído sobre o TensorFlow 2, o Keras disponibilizida todas as capacidades para atender o fluxo de trabalho de aprendizado de máquina, desde o gerenciamento do conjunto de dados até o treinamento e ajuste de hiperparâmetros das soluções [3]. Neste trabalho, será utilizada a plataforma TensorFlow e as API de alto-nível providas pelo Keras.

II. TRABALHOS CORRELATOS

No trabalho publicado em [4], os pesquisadores utilizaram redes neurais convolucionais para identificar as características faciais de suínos a partir da Image Data Generator disponível no Keras. O modelo apresentado neste caso, com o propósito de alterar o "modus operandi" desta indústria pecuária, que passaria a identificar os indivíduos não mais por RFID, mas por reconhecimento facial, apresentou taxa de reconhecimento próxima de 97%. Em [5], foi apresentada uma abordagem de aprendizado estatístico de modelos de tradução de imagens com textos associados a estas, onde identificou-se que o maior desafio estava relacionado à dificuldade de predizer a segunda palavra para descrição da imagem, a partir da primeira palavra para cada item da coleção de teste.

III. METODOLOGIA

A. *Coleção de dados*

O conjunto de dados utilizado neste trabalho foi o Flickr8K, que já contempla fotografias e descrições destas em língua inglesa. Este conjunto é relativamente pequeno, portanto, foi escolhido por não precisar de uma infraestrutura de máquina robusta, o que seria inconveniente para a conclusão deste estudo. Trata-se de coleção de referência para a descrição e pesquisa de imagens com base em frases, consistindo em 8.000 imagens, cada uma delas emparelhada com cinco legendas diferentes que fornecem descrições claras das entidades e eventos salientes [6]. Tais imagens foram escolhidas em seis grupos diferentes do Flickr e tendem a não conter pessoas ou locais conhecidos, mas foram selecionadas manualmente para representar uma variedade de cenas e situações [6].

A coleção disponibiliza por definição, um conjunto de treinamento com 6 mil imagens, um conjunto de desenvolvimento contendo 1 mil imagens e outras 1 mil imagens para testes.

B. *Métrica de Avaliação do Modelo*

A métrica para avaliar o desempenho do modelo foi a pontuação BLEU. Esta métrica é utilizada em traduções de texto para avaliar a tradução em relação a uma ou várias referências para traduções. Neste trabalho, foram comparadas cada descrição gerada em relação à todas as descrições de referência para a fotografia, depois foram calculadas as pontuações BLEU com a biblioteca python NLTK, que implementa este cálculo.

C. *Preparação do conjunto de fotografias*

O modelo pré-treinado Oxford Visual Geometry Group, ou VGG, foi o escolhido para interpretar o conteúdo das fotografias, através do acesso direto provido pela API do Keras. As características das fotografias foram pré-processadas usando o modelo VGG e salvas em um arquivo para uso posterior. Estes dados foram utilizados mais tarde para alimentar o modelo de interpretação de uma determinada fotografia no conjunto de dado, assim o treinamento dos modelos será feito mais rapidamente e consumindo menos recursos de máquina.

Como a proposta do modelo VGG é a predição de classes de fotografias, a última camada deste modelo não é necessária neste trabalho, pois não será feita classificação das imagens, apenas a descrição em forma de texto.

D. *Preparação do conjunto de textos*

O conjunto de texto possui uma ou várias descrições para cada uma das fotografias. Cada fotografia possui um identificador único que está relacionado no conjunto de texto que contém a descrição das imagens. Por isto, foi necessário um saneamento mínimo neste conjunto para diminuir o tamanho do vocabulário de palavras a ser utilizado de maneira a melhorar o tempo de treinamento do modelo. Basicamente, foram convertidas as palavras par letras maiúsculas, removidas as pontuações, as palavras com menos de dois caracteres e também palavras que continham números, salvando este novo conjunto em arquivo para uso posterior.

IV. O EXPERIMENTO

A. *Carga dos Dados*

Depois de preparar o conjunto de fotografias e o conjunto de textos, foi utilizado o conjunto de treinamento, monitorando o desempenho do modelo. O modelo desenvolvido gera uma descrição para uma determinada fotografia, sendo que é gerada uma palavra por vez para a descrição da fotografia. A sequência de palavras geradas anteriormente servirá de entrada para a execução corrente, portanto foi preciso definir uma palavra de início, identificando o início da descrição, e uma palavra de fim para identificar o final da descrição.

As descrições precisam ser codificadas em forma de números para serem usadas como entrada para o modelo de aprendizado, para isso, foi preciso criar um mapeamento de palavras para valores numéricos inteiros exclusivos. O Keras fornece uma classe, chamada Tokenizer que pode aprender este mapeamento a partir dos dados carregados. Esta classe foi utilizada para converter o dicionário de descrições em uma lista de caracteres e ajustar o Tokenizer a partir de uma descrição de fotografia carregada. Cada descrição de fotografia será dividida em palavras. O modelo receberá uma palavra e a fotografia, então gerará a próxima palavra. Em seguida, as duas primeiras palavras da descrição e a imagem serão fornecidas ao modelo como entrada, para gerar a próxima palavra. E desta forma o modelo será treinado.

