Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework

Ran Xu
Department of Computer Science
SUNY at Buffalo
rxu2@buffalo.edu

Caiming Xiong
Department of Statistics
UCLA
caimingxiong@ucla.edu

Wei Chen
Department of Computer Science
SUNY at Buffalo
wchen23@buffalo.edu

Jason J. Corso
Department of EECS
University of Michigan
jjcorso@eecs.umich.edu

Abstract

Recently, joint video-language modeling has been attracting more and more attention. However, most existing approaches focus on exploring the language model upon on a fixed visual model. In this paper, we propose a unified framework that jointly models video and the corresponding text sentences. The framework consists of three parts: a compositional semantics language model, a deep video model and a joint embedding model. In our language model, we propose a dependency-tree structure model that embeds sentence into a continuous vector space, which preserves visually grounded meanings and word order. In the visual model, we leverage deep neural networks to capture essential semantic information from videos. In the joint embedding model, we minimize the distance of the outputs of the deep video model and compositional language model in the joint space, and update these two models jointly. Based on these three parts, our system is able to accomplish three tasks: 1) natural language generation, and 2) video retrieval and 3) language retrieval. In the experiments, the results show our approach outperforms SVM, CRF and CCA baselines in predicting Subject-Verb-Object triplet and natural sentence generation, and is better than CCA in video retrieval and language retrieval tasks.

Introduction

More than 100,000 hours of videos are uploaded to YouTube everyday, and more than 100,000 new videos are added and shared in Facebook everyday. Most of those videos are paired with natural language descriptions, some of which are as simple as tags or as detailed as paragraphs. Those descriptions provide us with the possibility of joint video and human language understanding and thus support many promising applications, e.g. turning the surveillance video from last night into a page of incident report, teaching robots to recognize certain objects with human language, or recommending YouTube viewers more interesting video clips leveraging both text and video content analysis.

In such applications, there are three essential tasks, i.e. video retrieval (Song, Yang, and Huang 2011), language retrieval (Das, Srihari, and Corso 2013) and ultimately, natural language generation (NLG) from novel videos (Barbu et al. 2012; Das et al. 2013; Rohrbach et al. 2013; Gupta et al. 2009; Krishnamoorthy et al. 2013; Guadarrama et al. 2013; Thomason et al. 2014), as depicted in Fig. 1. All these problems have been pushed forward in the artificial intelligence, multimedia, computer vision, natural language processing and machine learning communities.

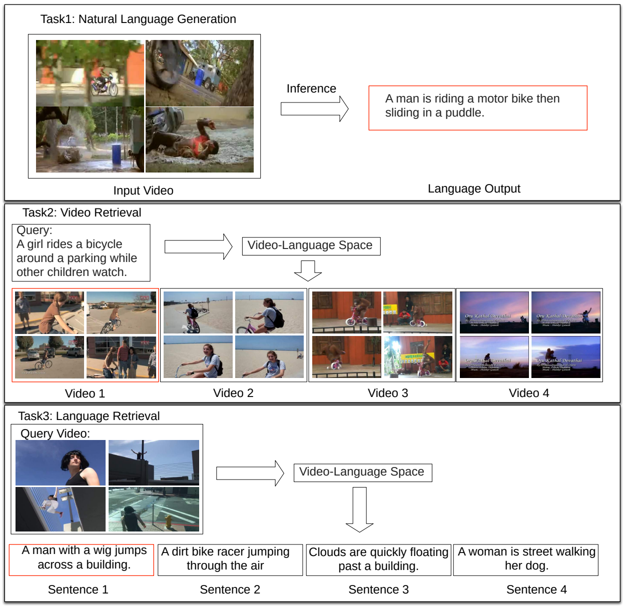

Task1: Natural Language Generation
Inference
A man is riding a motor bike then sliding in a puddle.
Input Video
Language Output
Task2: Video Retrieval
Query:
A girl rides a bicycle around a parking while other children watch.
Video-Language Space
Video 1
Video 2
Video 3
Video 4
Task3: Language Retrieval
Query Video:
Video-Language Space
A man with a wig jumps across a building.
A dirt bike racer jumping through the air
Clouds are quickly floating past a building.
A woman is street walking her dog.
Sentence 1
Sentence 2
Sentence 3
Sentence 4

Figure 1: Illustration of three important tasks, the upper box shows natural language generation, middle box shows video retrieval and bottom box shows language retrieval. The red box indicates the ground truth result.

