Literature Notes

Adam Lehavi
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089
alehavi@usc.edu

***Abstract*—This outlines Adam Lehavi's notes related to his research with Yoonsoo Nam.**

I. Working Model Idea

Take a video, use [1] to turn the video into key frames, use image summarization as surveyed in [2] to generate image summaries, and use an NLP model to turn a series of image summaries to a description.

II. Core Definitions and Acronyms

- RNN - Recurrent Neural Network
- mRNN - multimodal RNN
- RCNN - Region Convolutional Neural Network
- BRNN - Bidirectional RNN
- Semantics - referring to the arrangement and relation of words in sentence formation

III. Literature Review

A. Survey of Techniques for Labeling Video, Audio, and Text Data

Past approaches to video data:

1) Automatically annotate news videos through unsupervised learning using mining of similar videos. For a given video with a speech-recognized transcript, it first searches and ranks most similar videos and then mines those.
2) Automatically annotate of people passing surveillance cameras, annotating clothing color, height, and focus of attention.
3) Use TREC video retrieval evaluation.
4) Address the problem of automatic temporal annotation of realistic human actions in video.
5) Use an optimal graph (OGL) from multicues (partial tags and multiple features) to get results superior to other state-of-the-art methods.
6) Propose Interactive Self-Annotation framework, based on recurrent self-supervised learning.
7) State a recursive and semi-automatic annotation approach which proposes initial annotations for all frames in a video based on segmenting only a few manual objects.
8) Use Multiple Annotation Maturation (MAM)
9) Propose a novel Correlative Multi-Labe (CML) framework which simultaneously classifies concepts and models correlations in a single step on the TRECVID dataset

[3]

Note a lot of the papers are from the ACM international conference on Multimedia.

B. Video Summarization Based on ML

Process to summarize a video in terms of some smaller amount of frames/images [1]

C. Survey for Automatic Description Generation from Images

Datasets

- Pascal1K
- Flickr8K
- Flickr30K
- MS COCO

Recommendation is MS COCO, having 164K images each with 5 texts, collected judgements, and partial objects.

To read:

1) Karpathy and Fei-Fei (2015) MultRetrieval on Flickr8K/30K, COCO measured w/ BLEU, Meteor, CIDEr, mRank, R@k
2) Jia et al. (2015) Generation on Flickr8K/30K, COCO measured w/ BLEU, Meteor
3) Yagcioglu et al. (2015) VisRetrieval on Flikr8K/30K, COCO measured w/ Human, BLEU, Meteor, CIDEr

[2]

D. Deep Visual-Semantic Alignments for Generating Image Descriptions

Model inputs an image, and generates a space of objects and their regions which is then used to generate a description. Trained by having images and descriptions, and aligning objects in the description to regions. This data is then used to train a mRNN to generate the snippets.

Sentences written by people make frequent references to some particular but unknown location in the image.

Images detected with a RCNN pretrained on ImageNet and its Detection Challenge.

Word representations computed using BRNN.

Overall good results.

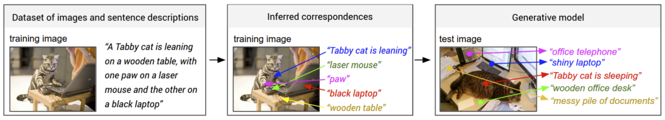
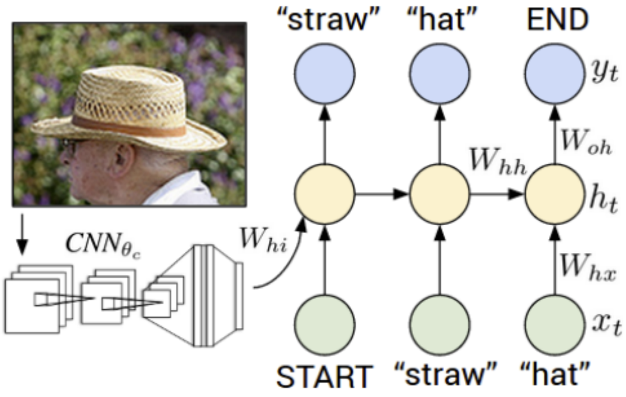


Figure 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle, Section 3.1) and then learns to generate novel descriptions (right, Section 3.2).

