Neural Image Captioning Model

Cameron Nichols(camnic), Chloe Zeien(czeien), Thomas He(mperform)

Algorithm

Encoder

In order to encode input photos into a compact representation, a Convolutional Neural Network (CNN)—more precisely, a ResNet-50 architecture that has been pre-trained on ImageNet—will be employed. The CNN's last fully connected layer, which usually functions as a classifier, will be eliminated so that the network can produce a feature vector that captures the image's content.

Decoder

The decoder will be a single-layer Long Short-Term Memory (LSTM) network. It will start with the feature vector from the encoder and produce a string of words to create the caption. Because LSTMs are recurring, they can take into account the context of the sequence, which makes them ideal for producing coherent phrases.

Implementation

Model Setup

To use transfer learning, we will start our encoder with a ResNet-50 model that has already been trained. Then, we will use the MS COCO dataset to further refine the model for contextual relevance.

Training and Attention

We will use an end-to-end training procedure that weights picture features for the LSTM utilizing a soft attention mechanism, and a combination loss function that maximizes attention efficacy and caption accuracy.

Evaluation

BLEU-4: To evaluate n-gram precision and sentence brevity.
METEOR: To assess alignment with human-like understanding, taking synonyms and grammar into account.
CIDEr: To measure the distinctiveness of captions in relation to human judgments.

References

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164, 2015.