Along the way towards these tasks, researchers have spent decades of efforts in video content understanding, including action classification (Wang et al. 2011; Sadanand and Corso 2012; Karpathy et al. 2014), detection (Tian, Sukthankar, and Shah 2013; Zhang, Zhu, and Derpanis 2013) and tagging (Yao et al. 2013; Moxley, Mei, and Manjunath 2010). We believe it is meaningful to push one step further and generate sentences for video because it is more natural to human perception and encodes spatio-temporal relationships and richer details from videos.

In transducing videos into sentences, a line of work has investigated marrying state-of-the-art computer vision (Li et

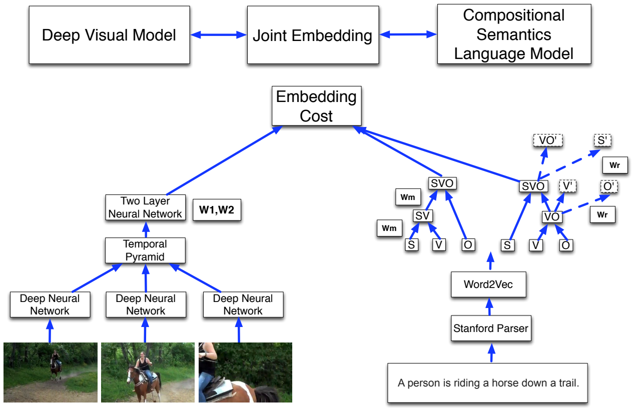


Figure 2: Overview of our unified framework with a joint Deep Video Model (Left) and a Compositional Language Model (Right).

al. 2011; Wang et al. 2011; Felzenszwalb et al. 2010) and natural language processing (Miller 1995) techniques. For example, (Guadarrama et al. 2013) use semantic hierarchies to trade-off the specificity and accuracy of subject, verb and object in YouTube videos "in-the-wild." (Rohrbach et al. 2013) propose a CRF to model relationships between different visual components and translate video to text in cooking videos. (Kulkarni et al. 2011) proposes a CRF model to incorporate object, attribute and preposition, where pair-wise relationships of object-attribute, object-preposition, etc. are modeled with a large text corpus. In the Experimental section, our baseline methods that use SVM and CRF follow this paradigm. A common methodology in the above models is to build the language model directly upon the output of the visual model without feedback to the visual model. Alternatively, another line of promising work builds a joint space, e.g., (Nakayama, Harada, and Kuniyoshi 2010) propose Canonical Contextual Distance (CCD) to construct image-text space and use KNN for image annotation, (Socher et al. 2013) use Recursive Neural Network to model a sentence and build an image-text space to do image and sentence retrieval. As (Das et al. 2013) discussed, the scalability of non-parametric methods with increasing of semantic space is unclear, and the generated text often lacks "semantic verification." In our efforts towards joint video-language modeling, we make two observations:

- Commonly used word similarity captures more syntactic expression than visually grounded semantics, e.g., in WordNet, the Lesk similarity between cat and kitten is 0.4 while the similarity between cat and dog is 1.04.
- Visually grounded semantics is highly compositional, such as "player rides bicycle" and "cook bakes bread", and it is meaningful to jointly learn such compositionality and video representation.

Inspired by such observations, we propose a unified framework with joint deep video and compositional language models to address the above points. Our framework consists of three parts: a compositional semantics language model, a deep video model and a joint embedding model.

