CS 231N Short Form Video Captioning with C3D and CLIP
Chenyang Dai (qddaichy@stanford.edu)

Objectives

The goal of this project is building video captioning system for the short form videos focusing on MSVD and MSR-VTT dataset. We mainly have three contributions:
1. Implement and train two encoder-decoder models using C3D and CLIP encoding while both models use GPT-2 for decoding.
2. Conduct various experiments regarding training strategies and find that pretraining with image captioning and fine tuning specific layers can greatly boost performance.
3. Conclude that adopting beam search, data augmentation and use more input frames can further improve the performance.

C3D Encoding

The C3D encoding model consists of 8 3D convolutional layers, 5 pooling layers and 3 fully connected layers. The final fully connected layers have a size of 15360 dimensions, which is then tokenized into 20 GPT-2 context tokens. GPT-2 uses context tokens to conditionally generate captioning in natural language.

CLIP Encoding

For CLIP encoding model, we have custom logic to parallel process video frames. There is no fusion logic for different frames’ CLIP feature vectors. We have a custom MLP layer that directly maps each frame feature to 4 GPT-2 tokens (4 x 5 frames = 20 tokens in total). The reason behind is that we want to defer the fusion logic to GPT-2 who has a lot more attention blocks and layers that can gradually fuse the information together.

GPT-2 Decoding and Loss Function

We concatenate the 20 context tokens with the tokenized ground truth captioning text to from one training example. Then we pad all examples to the same length. For each example, we build the corresponding attention mask allowing the model to pay attention to context tokens while masking out all future tokens or pad tokens.

The model is able to tell there are multiple important objects (a dog and a toy) and their relationship (playing with).

The model is able to capture different human motions like cleaning and climbing which is very impressive.

Reference
