Team Name:

TeamOfExplorer

Project Title:

Exploring model architectures for image caption task on COCO

Project summary

Multimodal machine learning, which combines information from multiple modalities such as text, image, and audio, has emerged as a promising research area in recent years. One of the most exciting applications of multimodal machine learning is image captioning. Image Captioning is the process of generating textual descriptions of an image. This task has great practical significance in areas such as assistive technology for the visually impaired, automated image indexing and retrieval, and personal digital assistants. The most common deep-learning based solutions address this task in two steps: 1. a visual encoder model capable of generating the most relevant feature abstracts from a given image 2. a language model capable of generating meaningful and syntactically correct sentences. In this project, our team will explore the combinations of various model structures and aim to build a good model of our own for the task of image caption.

Approach

We will split the task into two steps: visual encoding and language modeling (decoding).

For the visual encoding part, we will try extracting image features using:1.Non-attentive global CNN features;2.attention over grid of CNN features;3.MLP-mixer based encoder.

The features extracted from the visual encoding will be the input for the language modeling (decoding), where we will try LSTM-based RNN models and transformer based models.

If possible, we will also try some text-image fusion models which combine encoder and decoder into a single stream of Transformer layers such as the ViLBERT.

We will apply various model evaluation matrices as suggested by the coco image caption dataset: 1. BLEU 2.CIDEr.

Our pipeline will include:

1. Build a pipeline for data preprocessing (image transform and text encoding) and a dataloader

2. Implement interfaces for both visual encoding and language modelings based on Vinyals et al 2014 “Show and Tell: A Neural Image Caption Generator” using pytorch.

3. Implement other different models(CNN+attention,MLP-mixer) for the visual encoding interface in step 2. Refine the interface structure using polymorphism if required.

4. Implement other different models(such as Transformer-based Architectures) for the language interface in step 2. Refine the interface structure using polymorphism if required.

5. Load pretrained models, further training the pipeline to get acceptable performance for each combination.

6. Implement ViLBERT model or other fusion models from pretrained large image-text datasets, apply it to image captioning and check the performance.

7. Analyze experiment results and write up the project.

Related Work

The image caption task is a very classical task in computer vision and natural language processing. Some earliest works on image captioning were presented in 2015 by Vinyals et al. at ICML2015 and by Andrej Karpathy et al at CVPR2015, where both of them proposed a neural network-based model that consists of two main components: an encoder, which processes the input image and encodes it into a fixed-length vector representation, and a decoder, which generates the corresponding caption word by word based on the encoded representation of the image. In Vinyals's paper, the encoder is a CNN that is pre-trained on a large image classification dataset, while the decoder is a RNN that uses LSTM cells.Their results have similar performances with human, and has a BLEU-4=27.7 on COCO dataset. One of the most cited online source code for these two papers is neuraltalk2, which is implemented by Andrej Karpathy using Pytorch and Lua.

Starting from 2015, the task has generally been addressed with pipelines composed of a visual encoder and a language model for text generation. The advancement includes:

1.Attention mechanisms to allow the model to selectively attend to different parts of the input image when generating the caption. For example, the Show, Attend and Tell model (Xu et al., 2015) uses a visual attention mechanism to attend to different regions of the image while generating words in the caption. And the SCAN model by Lee et al. (2018) uses a hierarchical attention mechanism to attend to both image regions and words in the caption. The SCAN model achieved state-of-the-art performance on the COCO dataset at the time of publication.

2.Transformer-based models that can handle longer sequences of text than RNN-based models, and can better capture long-range dependencies in the image-caption pairs. For example, the ViLBERT model by Lu et al. (2019) combines a BERT-based language model with a VisualBERT-based image encoder to generate captions for images. The model achieved state-of-the-art performance on the COCO dataset at the time of publication.

3.Vision-language pretraining models. The models are trained on large-scale datasets that combine both image and language data. Then, the pretraining models then can be

used in the image caption tasks. For example, the ViLBERT model by Lu et al. (2019) and the OSCAR model by Li et al. (2020). The OSCAR model pre-trains a transformer-based language model on a dataset of over 4 million image-caption pairs. The pre-trained model can then be fine-tuned for image captioning. Their model has BLEU-4=41.7.

4. Multi-model learning: such as Kim et al. (2021) uses cross-modal skip-connections between a vision transformer and a language transformer, and they achieve BLEU-4=46.5, which is currently state of art based on https://paperswithcode.com/sota/image-captioning-on-coco-captions.

There is also a lot of work focusing on improving the mappings of the structures between feature space and text space to get good results. In this project, we will only focus on the generations of the feature space and the generation of the text space.

Datasets

We will train and test our models using Microsoft COCO dataset, which consists of images of complex scenes with people, animals, and common everyday objects in their context.

List your Group members.

kyu311,qqi33,rzhang668,yshang42

References:

1. “From Show to Tell: A Survey on Deep Learning-based Image Captioning” by Matteo Stefanini et al (2021)
2. "Microsoft COCO Captions: Data Collection and Evaluation Server" by Chen et al (2015)
3. “Microsoft COCO: Common objects in context” by Tsung-Yi Lin et al (2014)
4. "MLP-Mixer: An all-MLP Architecture for Vision" by Ilya Tolstikhin, et al (2019)
5. “Show and Tell: A Neural Image Caption Generator” by Oriol Vinyals et al (2015)
6. "Deep Visual-Semantic Alignments for Generating Image Descriptions" by Andrej Karpathy and Li Fei-Fei (2015)
7. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention” Kelvin Xu, et al (2015)
8. “Stacked Cross Attention for Image-Text Matching” by Kuang-Huei Lee et al (2018)
9. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks” by Jiasen Lu, et al (2019)
10. “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks” by Xiujun Li et al (2020)
11. “mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections” by Chenliang Li et al (2022)
12. Github Neuraltalk2: https://github.com/karpathy/neuraltalk2
13. https://paperswithcode.com/sota/image-captioning-on-coco-captions