Firstly, we propose a compositional semantics language model that enforces semantic compatibility between essential concepts, especially visually meaningful concepts in videos. We assume that the essential semantic meaning of a video can be captured by <Subject, Verb, Object> triplet. First, we parse natural sentence descriptions into SVO triplets that represent each subject, verb and object respectively, then we leverage a continuous language model (Mikolov et al. 2013) to represent each element of SVO with a continuous vector. Based upon the initial word vector, we construct our language model in the dependency-tree structure. The right side of Fig. 2 shows our model, S, V and O are leaf nodes, and there are two structures in our particular problem. To compose leaf nodes towards higher layers, a composition function is applied to two leaf nodes, note that weight W_m in Fig. 2 is recursively used until the root node is composed. Therefore, the root node is the representation of an SVO triplet in video-text space. With our language model, the compositionality of nodes can be explicitly modeled by the weight of the composition function.

Secondly, inspired by current advances of deep learning (Donahua et al. 2013; Krizhevsky, Sutskever, and Hinton 2012), we present a deep video model: as Fig. 2 shows, visual features are extracted with a deep neural network (Donahua et al. 2013) from a sequence of frames of each video, we use a temporal pyramid scheme to capture motion information, then build a two-layer neural network to map visual features to video-text space. $W1$ and $W2$ are weights in the two-layer neural network.

Finally, we propose the joint embedding model that minimizes the distance of the outputs of the deep video model and compositional language model in video-text space, and update these two models jointly in the unified framework.

Related Work

Recently, there are many new works related to our problem. In this section, we mainly review two relevant topics, 1) video to text and 2) multi-modal embeddings.

Video to text In describing videos with sentences, (Krishnamoorthy et al. 2013; Motwani and Mooney 2012) are earlier papers that focus on smaller data sets. (Barbu et al. 2012) build a system that leverages on object detection, object tracking and human pose estimation in order to render linguistic entities and generate sentences. (Ramanathan, Liang, and Fei-Fei 2013) use a topic-based semantic relatedness measure to help action and role classification, (Yu and Siskind 2013) use HMMs to track sentences and learn the representation for word meanings with the help of video clips. The above two papers separately show language helping vision tasks and vision helping language tasks, but not a joint model to push the other direction.

Multi-modal embeddings The only paper we aware of that constructs multi-modal space for video and language is (Das, Srihari, and Corso 2013), which uses latent topics to discover the most probable related words from text and furthermore, translate words to probable frames, while bounding box annotation of objects are needed. Additionally, there exist a number of papers on image and language embedding:

(Frome et al. 2013) propose a deep visual-semantic embedding model to bridge image and text. (Socher and Fei-Fei 2010) present a semi-supervised model to segment and annotate images that finds mapping between segment-wise visual words and textual words based on kCCA.

Our approach differs from above methods in that we explore the compositionality of word relations within the joint model and we can also infer the representation of words due to our tree structured language model and SVO assumption.

An Unified Framework with Joint Video-Language Model

In this section, we propose a joint model for video-language embedding. In our joint architecture, the goal is to learn a function $\mathbf{f}(\mathcal{V})$,

$$\mathbf{f} : \mathcal{V} \rightarrow \mathcal{T} \tag{1}$$

where $\mathcal{V}$ represents the low-level features extracted from video, and $\mathcal{T}$ is the high-level text description of the video. Since video $\mathcal{V}$ represents the low-level signals and language $\mathcal{T}$ shows high-level human expression, we need a bridge to connect these two levels of information. Thus, we propose a joint model $\mathcal{P}$ which consists of three parts: compositional language model $M_L : T \rightarrow T_f$, deep video model $M_V : V \rightarrow V_f$ and an joint embedding model $E(V_f, T_f)$, such that

$$\mathcal{P} : M_V(V) \longrightarrow V_f \leftrightarrow E(V_f, T_f) \leftrightarrow T_f \longleftrightarrow M_L(T) \tag{2}$$

where V_f and T_f are the output of the deep video model and compositional language model respectively. Using our joint model $\mathcal{P}$ with three parts, video and corresponding language descriptions can be integrated into our unified framework. Next, we explain each part of the joint model in detail.