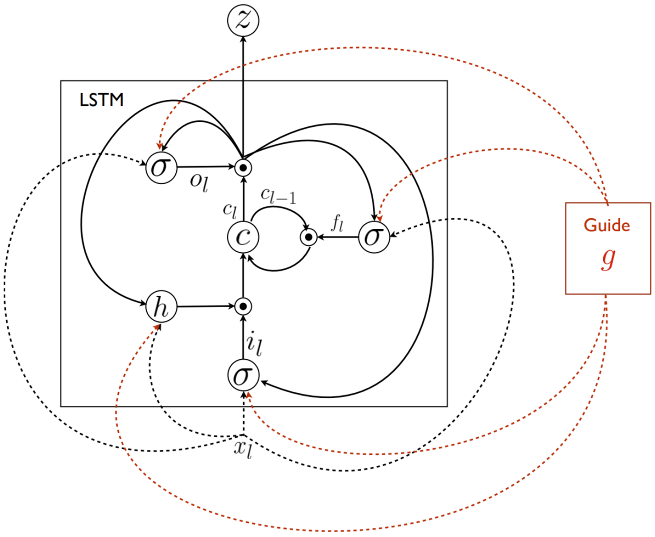


[4]

E. Guiding LSTM for Image Caption Generation

Use a CNN for encoding and an LSTM for the decoding step, but with some guiding factor, leading to what they attempt to coin as a gLSTM.

Hard-Attention seems to perform better on COCO.



[5]

F. Deep Learning for Video Captioning

Video V gets decomposed to frames, as well as motion, audio, and semantics. Sentence Y gets decomposed to words, so feature extraction is as such:

$$V = \{f_1, f_2, \ldots, f_N\} \tag{1}$$

$$Y = \{y_1, y_2, \ldots, y_T\} \tag{2}$$

$$\mathcal{F} = \{F_V, F_M, F_A, F_S\} \tag{3}$$

$$\mathcal{F} = f_{feat}(V) \tag{4}$$

$$F_t = f_{aggr}(\mathcal{F}, s_t) \tag{5}$$

Possible idea is to replace V with key frames from [1] and see if using this can improve training / results.

Extraction:

- Visual: CNNs like VGG and Inception Networks
- Motion: 3D CNNs from spatiotemporal features
- Audio: Mel Frequency Cepstral Coefficients (MFCC) used to get features, and bag-of-audio-words acquire fixed length audio features
- Semantic: video-level category information

Aggregation:

- LSTM/GRU is simplest since modalities have variate length.
- Temporal Attention: hLSTM
- Spatial Attention: MAM-RNN considers prior frames for spatial attention and SAM distinguishes foreground and background
- Multimodal Feature Fusion: Rarely explored, MMVD, AttenstionFusion, and MA-LSTM

Scoring:

- BLEU: used in past, not differentiable
- CIDEr: based on n-grams like BLEU
- SPICE: shown to generate better captions when used with CIDEr

Datasets:

- MSR-VTT contains 10,000 clips from 20 categories, with the most descriptions and largest vocab size. Recommended.

Method	T	B@4	M	C
MMVD [Ramanishka *et al.*, 2016]	M	40.7	28.6	46.5
Attention Fusion [Hori *et al.*, 2017]	M	39.7	25.5	40.0
MA-LSTM [Xu *et al.*, 2017]	M	36.5	26.5	41.0
HACA [Wang *et al.*, 2018c]	M	43.4	**29.5**	49.7
Temporal Att. [Yao *et al.*, 2015]	A	34.8	25.1	36.7
hLSTMat [Song *et al.*, 2017]	A	38.3	26.3	-
MGSA [Chen and Jiang, 2019a]	A	**45.4**	28.6	50.1
LSTM-E [Pan *et al.*, 2016]	S	36.1	25.8	38.5
M&M TGM [Chen *et al.*, 2017]	S	44.3	29.4	49.3
RL Ent [Pasunuru and Bansal, 2017]	R	40.5	28.4	**51.7**

[6]

References

[1] W. Ren and Y. Zhu, "A video summarization approach based on machine learning," in *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2008, pp. 450–453.

[2] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[3] S. Zhang, O. Jafari, and P. Nagarkar, "A survey on machine learning techniques for auto labeling of video, audio, and text data," *CoRR*, vol. abs/2109.03784, 2021. [Online]. Available: https://arxiv.org/abs/2109.03784

[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[5] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2407–2415.

[6] S. Chen, T. Yao, and Y.-G. Jiang, "Deep learning for video captioning: A review." in *IJCAI*, vol. 1, 2019, p. 2.