Multi-Modal and Transformer-Based Image Captioning

Chris Kim
chris.c.kim@stanford.edu

Thomas Jiang
twjiang@stanford.edu

Ruth Ann Armstrong
ruthanna@stanford.edu

BACKGROUND/INTRODUCTION

Motivation

- Image captioning is a widely used benchmark for computer vision that has important real-world applications.
- Over 12 million individuals over the age of 18 in the US are affected with visual impairment issues, making image captioning an important area of research.
- As such, improving caption generation models has important real-world benefits.
- From a technical perspective, image captioning is also interesting as it incorporates two areas of machine learning: computer vision and natural language processing.

Existing Approaches

- Task Agnostic Multi-Modal Architecture: One for All (OFA), Wang et al. (2020)
- It is a single model that can be used to generate captions.
- Incorporating a decoder architecture results in better accuracy.

Transformers

- Transformer-based models have been successful in a wide range of tasks.
- Each sub-object is processed independently, which is effective for generating accurate text summaries.

METHODS

Model Architecture

- Encoder architecture:
  - 3-layer CNN with output channels 12, 12, 24 with size 3 kernel
  - 2 max-pooling layers of kernel size 2
  - Dropout layer with probability 0.2
  - Fully connected layer
  - Logarithmic softmax

- Transformer architecture:
  - Layered CNN with sequential LSTM layers

- The architectures are connected as the CNN encoder output serves as the initial input to the encoder LSTM.

- We also experimented with using pretrained CNNs including ResNet, VGG19-LSTM and ResNet-18 and with replacing LSTM modules with RNNS.

- Incorporating attention mechanisms into models can improve performance.

- Attention mechanisms can be used to focus on specific parts of the image.

- Different attention mechanisms to adapt to different parts of the image.

- Attention to wrong object:

- Predicted caption: “Two people are standing on the beach.”
- Sampled true captions: 
  - “Two people are standing on the beach.”
  - “Two people are standing on the beach.”

- BLEU-4: 0.92

- Sampled True Captions: “Two kids are running through the brush.”
- Predicted caption: “Two kids are running through the brush.”

- BLEU-4: 0.87

- Sampled True Captions: “A dog jumps over a chain.”
- Predicted caption: “A dog jumps over a chain.”

- BLEU-4: 0.8

CONCLUSION AND FUTURE WORK

- Our best model, a ResNet18-Transformer achieved a BLEU-1 score of 0.879 and a BLEU-4 score of 0.543.

- Our transformer architecture with 2 Encoder Layers and 4 Decoder Layers performed best, outperforming deeper and larger transformer architectures.

- This is likely because the size of our dataset was relatively small.

- For future work, we are interested in exploring better ways to adapt state-of-the-art architectures to settings with low amounts of data similar to that of Flickr8K.