Compositional Semantics Language Model

The sentence description generated by a human is a unique and important resource in video or image to text. Generally, researchers consider each word in text description as a discrete index: (Guadarrama et al. 2013) use distributional clustering (Pereira, Tishby, and Lee 1993) based on word co-occurrence in particular syntactic contexts to build word hierarchy, (Das et al. 2013) use topic model to verify predicted words. We use a continuous space word representation (Mikolov et al. 2013) to initialize because it captures a large number of syntactic and semantic word relationships. For example, with such representation, $vec(\text{“}Germany\text{”})$ + $vec(\text{“}capital\text{”})$ is close to $vec(\text{“}Berlin\text{”})$. Due to the great variance of human input sentences, a continuous word vector is more suitable to capture semantic similarity.

First, we use the Stanford Parser (Klein and Manning 2013) to parse all sentences, then we choose $nsubj$ to find subject-verb pair, and choose $dobj$ to find verb-object pairs in order to obtain SVO triplets. Then, each word in SVO is mapped to a continuous d-dimensional vector by model $M_L^1(\cdot)$ which is the same as (Mikolov et al. 2013),

$$M_L^1 : T \longrightarrow [m_s, m_v, m_o] \tag{3}$$

where m_s, m_v and m_o are corresponding continuous word vector of SVO. To obtain an embed feature representation for SVO, we propose a novel compositional language model $CLM(\cdot)$ with recursive neural network as in Fig. 2.

In the compositional language model $CLM(\cdot)$, the goal is to learn a representation of $[m_s, m_v, m_o]$ for sentences. Assume we are given the continuous vectors $[m_s, m_v, m_o]$ of SVO for sentence, first, infer the representation for SV as m_{sv}: $([m_s, m_v] \rightarrow m_{sv})$, then combining m_{sv} with m_o, obtain the embedded language representation m_{svo}: $([m_{sv}, m_o) \rightarrow m_{svo})$ as the sentence representation T_f in embedding model.

Concretely, we adopt the Forward Propagation algorithm to infer the sentence representation, as follows.

As in Fig. 2, the SVO triplet is <person, ride, horse>, and corresponding word vectors are denoted as m_s, m_v and m_o, our model compute parent word vector from dependent child nodes with a composition function f parameterized by W_m, e.g., we can compute the parent node m_{sv} via:

$$m_{sv} = f(W_m[m_s; m_v] + b_m) \tag{4}$$

where $W_m \in \mathbf{R}^{d \times 2d}$ is a parameter matrix, and $b_m \in \mathbf{R}^d$ is bias. In all experiments $d = 300$. Similarly, we apply the composition function recursively and compute the parent node of m_{sv} and m_o via:

$$m_{svo} = f(W_m[m_{sv}; m_o] + b_m) \tag{5}$$

We use $tanh(\cdot)$ as composition function $f(\cdot)$, since $tanh(\cdot)$ performs well in most deep neural network.

To measure how well a parent node can represent the child nodes, we reconstruct the child nodes with a matrix $W_r \in \mathbf{R}^{2d \times d}$ and reconstruct m_s^{rec} and m_v^{rec} with:

$$[m_s^{rec}; m_v^{rec}] = W_r m_{sv} + b_r \tag{6}$$

$$[m_{sv}^{rec}; m_o^{rec}] = W_r m_{svo} + b_r \tag{7}$$

The reconstruction error of one non-terminal node p with our compositional language model is:

$$\begin{aligned} E_{rec}(p|W_m, W_r) = & \frac{n_1}{n_1 + n_2} \|m_1 - m_1^{rec}\|_2^2 \\ & + \frac{n_2}{n_1 + n_2} \|m_2 - m_2^{rec}\|_2^2 \end{aligned} \tag{8}$$

where m_1 and $m2$ are children of p, m_1^{rec} and m_2^{rec} are reconstructed word vectors of m_1 and $m2$. n_1 and n_2 are number of nodes in both branches under p, which are used to weigh the branch with more children higher.