No futuro, quando o modelo for usado para gerar as descrições, as palavras geradas serão concatenadas e fornecidas de forma recursiva como entrada do modelo, de modo a gerar uma descrição para uma fotografia.

B. *Definição do Modelo*

O modelo utilizado segue é baseado no "merge-model"[7].

- Extrator de Características de Fotografias: trata-se do modelo VGG com 16 camadas, pré-treinado com o conjunto de dados ImageNet, disponível através do Keras.
- Processador de Sequência: É uma camada de incorporação de palavras para tratamento das entradas de texto, seguida por uma camada de rede neural recorrente Long Short-Term Memory (LSTM).
- Tradutor/Decodificador: o Extrator de Características e o Processador de Sequência geram um vetor de comprimento fixo. Eles são mesclados e processados por esta camada, chamada 'Dense', ou Densa, para fazer a previsão final.

A figura 1 apresenta um exemplo da estrutura do modelo e formas das camadas:

Layer (type)	Output Shape	Param #	Connected to
input_18 (InputLayer)	[(None, 34)]	0	
input_17 (InputLayer)	[(None, 4096)]	0	
embedding_7 (Embedding)	(None, 34, 256)	1940224	input_18[0][0]
dropout_14 (Dropout)	(None, 4096)	0	input_17[0][0]
dropout_15 (Dropout)	(None, 34, 256)	0	embedding_7[0][0]
dense_23 (Dense)	(None, 256)	1048832	dropout_14[0][0]
lstm_7 (LSTM)	(None, 256)	525312	dropout_15[0][0]
add_31 (Add)	(None, 256)	0	dense_23[0][0] lstm_7[0][0]
dense_24 (Dense)	(None, 256)	65792	add_31[0][0]
dense_25 (Dense)	(None, 7579)	1947803	dense_24[0][0]

Total params: 5,527,963
Trainable params: 5,527,963
Non-trainable params: 0

Figura 1 - Estrutura do modelo e formas das camadas

C. *Ajuste do Modelo*

O desempenho do modelo foi monitorado no conjunto de dados de desenvolvimento, quando o desempenho melhorar ao final de cada época, o modelo é salvo em um arquivo para ao final, escolhermos o modelo que obteve o melhor desempenho para aplicá-lo ao conjunto final de treinamento como modelo final. Isto foi feito utilizando o ModelCheckpoint através do Keras, definindo o monitoramento da perda mínima no conjunto de validação e salvando em arquivo o modelo que alcançar a menor perda.

V. RESULTADOS

O modelo foi ajustado por 10 épocas, dado o tamanho do conjunto de dados e os recursos de máquina disponíveis. O melhor resultado foi ao final da época 4, com uma perda de 3.543 no conjunto de treinamento e 3.877 no conjunto de desenvolvimento. A partir da época 5, não houveram mais melhoramentos no modelo.

Após o ajuste do modelo, avaliou-se o seu desempenho no conjunto de dados de testes. A avaliação teve como base a geração de descrições para todas as fotografias do conjunto de teste e a avaliando tais previsões com uma função de custo padrão. Primeiramente, foi avaliado a capacidade de o modelo treinado gerar uma descrição para uma fotografia, utilizando as palavras de início descrita no item V, para servir como gatilho para o processo recursivo de chamada do modelo para gerar as palavras seguintes e compor a descrição da fotografia, até que a palavra que indica o fim da descrição seja encontrada ou o limite de tamanho da descrição seja alcançado.

A fotografia abaixo, que não faz parte do conjunto disponível pelo Flickr foi submetida para que o modelo a

descrevesse. Pelos dados de treinamento, o modelo identificou que se tratava de um cachorro, porém não conseguiu especificar na fotografia a quantidade de animais, na imagem um e o modelo descreve dois.

Figura 2 - Fotografia submetida ao modelo

```
>exemplo.jpg
startseq two dogs are running on the grass endseq
```

Figura 3 - Descrição retornada pelo modelo

REFERENCES

[1] E. Bergelson and D. Swingley, “At 6–9 months, human infants know the meanings of many common nouns,” Proceedings of the National Academy of Sciences, vol. 109, no. 9, pp. 3253–3258, 2012.

[2] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma and W. Gao, “Direct speech-to-image translation”, Journal Of Selected Topic On Signal Processing, Vol., No., Janeiro 2020, pp.01.

[3] M. Heller, “What is Keras?The deep neural network API explained”, Infoworld, Janeiro 2019, https://www.infoworld.com/article/3336192/what-is-keras-the-deep-neural-network-api-explained.html.

[4] K. Wang, C. Chen, Y. He, “Research on pig face recognition model based on keras convolutional neural network”, IOP Science Conference Series: Earth and Environmental Science, Agosto 2020.

[5] K. Barnard, P. Duygulub and D. Forsythc, “Recognition as translating images into Text”, IOP Science Conference Series: Earth and Environmental Science, Novembro 2002.

[6] M. Hodosh, P. Young, J. Hockenmaier, “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”, Journal of Artificial Intelligence Research, Agosto 2013.

[7] R. Nicole, “Where to put the image in a n image caption generator”, Natural Language Engineering, Março 2017.