11-775 Project Proposal: Localizing Moments in Video with Natural Language

Anusha Prakash
anushap@andrew.cmu.edu

Anuva Agarwal
anuvaa@andrew.cmu.edu

Bhavya Karki
bkarki@andrew.cmu.edu

Pravalika Avvaru
pavvaru@andrew.cmu.edu

Abstract

We propose to build a system that retrieves specific temporal segments from a video given a natural language text description. This system can be applied to a wide range of applications like retrieving unforgettable memories from a long personal holiday video , searching for educational videos from a library of videos etc. We want to implement the video-language model in [1] to identify moments in a video in which referring expressions and video features from corresponding moments are mapped using the same embedding space. Improving the joint video - language model, up-scaling the vocabulary of referring expressions it supports and extending the algorithm to work on longer videos make this project challenging and interesting.

1 Introduction

Retrieving relevant videos given a natural language have been long studied. Whereas, retrieving a temporal segment of a video given a natural language description of the segment is quite challenging and new. Requiring to understand both language and video makes the task more difficult. Consider an example: A baby stands up, falls down and again stands up to start walking. If we want to infer to a particular temporal segment - "when the baby resiliently started walking again after falling down", it's not enough to use action, object or attribute keyword. Instead, moment must be defined by when and how specific actions are affected and dependent to other actions in the video. Such moments will identify the start and end points within the video which captures the temporal moment. Most of th existing methods [2, 3] retrieve videos using natural language but they don't identify the moment within the video. Also, most of the related work deal with action recognition in a themed set of videos such as a particular sport, or baking, or any particular activity. To localize moments in a generic set of videos, [1] develop a novel video-language model which uses global video features, specific moment features along with language features. They use a Moment Context Network (MCN) to include the global video feature to provide temporal context and a temporal endpoint feature to indicate the moment in the video. They introduce the Distinct Describable Moments (DiDeMo) dataset which consists of over 40,000 pairs of localized video moments and corresponding natural language. We will be using this paper as the baseline and will work on enhancing the solution through a better architecture. We wish to extend this algorithm to work on longer videos and larger vocabulary as well, which is not effectively tackled in this paper. Also, we will explore using a different set of visual features apart from the features this paper was tested on.

2 Baseline or Initial Experiments

For the baseline, we intend to implement the Moment Context Network (MCN) described in [1]. It relies on local and global features. We plan to extract visual temporal context features which encode the video moment by integrating both local features, which depicts what occurs within a specific moment, global video context, which provide context for a video moment and temporal endpoint features, which indicate when a moment occurs within the video.

We plan to employ Deep CNNS to extract the high-level video features for each video keyframe and then use this to extract the local and global features. RGB frames and optical flow frames are the two sources of visual input modalities we plan to explore. Temporal context features are extracted by inputting local video features, global video features, and temporal endpoint features into Neural Network.For a given natural language description, we extract the language features using an LSTM network. The Joint Video and Language Model is the sum of squared distances between embedded appearance, flow, and language features.

We plan to employ the Ranking loss to train our model and the Distinct Describable Moments (DiDeMo) dataset the paper proposes to train and evaluate our model.

3 Proposed Approach

Our goal is to improve the joint video - language model through a better architecture and extending the algorithm to work on longer videos. Our extended goal would be to up-scale the vocabulary of referring expressions the system supports. This section is divided into the following subsections to neatly chart out the different phases in this study.

3.1 Data Pre-processing

As in the paper, we will split the DiDeMo videos into training (8,395), validation (1,065), and testing (1,004) sets. Videos from a specific Flickr user only appear in one set. For recognition, we will experiment with various optimizers such as SGD, Adam, etc. Candidate moments come from the temporal segments defined by the gifs used to collect annotations. A 30 second video is to be broken into six five-second gifs. Moments can include any contiguous set of gifs, so a 30-second video contains 21 possible moments.

3.2 Models and Ablation Study

We plan to employ LSTM, GRU and Hierarchical RNN approaches to improve the architecture to enhance the model performance. We also plan to perform comparative studies in terms of time vs accuracy trade-off upon varying the layers, hyper-parameters in the network architecture.

Another area we intend to address is the joint-embedding space of the video and language model. We want to try out different distance metrics to build the joint model.

We want to use different features other then the RGB and optical flow and explore what each feature is trying to capture and explore the best combinations of features.

To address up-scaling the vocabulary part, we plan to pre-train on Moments in Time [4] dataset and select relevant actions and find a common embedding space. We want to explore a different linguistic embedding space and perform semantic word2vec matching and perform ablation studies on the same.

3.3 Risks and Pitfalls

The proposal has been drafted keeping the various pitfalls that we could encounter while completing this project. Given the huge collection of videos, we may choose to reasonably scale down on the size according to infrastructure and time constraints. Also, the feasibility of an approach could be determined by its training time, and if not feasible we would choose an alternative method of

performing the experiment. The availability of the moments-in-time dataset could be in question since the dataset is yet to be released as a part of the ActivityNet 2018 challenge and the usage requires an application that we have sent out.

We do not guarantee a performance improvement with the ablation studies and experimenting with the models. However, having conducted the experiments the performance of the features and models will provide a strong proof as to the plausibility of using that approach for extracting moments from the DiDeMo dataset.

3.4 Evaluation Metrics

We will use the same metrics to evaluate our work as the DiDeMo paper. They measure the performance of each model with Rank@1 (R@1), Rank@5 (R@5), and mean intersection over union (mIoU). They compute the score for a prediction and each human annotation for a particular description/moment pair instead of consolidating all human annotations into one ground truth.