Deep Video Model

Our visual model $M_V(V)$ applies features from a deep neural network (Donahua et al. 2013) trained with ImageNet (Deng et al. 2009). We extract one frame per second from video and compute a 4096-dimensional feature per frame, and find 7th layer output after ReLU performs the best. Then, we apply a temporal pyramid scheme to summarize the feature sequence and capture motion information.

Denote visual feature as x, we train a two-layer neural network to project x to embedding space:

$$V_f(x) = W_1 f(W_2 x) \tag{9}$$

We use standard non-linear function $tanh(\cdot)$ as $f(\cdot)$, $W_2 \in \mathbf{R}^{h\times 4096}$ and $W_1 \in \mathbf{R}^{300\times h}$ are parameter matrix that map $x \in \mathbf{R}^{4096}$ to joint space. We tested different h and find it insensitive, in our experiments we set h as 1000. We have tested single layer neural network and single linear mapping, and find two-layer neural network works the best. Our baseline method with CCA confirms a similar finding.

Joint Embedding Model

The Compositional Semantics Language Model captures high-level semantics information that can help constrain the visual model, and the visual model on the contrary, provides video evidence to support word selection. Most existing methods focus on the second bottom-up path using visual evidence to support word generation, while the first top-down path is largely neglected. In our joint embedding model, we define an objective function to take into account video-language embedding error E_{embed} and language model reconstruction error E_{rec}. E_{embed} is based on least squares to implement both bottom-up and top-down paths simultaneously:

$$\begin{aligned} &E_{embed}(V,T) = \\ &\|W_1 f(W_2 x_i) - CLM(m_{s,i}, m_{v,i}, m_{o,i}|W_m)\|_2^2 , \end{aligned} \tag{10}$$

where $m_{s,i}$ represents S word vector of i-th video. The objective function is:

$$\begin{aligned} J(V,T) =& \sum_{i=1}^{N} (E_{embed}(V,T) + \sum_{p\in\mathbf{NT}} E_{rec}(p|W_m, W_r)) \\ &+ r. \end{aligned} \tag{11}$$

where $\mathbf{NT}$ is the non-terminal set of one tree structure. Suppose our training set contains N videos, each paired with M sentences, and each SVO triplet has t tree structures. Let θ be a general notation of model W_1, W_2, W_m or W_r, the regularization term $r = \lambda/2\ \|\theta\|_2^2$. In practice, we use the mean word vector over all sentences as ground truth in each video to represent the training sentence.

According to this embedding model, we update the visual model and the compositional language model jointly. And we implement both bottom-up and top-down paths in the embedded model, thus our model is able to process both video-to-text and text-to-video tasks.

Learning and Inference

Learning the Joint Video-Language Model

To estimate parameters W_1, W_2, W_m and W_r, we initialize m_s, m_v and m_o with word vectors, W_1, W_2 and W_m, W_r with random zero-mean matrix normalized by column dimension of the matrix. We propose a coordinate descent method to optimize the objective function. To start, we first fix W_m, W_r and optimize W_1 and W_2, the gradient is:

$$\begin{aligned} grad_{W_1} = \frac{1}{N}\sum_{i=1}^{N} & 2(W_1 f(W_2 x_i) - \\ & CLM(m_{s,i}, m_{v,i}, m_{o,i}|W_m)) df(W_2 x_i)^\intercal + \lambda W_1. \end{aligned} \tag{12}$$

$$\begin{aligned} grad_{W_2} = \frac{1}{N}\sum_{i=1}^{N} & 2W_2^\intercal (W_1 f(W_2 x_i) - \\ & CLM(m_{s,i}, m_{v,i}, m_{o,i}|W_m)) df(W_2 x_i) x_i^\intercal + \lambda W_2. \end{aligned} \tag{13}$$

where $df(\cdot)$ is derivative function of $tanh(\cdot)$.

