# 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 computer vision task that has **important real world applications**. Over 12 million individuals over the age of 40 are affected with vision impairment issues, and rely on image captions to navigate the internet. 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. (2022)**
OFA is a single model that generalizes to multiple tasks by incorporating a shared representation space for text and images.

**Leveraging information from pretrained language models: Fusion Model, Kalimuthu et al. (2020), VisualGPT Chen et al. (2021)**
These models integrate language distribution knowledge contained in large pretrained language models to boost model efficacy.

**Reinforcement learning: Visual Reserved Model, Wei et al. (2021), Partial Off-Policy Learning, Shi et al. (2021)**
Wei et al. combine the inclusion of past visual context when producing captions and a policy-gradient RL algorithm, while Shi et al. use an off-policy learning scheme to increase the diversity of produced captions.

**LSTMs and RNNs: m-RNNs, Mao et al. (2014), Bidirectional LSTMs (2016)**
RNNs leverage a recurrent neural architecture to produce sequences of texts based on features input into the model. LSTMs improve upon RNNs by including memory cells that are better at retaining information over longer ranges.

## PROBLEM STATEMENT

Given an image as input, our model aims to produce a sequence of text tokens which accurately describe the image. Captions are evaluated using BLEU, which counts matching n-grams in the target text to n-grams in the reference caption.

Image → Model → Predicted Caption

True Caption } BLEU Evaluation

## DATASET

**Flickr8K**

**8,091 images of varying sizes**

**37,197 captions**

[A blonde horse and a blonde girl in a black sweatshirt are staring at a fire in a barrel., A girl and her horse stand by a fire. A girl holding a horse 's lead behind a fire., A man , and girl and two horses are near a contained fire., Two people and two horses watching a fire.]

**Zero-padding to max dim, 500x500**

**Normalization**

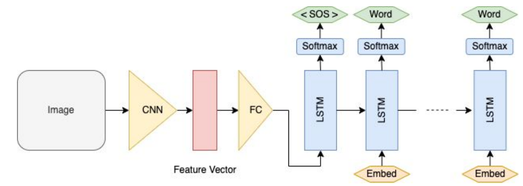
**Rescaling to 224x224 resolution for compatibility with pretrained CNNs**

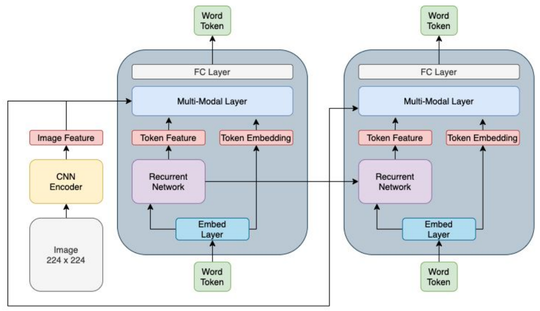
**Punctuation removal and lower-casing**

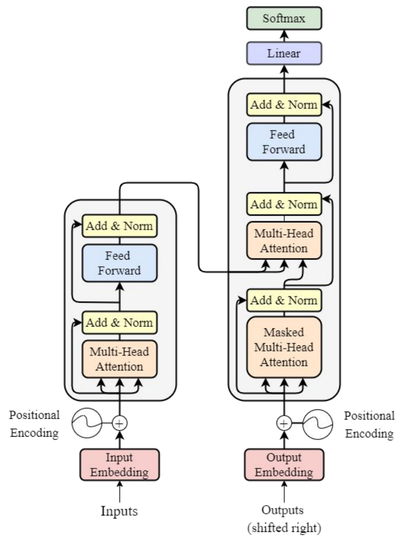
**Tokenization, splitting on whitespace + zero padding to length 20 + '<SOS>' and '<EOS>' tags**

**Generation of one-hot embedded word vectors for top *n* most common words**

## METHODS




### Baseline: CNN + LSTM/RNN, Models with Pretrained Encoder Networks

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

Decoder architecture
- L-layered RNN 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 AlexNet, VGG16-LSTM and ResNet-18 and with replacing LSTM modules with RNNs

### Multi-modal Architecture

- Accepts a concatenated input vector containing image features, word token features and word token embeddings
- Image features are extracted from pretrained CNN encoder modules
- Recurrent network layer consisted of variations in the basic RNN and LSTM cells and Transformer modules
- The output of the multi-modal layer is passed through a final fully-connected linear layer with a softmax activation to produce predicted word tokens

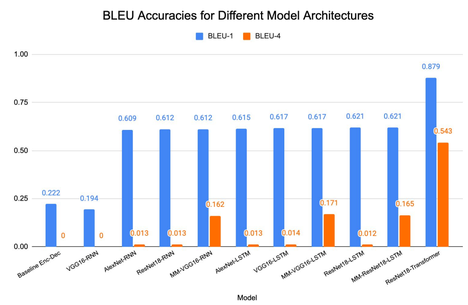
### Transformer

Encoder
- Image features are extracted from a pretrained ResNet18 model
- ResNet18 fine-tuned to remove last two layers for direct image features

Decoder
- Multi-headed attention used to grant importance to specific features
- Positional encoding was used in the model to capture token order
- Cross entropy loss between target sequence and caption sequence.

Trained for 50 epochs with lr = 0.00001 and dropout = 0.2, experimenting with n_heads and num_layers for encoder and decoder

## EXPERIMENTS AND ANALYSIS




### Quantitative Analysis

The baseline encoder-decoder + VGG16-RNN significantly underperforms comparatively (this model was trained on 500 images for model iteration)

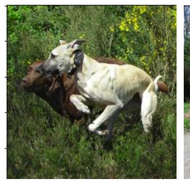
Different CNN encoder modules leads to minor changes in scores

There is no significant difference in performance between RNN and LSTM modules

Multimodal models had relatively mediocre performance

The transformer architecture significantly outperformed all other models

**BLEU-4:** 0.826
**Sampled True Captions:** ['two kids playing on the beach, close to the water','two children are standing on the shore next to a body of water']
**Predicted Caption:** 'two people are standing on the beach'



**BLEU-4:** 0.947
**Sampled True Captions:** ['two dogs run through the brush', 'a white dog races a brown dog in a field of grass']
**Predicted Caption:** 'two dogs are running through the grass'

**BLEU-4:** 0.262
**Sampled True Captions:** ['a dog jumps over a chain', 'a brown dog jumps over a chain']
**Predicted Caption:** 'two dogs are running through the grass'

### Common Caption Errors for Transformer Models

**Repeated words and phrases:** Some captions featured repeated words and phrases, which might have been due to the **attention mechanism incorrectly focusing** on the **same part of the image multiple times**.

**Attention to wrong object:** In some cases, the attention mechanism paid attention to wrong portions of the image. A possible solution for this might be the use of **larger pretrained CNNs** which are better at extracting **more detailed image features**.

**Different images with similar captions:** Some images with **similar color schemes** had **similar captions** which were sometimes incorrect. Further finetuning or a larger CNN might help the model **differentiate better** between these types of images.

## 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 .