Natural Language Processing with Deep Learning
CS224N/Ling284

Comparing Attention-based Neural Architectures for Video Captioning

Jason Li, Helen Qiu
jasonkli@stanford.edu, shiqiu21@stanford.edu
CS224N-Winter2019, Stanford University

Introduction & Related Work

- Video captioning is the task of understanding the visual content of video sequence and translating that understanding to an appropriate caption.
- Automatic video description generation can have many practical applications in daily scenarios such as video retrieval, blind navigation, and automatic video subtitling [1].

- Early works in video captioning extracted features from 2D-CNNs for each frame and averaged them before inputting into an LSTM.
- Venugopalan [2] et al. used a sequence-to-sequence model with per-frame 2D-CNN features as the input sequence.
- Later works incorporated attention, with a particular emphasis on temporal attention to attend to different features over time.

In our work, we explore spatio-temporal attention over features extracted from a pre-trained P3D (Pseudo-3D ResNet). We also compare the performance of LSTMS to Transformers.

Data

Dataset: MSVD (Microsoft Video Dataset)

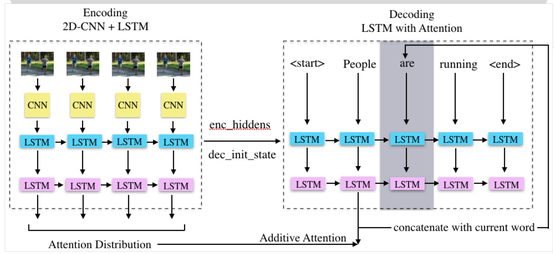
- 1969 videos – average of 10 seconds
- Average of 40 captions each
- We used 5 captions per video for training
- 1200 training, 100 validation, 669 training split
- Captions are usually 5-10 words long – lack of diversity with many similar captions across videos

References

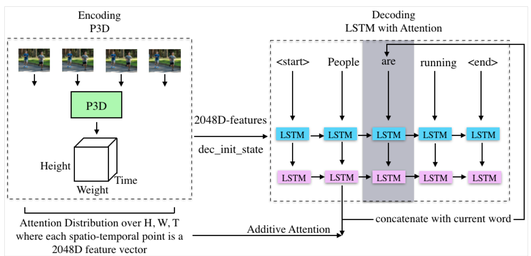
[1] Song, J., Gao, L., Liu, L., Zhu, X., & Sebe, N. (2018). Quantization-based hashing: a general framework for scalable image and video retrieval. Pattern Recognition, 75, 175-187.
[2] [Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R.J., Darrell, T., & Saenko, K. (2015). Sequence to Sequence – Video to Text. 2015 IEEE International Conference on Computer Vision (ICCV), 4534-4542.

Approach

1. ResNet-LSTM Encoder + LSTM Decoder

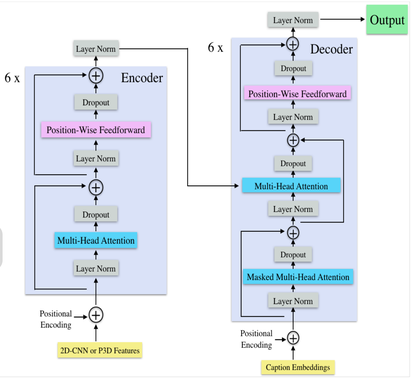


2. P3D Encoder + LSTM Decoder



3. Combined P3D/ResNet-LSTM Encoder with LSTM Decoder

4. Transformer with ResNet Features



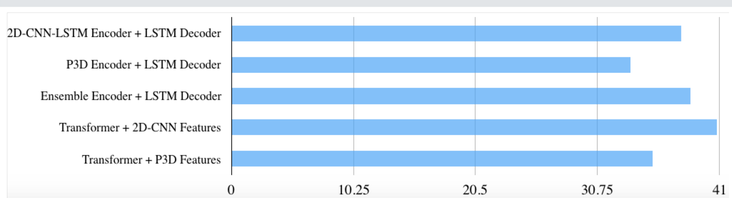
5. Transformer with P3D Features

Additive Attention

$$e_i = v^T tanh(W_1 h_i + W_2 s) \in R$$

Results

BLEU Score	2D-CNN-LSTM Encoder + LSTM decoder	P3D Encoder + LSTM decoder	Ensemble Encoder + LSTM decoder	Transformer + 2D-CNN Features	Transformer + P3D Features
Train	71.85	64.11	66.73	61.47	54.25
Test	37.75	33.51	38.55	40.78	35.4



Of the 5 models we experimented with, Transformers with ResNet features performed best, followed by combined P3D/ResNet features with an LSTM.

Analysis

BLEU Score	2D-CNN-LSTM Encoder + LSTM decoder	P3D Encoder + LSTM decoder	Ensemble Encoder + LSTM decoder	Transformer + 2D-CNN Features	Transformer + P3D Features
"Man"	38.75	36.5	38.74	44.6	37.27
"Woman"	32.04	23.18	34.59	33.68	22.48

- Skew in dataset for "man" vs "woman" (541 vs. 152 in training and 291 vs. 87 in validation)
- We compared performance on subset that included "man" or some close variation, but not "woman", and vice versa
- Performance significantly better for "man" subset
- Particularly large discrepancy for P3D – possible that it does not discern "man" vs. "woman" since it was pre-trained on activity recognition

Main Findings & Future Work

Main findings:

- Attention over P3D features did not perform as well as 2D-CNN features -- pre-training on action recognition may limit what features are captured, as shown by "man" vs "woman" analysis
- Ensemble model shows promise in providing different, but relevant, information for video captioning
- Transformers outperformed LSTMs

Future Work:

- Further experiment with combining P3D features with LSTM features, such as with Transformer
- Substitute ResNet-152 with Faster-RCNN for feature extra
- Compare with other datasets – image features seem to be sufficient for the MSVD dataset, but what about others?
- Incorporate additional features, such as optical flow, as done in other works
- Look into ways to address bias towards more frequent words