Then, we fix W_1, W_2 and optimize W_m, W_r. Starting from the top node, we use backward propagation (Goller and Kuchler 1996) through the tree structure to compute the gradient. In practice, we attach the bias term b_m, b_r with the W_m, W_r matrix and learn them together. We iteratively optimize the visual model and the language model with L-BFGS to minimize the objective function.

Inference

One advantage of our model is to infer the SVO representation and thus predict the SVO directly. Our first intuition is to consider each word itself as model and initialize a word vector of SVO randomly, but preliminary experiments do not show good results, so we leave it as future work and alternatively design a strategy to initialize SVO representation.

Given the visual model and the language model, we can describe a video with sentences by projecting a test video to the video-language space and finding the k nearest triplet vector m_{svo} in a large sentence pool. We build such a sentence pool from all sentence descriptions of training videos, then they are parsed with the Stanford Parser, mapped to word vector, and composed to m_{svo} via forward propagation with trained language model. We initialize the word representation with the most frequent word vector in top k neighbors of testing video. Then, we estimate m_s, m_v and m_o for test video by optimize objective function:

$$\begin{aligned} &(m_s, m_v, m_o) = \\ &\mathop{\arg\min}_{m_s, m_v, m_o} \|W_1 f(W_2 x) - CLM(m_s, m_v, m_o|W_m)\|_2^2 . \end{aligned} \tag{14}$$

We use forward-backward propagation to compute gradient of word vector and optimize with L-BFGS.

Experiments

In the following subsections, we introduce the experimental setup, three baseline methods, and evaluation results on three tasks: 1) SVO prediction and natural language generation, and 2) video retrieval and 3) language retrieval.

Experimental Setup

Data set The data set we use is YouTube videos collected by Chen et al. (Chen and Dolan 2011), which contains 1970 short video clips and paired with multiple language descriptions. We split the data set as 1297 training videos and 670 testing videos, same as (Guadarrama et al. 2013).

Defining Ground Truth for SVO. Given extracted sentences, we use the Stanford Parser to extract SVO triplet and use Porter Stemming algorithm to stem words. Then, we filter all labels that don't appear in the description of at least 5 videos, we further filter remaining labels that have no match to WordNet or pre-trained 3 million words and phrases provided by (Mikolov et al. 2013). Then, we compute Lesk similarity (Pedersen, Patwardhan, and Michelizzi 2004) for all remaining word pairs in SVO group. At last, we use spectral clustering on Lesk similarity to get 45 subject clusters, 218 verb groups and 241 groups. Words in same cluster are regarded as synonyms. We extract the "Most Common" SVO triplet from all candidate triplets in one video as ground truth.

Baseline Methods

To fully evaluate our model, we designed three non-trivial baseline methods, i.e. 1) SVM, 2) Conditional Random Field (CRF) model and 3) Canonical Correlation Analysis (CCA).

SVM We apply the state-of-the-art motion descriptor, i.e. dense trajectory (Wang et al. 2011) and object descriptor, i.e. ObjectBank (Li et al. 2011) as visual features. We random sample 100,000 visual features from each channel of dense trajectory descriptors and construct a 4000-dimensional codebook with K-means, then we encode each video as histogram of cluster centers. For object description, we use the ObjectBank default 177 object models trained from ImageNet and 20 object models trained from Pascal 2007 data set. Finally, we train a RBF kernel SVM with all visual features.

