Advanced Image Captioning: Capstone Project

Please make this document anonymous. Your team name should be anonymous.

Team name: *JELLYFAM*

Overview of Project

Image captioning is the task of generating a textual description for a given image. For our project, we hope to develop an image captioning model that can generate accurate and meaningful captions to photos by exploring various deep learning techniques. We will systematically experiment with different aspects of the model, such as data preprocessing, feature extraction, model architectures, and word embedding to find the best way to produce such captions. Below, we will describe the steps to our project and how we will experiment at each step

Steps of the Project

We believe there are a few steps to the project that will require us to experiment at each step. In this proposal, we will explain how we plan to deal with each step, and what experiments we plan to conduct

- Dataset Extraction and Preprocessing
- Model Architecture
- Evaluation

Dataset Extraction and Preprocessing

In order, to find our data, we plan to take advantage of some of the pre-exisiting datasets built for image captioning. This includes

- MS COCO: A large dataset containing 330K images with 5 captions per image
- Flickr30k: A dataset containing 30K images with 5 captions per image
- Flickr8k: A smaller dataset containing 8K images with 5 captions per image

Depending on our compute resources and how our own machines handle the data, we will select one of the three to work with. Since we are not scraping our own data, the only work we will have to do on the preprocessing end is preprocessing our images, preprocessing our captions and building a vocabulary/embeddings for our captions. We believe we can experiment at each level, and will test as many options based on if time permits.

We hope to use Google Collab to meet our GPU needs

Image Processing

Some ways we plan to experiment with our image processing are:

- Image resizing: Experiment with different resizing methods, such as preserving the aspect ratio, center cropping, or padding the images to a fixed size
- Image normalization: Test different normalization techniques, such as scaling pixel values to the range [0, 1], subtracting the mean and dividing by the standard deviation
- Data augmentation: Apply data augmentation techniques, like random flipping, rotation to increase the diversity of the training set and improve generalization

We will also decide if we want to handle images with or without color

Caption Processing

In order to process our captions, we plan to use the spaCY nlp library, which allows more complex tokenization (lowercasing, lemmatization, remove stop words, remove punctuation and extra white space ,use only top 1000 most frequent words, and replace the rest with OOV, replace numbers with NUM). We will experiment with parameters of the vocabulary (size, frequency). With regards to the word embeddings of our captions we can experiment a bit with using non-contextualized word embeddings (word2vec) and contextualized word embeddings (BERT).

Model Architecture

We are considering a few architectures to use for this assignment, but plan to experiment greatly with each architecture in order to find the best model available.

The architectures we plan to experiment with are

- CNN + RNN
- CNN + LSTM
- CNN + Transformer
- CNN + BERT Transformer
- LSTM only
- RNN only
- CNN + GRU
- GRU only

In this architecture, the CNN will serve as the feature extractor and then be fed into an RNN/LSTM/Transformer in order to produce a captioning sequence. Given that we may use BERT embeddings, we included using architectures such as the BERT transformer in order to take advantage of contextualized word embeddings.

For feature extraction, we will also experiment with using different pre-trained CNN's (VGG, ResNet, Inception, DenseNet, or MobileNet) as well as our own CNN models. We can also experiment with using multiple CNN's to extract features and concatenating those features to feed into the caption generation portion

For our LSTM/RNN/GRU/Transformers models, we will experiment regularization techniques (L1 or L2), dropout layers, varying the number of LSTM layers, different activation functions, and bidirectional LSTMs.

Evaluation

To evaluate our performance, we will be using cross-validation across multiple folds of data to ensure our data isn't overfitting. To evaluate the accuracy of our captions with the true captions, we can use metrics such as BLEU.

Ethics and Impact

Three stakeholders that we can imagine being affected by this project are a) people who are visually-impaired and utilize captions to understand images. This type of technology is very important and can influence how visually impaired people can understand the world around us. Another stakeholder affected by this project are social media sites, as they will need to be aware of when captions are genuine or generated artificially, as image captioning software could theoretically be trained to be used in a harmful

manner/reinforce exisiting biases. Another stakeholder that this project could impact are photographers, as this technology being used incorrectly could mischaracterize their work within a caption.

The socio-historic context that this project lives in is that it is very difficult from a human perspective, to continually caption images as that requires a lot of manual work. In order to reduce that work, and potentially increase accessibility of understanding said photos, our software aims to help reduce the effort needed to produce accurate image captions

Division of Labor + Progress

We believe the best way to go about this project is to start small and then continue adding complexity in the form of more sophisticated models. In terms of the code, we expect the code between models to be largely the same outside of changing small portions in each instance. We can report our progress as we go.

We expect the division of labor to look like as following:

1 Member: Handles image preprocessing 1 Member: Handles caption preprocessing 1 member implements implementing evaluation criteria and accuracy/cross validation Each of the members will have to experiment with 2 of the 8 architectures listed in Model Architecture. The member who doesn't do one of the 3 above, will handle the more-complex architectures.

We anticipate the most difficult parts of the project being handling all the data, finding a way to work effectively between 4 partners, and modularizing the code effectively to make building models simpler.

Skills

All of our team members are senior-year CS students who have all taken Deep Learning, Machine Learning and Computer Vision. Members of the team have also taken the Computational Lingustics course and the Artificial Intelligence course. All members of the team have completed the basic requirements for an scB at Brown