NLP project proposal

Naïm LEHBIBEN, Baptiste ZLOCH, Matthieu MONNOT

April 2024

The project will consists in three parts:

1 The dataset

We will leverage the Flickr8k dataset[1], containing **8000 image**s, each accompanied by **five** corresponding captions. To capture the essence of each image, we will extract visual features using a pre-trained visual model like a Convolutional Neural Network (CNN) model, such as VGG16 or ResNet-50. This process effectively transforms the image data into a lower-dimensional vector representation. Each caption will undergo pre-processing steps like tokenization, lemmatization, and stop word removal. Following this, we will employ word embedding techniques like Word2Vec or GloVe (as seen during the lectures) to convert each word into a dense vector, enabling the model to grasp semantic relationships between words.

2 The model

We will explore a deep learning architecture known as an Encoder-Decoder framework. The encoder will process the image features extracted by the visual model[2], while the decoder will be responsible for generating a sequence of words representing the image caption. To handle the sequential nature of text data, recurrent neural network and especially Long Short-Term Memory (LSTM) cells are likely to be utilized within both the encoder and decoder. The model will be trained on a designated portion of the Flickr8k dataset. A separate validation set will be employed to monitor performance during training and prevent overfitting. The model will be optimzed using Adam optimizer for adaptive learning rate and we will employ regularization in order to avoid overfitting. To assess the model's effectiveness, we will utilize the BLEU (BiLingual Evaluation Understudy) score. BLEU compares the generated caption to the five reference captions available for each image, evaluating its n-gram precision.

3 The baseline

To establish a benchmark for comparison, we could leverage several baseline approaches, our preferred one: Nearest Neighbor Search: We will search the training dataset for images with visual features similar to the query image and retrieve their corresponding captions.

Comparing the BLEU score achieved by our proposed model with these baseline approaches will demonstrate the efficacy of our NLP approach in image captioning.

[1] The dataset is available here : https://www.kaggle.com/datasets/adityajn105/flickr8k

[2] It could be a CNN or more complex model