CRF In order to compare with our compositional semantics language model, we propose a CRF model to incorporate subject-verb and verb-object pairwise relationship. We minimize the following energy function over labeling L of video Vid and sentence Tex:

$$E(L;V,T) = \alpha_1\psi(S;Vid) + \alpha_2\psi(V;Vid) + \qquad (15)$$
$$\alpha_3\psi(O;Vid) + \alpha_5\psi(S,V;Tex) + \alpha_4\psi(V,O;Tex)$$

Unary potentials $\psi(S;Vid), \psi(V;Vid), \psi(O;Vid)$ are represented by probability outputs of SVM score over SVO. Pairwise potential $\psi(S,V;Tex), \psi(V,O;Tex)$ are learned from word co-occurrence. Specifically, for each subject, verb or object, we use the same training sentence pool described in Inference section to get all pairwise co-occurrence statistics. Gibbs Sampling is used as inference algorithm.

CCA CCA (Socher and Fei-Fei 2010) has been used to build an image-text space, we use CCA as a baseline to build the video-language space and compare video retrieval/text retrieval with our method. We use same deep video feature and average of word vector to learn the joint space. Then, we apply the same strategy as in the Inference section and use the mapped sentence representation to search the k nearest neighbors in large sentence pool and vote the most frequent triplets.

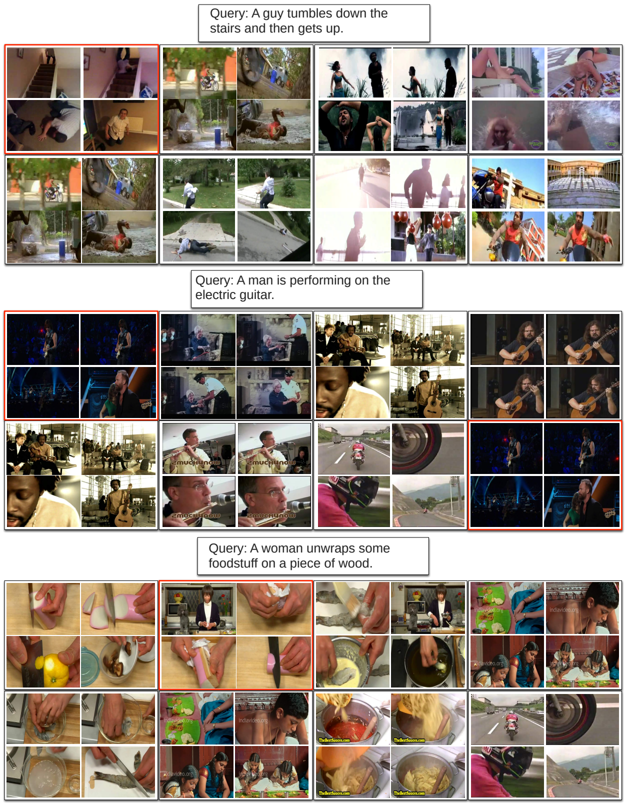

Query: A guy tumbles down the stairs and then gets up.
Query: A man is performing on the electric guitar.
Query: A woman unwraps some foodstuff on a piece of wood.

Figure 3: Video retrieval examples. For each query sentence, the top row shows top four videos retrieved with our method, the bottom row shows top four videos retrieved with CCA. Video in red bounding box indicates it is the ground truth video.