4 Related Work

Over recent years, there has been an increased interest in jointly modeling images/videos and natural language sentences. The challenge lies in encoding temporal aspects of the video, to find contextual relationship between frames and having to process sheer volume of the video data. [5] implemented joint models taking a combination of the simple average or weighted average over video FC-7 features, and Language LSTM or Word2Vec features on MS-COCO and LSMDC16 Movies dataset. [6] tackle the challenge of generating unambiguous descriptions of a specific object or region in an image, which can comprehend such an expression to infer which object is being described. For generating the descriptions, they use RNN to generate descriptive sentences and a bean search to find the most probable sentence that distinguishes the input region from other candidate regions. For the description comprehension part, they adopt a ranking-based approach to generate a set of region proposals and then ask the system to rank these by probability. [7] propose a Compositional Modular Networks (CMNs) - an end-to-end trained model that learns language representation and image region localization jointly. The model localizes a referential expression by grounding the components in the expressions and exploiting their interactions such that the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them. The score of a region is determined by simultaneously looking at whether it matches the description of the subject entity and whether it matches the relationship with another interacting object entity mentioned in the expression. [8] developed a model that works on a finer level and embeds fragments of images and fragments of sentences (typed dependency tree relations) into a common space. [9] introduce a neural network-based embedding models for video, sentence, and image inputs whose parameters can be learned jointly. They utilize web image search in a sentence embedding process to disambiguate fine-grained visual concepts. Another related work [10] encode temporal data using a new concept called rank pooling where they learn a ranking machine that captures the temporal evolution of the data and uses the parameters of the latter as a representation.

5 Data & Technical Requirements

A major challenge when designing algorithms to localize moments with natural language is that there is a dearth of large-scale datasets which consist of referring expressions and localized video moments. To mitigate this issue, the authors of [1] introduce the Distinct Describable Moments (DiDeMo) dataset which includes over 10,000 25-30 second long personal videos with over 40,000 localized text descriptions. They randomly selected over 14,000 videos from YFCC100M [11] which contains over 100,000 Flickr videos with a Creative Commons License.

Videos in DiDeMo represent a diverse set of real-world videos, which include interesting, distinct moments, as well as uneventful segments which might be excluded from edited videos. Because videos are curated from Flickr, DiDeMo reflects the type of content people are interested in recording and sharing. Consequently, DiDeMo is human-centric with words like “baby”, “woman”, and “man” appearing frequently. Since videos are randomly sampled, DiDeMo has a long tail with words like “parachute” and “violin”, appearing infrequently (28 and 38 times). Important, distinct

moments in a video often coincide with specific camera movements. For example, "the camera pans to a group of friends" or "zooms in on the baby" can describe distinct moments. Many moments in personal videos are easiest to describe in reference to the viewer (e.g., "the little boy runs towards the camera"). DiDeMo contains more sentences with temporal indicators than natural language object retrieval and video description datasets, as well as a large number of spatial indicators. DiDeMo has a higher percentage of verbs than natural language object retrieval datasets, suggesting understanding action is important for moment localization in video.

The code to access the data and run the models is available at `https://github.com/LisaAnne/LocalizingMoments`. Their annotations include descriptions which are temporally grounded in videos. Each video is split into 5-second temporal chunks. The first temporal chunk corresponds to seconds 0-5 in the video, the second temporal chunk corresponds to seconds 5-10, etc. The dataset can be accessed through json files which contain the following fields: annotation_id (annotation ID for description), description (description for a specific video segment), video (video name), download_link (a download link for the video), num_segments (some videos are a little shorter than 25 seconds, so were split into five temporal chunks instead of six) and times (ground truth time points marked by annotators).

As a part of one of our experiments, we will need to obtain the Moments-in-Time dataset [4]. The dataset is a large-scale human-annotated collection of one million short videos corresponding to dynamic events unfolding within three seconds trained on 339 classes. The dataset which we hope to obtain comprises of 200 classes, 100000 training videos, 20000 testing videos and 10000 validation videos. The advantage of training on these videos are the large coverage and diversity of events in both visual and auditory modalities.

We have been provided with AWS credits and with these credits we requested for GPU p2.xLarge (1 quad-core, GPU instance with 61 GB RAM that is priced at $0.900 per hour). We will request for more instances on a need basis.

6 Proposed Timeline

Keeping the time complexity in mind and experiments we intend to conduct over the course of this project, we propose the following timeline:

- Proposal : March 10th
- Baseline Implementation : March 17th
- Ablation study about embedding space : March 25th
- Ablation study with various features : March 31st
- LSTM approach : April 8th
- Hierarchical RNN model : April 16th
- Results and Analysis : April 20th
- Conclusion and Report : April 25th

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. *CoRR*, abs/1708.01641, 2017.

[2] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *CoRR*, abs/1609.08124, 2016.

[3] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. *CoRR*, abs/1505.01861, 2015.

[4] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*, 2018.

[5] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *CoRR*, abs/1609.08124, 2016.

[6] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283, 2015.

[7] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. *CoRR*, abs/1611.09978, 2016.

[8] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. *CoRR*, abs/1406.5679, 2014.

[9] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. *CoRR*, abs/1608.02367, 2016.

[10] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. Action recognition with dynamic image networks. *CoRR*, abs/1612.00738, 2016.

[11] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 1(8), 2015.