Video-Based Activity Search with Siamese Sequence Models

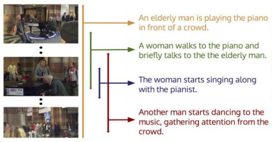
Milind Jagota, Erik Jones, Zhaoyu Lou; {mjagota, erjones, zlou}@stanford.edu

Abstract

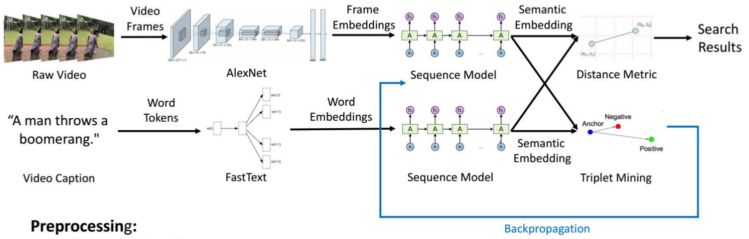
We present a video clip search platform which matches the semantic information in a description to the semantic content contained in a video clip. We train two distinct GRUs in a Siamese network to learn a semantic embedding space for the videos and captions. We use a triplet-based approach for one shot learning. After training, we obtain a top-20% performance of 99.5% on our validation. Our evaluation is based on the search ranking of the ground truth clip for a caption, relative to all other clips. Our search ranking is based on a metric between word and clip vectors in semantic space. This result demonstrates the feasibility of successful clip search engines in the future.

Data and Features

We used the ActivityNet dataset. Our data consisted of 37,000 video clips and associated descriptions from the Stanford Vision Lab, representing 849 hours of video and 203 different activity types. Video frames were converted to feature embeddings using AlexNet, and descriptions were converted to word embeddings using Facebook's FastText model. We used a subset of 10,000 video clips and their captions for training, and a separate subset of 1,000 clips for validation.


An elderly man is playing the piano in front of a crowd.
A woman walks to the piano and briefly talks to the the elderly man.
The woman starts singing along with the pianist.
Another man starts dancing to the music, gathering attention from the crowd.

System Design and Models



Preprocessing:

- Video sliced into frames
- 500 dimensional frame embeddings extracted from last layer of AlexNet
- Caption tokenized into words
- 300 dimensional word embeddings
- FastText skip-gram model

Architecture:

- Siamese sequence models
- Learns mapping to semantic space
- RNN, LSTM, Bi-LSTM, GRU
- 400 dimensional hidden state

Learning:

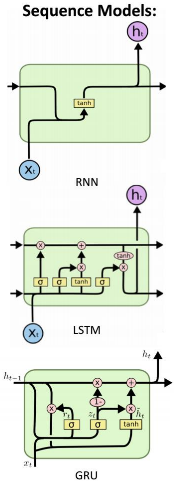
- Triplet-based one shot learning
- Online generation of hard triplets
- Clip and caption used as anchor and positive
- All negatives resulting in hard triplets used

$$Loss = \sum_{i=1}^{N} \left[\|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2 + \alpha \right]_+$$

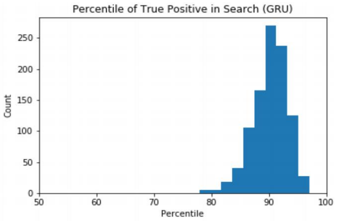
Evaluation:

- Video semantic embeddings saved as matrix
- Query converted to semantic embedding vector
- Outputs argmax of matrix-vector product

Sequence Models:



Results



Model	Top-20%	Top-10%	Median
RNN	0%	0%	52.3%
LSTM	97.3%	53.8%	90.5%
Bi-LSTM	94.6%	41.8%	89.2%
GRU	99.5%	56.6%	90.5%

Evaluation Example:

Query: A man … trying to shovel snow…

Captions of Resulting Clips:

1. A man breaks up ice with a spade…
2. … woman continues taking snow …
3. A man is seen kneeling off a roof…
4. A man shovels snow…

Our best model was the GRU-based network, which had the true positive in the top 20% returned for almost every caption.

Note that these metrics underestimate the true accuracy, since many clips besides the one true positive are semantic matches.

Discussion and Future

Our model achieves remarkable accuracy, showcasing the feasibility of a text-to-video search engine. The LSTM and GRU considerably outperformed the vanilla RNN, demonstrating their power in capturing long term dependencies in the data.

An interesting result we found was that increasing the triplet margin monotonically increased network performance within our testing range. This suggests that using all triplets rather than only hard triplets may be better for this task.

More generally, our problem differs from the person re-identification task that triplet loss was developed on. Unlike faces, which either match or don't, two sentences can be semantically similar without being identical. This is evidenced in our results; the returned rank of the ground truth is often lower than similar clips. We are curious about the efficacy of triplet loss in learning over such pairs.

Lastly, it would be interesting to integrate the model into a end to end clip search engine. This would likely require training on a larger and more diverse dataset.

References

[1] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606.*

[2] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017, October). Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision* (Vol. 1, No. 2, p. 6).

[3] Olah, C. (2015, August 27). Understanding LSTM Networks. Retrieved March 19, 2018, from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[4] Heilbron, F. C., Escorcia, V., Ghanem, B., & Niebles, J. C. (2015, June). Activitynet: A large-scale video benchmark for human activity understanding. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 961-970). IEEE.