Uso de rede neural para descrever imagem na língua portuguesa

1st Caio Barata Pinho
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
caio.pinho@edu.ufes.br

2nd David Pereira de Araújo
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
david.araujo@edu.ufes.br

***Resumo*—Este trabalho tem como objetivo demonstrar a utilização de uma rede neural na tarefa de descrever uma imagem em língua portuguesa de forma automática. Para realizar esse estudo, foi usada uma abordagem de aprendizado profundo especializada para trabalhos com imagens e vídeos. Por meio desse estudo foi possível verificar que é possível descrever imagens automaticamente. Foram realizados vários treinos e testes do modelo e serão apresentados nas seções pertinentes. Neste trabalho foi utilizada a API Keras e encontrados outros trabalhos similares que utilizaram outras abordagens para trabalhar com imagens.**

Palavras-chave— rede neural, CNN, imagem, legenda, visão computacional

I. Introdução

Visão computacional é uma abordagem da grande área de Inteligência Artificial (IA), que vem ganhando bastante atenção nos últimos anos e sendo utilizada como um recurso importante na produção de filmes e jogos eletrônicos. No entanto, o campo de visão computacional ainda é um desafio para pesquisadores na busca de soluções para determinados problemas. Por exemplo, criar legendas de forma automática de uma imagem é um grande desafio que a visão computacional pode resolver e recentemente ganhou forças para avançar nessa tarefa de descrever objetos de uma imagem automaticamente. Redes neurais profundas, em especial uma delas chamada rede neural convolucional, tem apresentado resultados satisfatórios em trabalhos recentes nesse campo de pesquisa com avanços significantes em trabalhos que descrevem imagens.

A rede neural convolucional pode ser uma grande aliada da visão computacional nas tarefas de descrever uma imagem, ou seja, gerar uma legenda de forma automática para uma determinada imagem. Redes Neurais Convolucionais ou em inglês Convolutional Neural Networks (CNNs) são tipos de redes neurais estruturadas para receber imagens como entrada. Essas imagens, quando interpretadas por um computador são transformadas em valores armazenados em uma matriz, ou podemos dizer tensor. Dessa forma, as CNNs mantém as características espaciais de uma imagem, como a altura e largura, e cores. No funcionamento de uma CNN há o processo de extração de características de imagens, que tem como objetivo principal a aplicação de filtros nas imagens de entrada da rede. O processo realizado por esses filtros tem um nome, é uma convolução. Por isso o nome de rede neural convolucional.

II. Trabalhos Correlatos

A. Outras abordagens

Nesta seção, será apresentado um breve histórico de trabalhos relacionados com geração automática de legenda de imagens. Na literatura pertinente, é possível encontrar vários trabalhos que utilizaram outras abordagens na geração de legenda de imagens. Por exemplo, redes neurais recorrentes em inglês Recurrent Neural Network (RNN) em combinação com CNN. O trabalho de [1], utilizou a combinação de vários métodos, CNN, RNN e em inglês Long Short Term Memory (LSTM), para trabalhar com geração de legenda para imagens.

Já [2], utilizou o mecanismo de atenção para gerar legenda de imagens. E no mesmo trabalho é citado pelos os autores a utilização de outros métodos em outros trabalhos, como por exemplo, a rede Memória de Curto e Longo Prazo ou simplesmente LSTM, que também é uma rede neural recorrente.

III. Metodologia

Neste trabalho, foi utilizado o modelo de rede neural convolucional com a API Keras para realizar a tarefa de descrever uma imagem na língua portuguesa de forma automática. Para realizar os experimentos e testes foi utilizado o dataset FLICKR30K. Esse conjunto de dados foi dividido em partes de treino e testes. Durante a etapa de testes foi possível obter vários resultados, no entanto, por falta de recursos computacionais os resultados não foram tão bons como o esperado.

A. Um pouco sobre CNN

A CNN é um rede neural profunda [3], ou seja, um algoritmo de aprendizado profundo que pode ter como dado de entrada uma imagem, atribuir pesos e vieses, rede essa especializada e utilizada para reconhecimento de imagens, problemas de classificação de imagens, etc. Uma CNN é constituída pelas etapas de:

- Rede neural para extrair características da imagem de entrada;
- Rede neural para classificar a imagem de entrada a partir das características;

- Uma saída da CNN que apresenta a possível classe da imagem de entrada.

B. Arquitetura do Modelo Proposto

A arquitetura do modelo proposto é simples e no modelo do encoder possui apenas quatro camadas conforme mostrado na Figura 1.

```
Model: "cnn__encoder_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inception_v3 (Functional)    (None, None, None, 2048)  21802784
_________________________________________________________________
model (Functional)           (None, None, None, 2048)  21802784
_________________________________________________________________
reshape (Reshape)            multiple                  0
_________________________________________________________________
dense (Dense)                multiple                  1049088
=================================================================
Total params: 22,851,872
Trainable params: 1,049,088
Non-trainable params: 21,802,784
_________________________________________________________________
```

Figura 1. Encoder

A Figura 2 apresenta o modelo do decoder possui também apenas quatro camadas conforme mostrado na Figura 2.

```
Model: "rnn__decoder_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      multiple                  5120512
_________________________________________________________________
gru_1 (GRU)                  multiple                  984576
_________________________________________________________________
dense_6 (Dense)              multiple                  2570257
_________________________________________________________________
bahdanau_attention_1 (Bahdan multiple                  197377
=================================================================
Total params: 8,872,722
Trainable params: 8,872,722
Non-trainable params: 0
_________________________________________________________________
```

Figura 2. Decoder

C. Camada convolucional

A camada convolucional é vista como o resultado da aplicação dos filtros mencionado em seção anterior. Esse filtro, poder ser chamado de kernel e também é uma matriz com valores. Existem diversos tipo de filtros, seus valores dependem do objetivo do filtro. Alguns filtros mais conhecidos são: filtro para detecção de bordas, desfoque, nitidez, etc.

A camada de convolução gera mapas de características processando a imagem através de filtros (ou kernels); para obter as saídas dessa camada, é necessário passar os mapas por funções de ativação (por exemplo, ReLU) e essa mesmas camadas produzem um número de mapa de características igual ao número de filtros utilizados.

D. API Keras

A API Keras é uma tecnologia de alto nível do TensorFlow para trabalhar com modelos de aprendizado profundo [4]. Ela é usada para prototipagem rápida que, permite construir, treinar, avaliar e executar de forma fácil os diversos tipos de redes neurais artificiais. TensorFlow é uma tecnologia criada pela Google e é uma das bibliotecas integrada com a API Keras. Existem diversos já modelos implementados e fornece conjuntos de dados. A API pode manipular modelos com topologia não linear, camadas compartilhadas e até mesmo várias entradas ou saídas. A ideia principal do modelo de aprendizado profundo geralmente é trabalhar com um gráfico de camadas.

IV. Experimentos e Resultados

A. Validação do modelo

Na Figura 3 é apresentado o resultado da validação do modelo que contém a Loss e as Épocas.

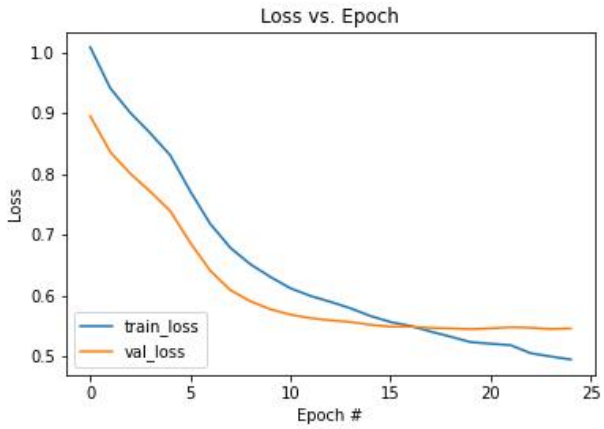

Figura 3. Loss vs Epoch

Após os ajustes e consolidação do modelo, foram realizados diversos testes com o conjunto de dados FLICKR30K. As Figuras 4 e 5 apresentam os resultados experimental do modelo proposto para descrever uma imagem. Foram utilizadas imagens disponíveis na internet para esses experimentos. Na Figura 4, o resultado apresentado pelo o modelo foi: ”os homens de uniformes amarelo e azul estão jogando futebol”. Já na Figura 5, o resultado apresentado foi: ”um ciclista está andando de bicicleta”.

Figura 4. Resultado do primeiro teste

A Figura 5 foi uma das imagens utilizadas nos testes com resultado satisfatório.

Figura 5. Resultado do segundo teste

V. Conclusão

Com este estudo, foi possível verificar que é possível descrever uma imagem de forma automática utilizando redes neurais profundas utilizando tecnologias gratuitas e disponíveis para a comunidade científica e acadêmica. Embora o modelo proposto sendo bastante simples e mesmo com a precisão dos resultados não chegando a um percentual de 90% nos resultados, foi verificado a possibilidade de utilizar a rede CNN para a tarefa de gerar legenda para uma imagem, sendo um possível caminho para trabalhos futuros e experimentos utilizando redes neurais para tarefas com imagens.

Referências

[1] M. Tanti, A. Gatt, and K. P. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?," Proceedings of The 10th International Natural Language Generation conference, Santiago de Compostela, Spain, pp. 51–60, Agosto 2017.

[2] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," Proceedings of the 32nd International Conference on Machine Learning, PMLR. Lille, France, vol. 37, pp. 2048–2057, Julho 2015.

[3] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," The MIT Press, 800 pp, 2016.

[4] A. Gulli and S. Pal, "Deep learning with Keras," Packt Publishing Ltd. 2017.