Table 1: SVO prediction accuracy with Prior, SVM, CRF, CCA, Our model

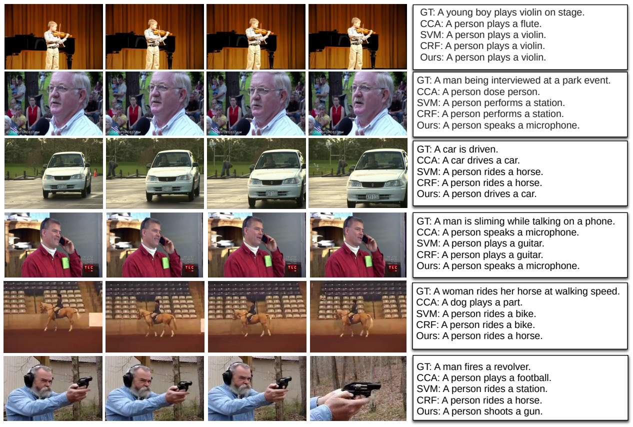
Method	Prior	SVM	CRF	CCA	Ours
S(%)	77.01	77.16	77.16	77.16	**78.25**
V(%)	14.63	22.39	22.54	21.04	**24.45**
O(%)	4.18	9.10	9.25	10.99	**11.95**

Results

SVO triplet prediction For SVO prediction, we use binary (0-1 loss) to measure the accuracy. Following (Guadarrama et al. 2013), we get 45 subjects, 218 verbs and 241 objects as ground truth classes. We evaluate 5 baseline methods, 1) "Prior" that uses the prior distribution of SVO triplet, which means simply find the most common SVO in training set as prediction results. 2) SVM, 3) CRF, 4) CCA, and our joint model.

From Table. 1, it is clear that our joint model outperforms all our baselines in predicting subject, verb and object. It demonstrates by capturing compositional semantics and building both top-down and bottom-up connection between video and language, our approach is able to show some advantages over pure bottom-up methods and multi-modal embedding without considering compositionality.

Natural Language Generation Fig. 4 shows sentences



GT: A young boy plays violin on stage.
CCA: A person plays a flute.
SVM: A person plays a violin.
CRF: A person plays a violin.
Ours: A person plays a violin.

GT: A man being interviewed at a park event.
CCA: A person dose person.
SVM: A person performs a station.
CRF: A person performs a station.
Ours: A person speaks a microphone.

GT: A car is driven.
CCA: A car drives a car.
SVM: A person rides a horse.
CRF: A person rides a horse.
Ours: A person drives a car.

GT: A man is sliming while talking on a phone.
CCA: A person speaks a microphone.
SVM: A person plays a guitar.
CRF: A person plays a guitar.
Ours: A person speaks a microphone.

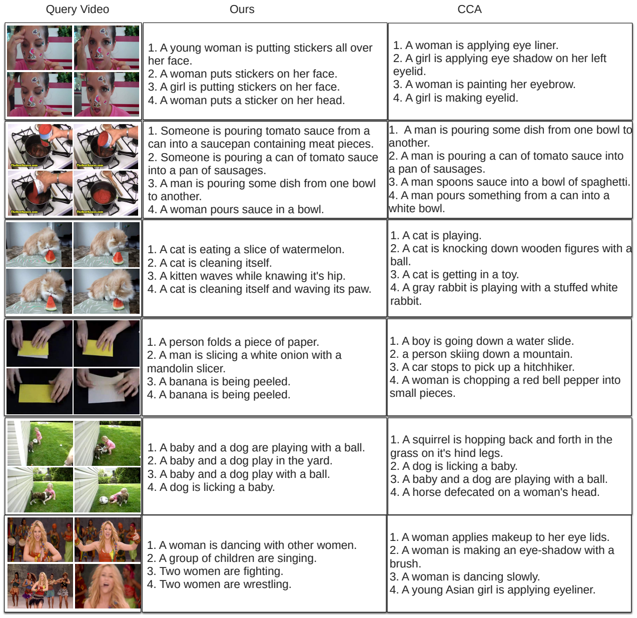
GT: A woman rides her horse at walking speed.
CCA: A dog plays a part.
SVM: A person rides a bike.
CRF: A person rides a bike.
Ours: A person rides a horse.

GT: A man fires a revolver.
CCA: A person plays a football.
SVM: A person rides a station.
CRF: A person rides a horse.
Ours: A person shoots a gun.

Figure 4: Natural language generation of different methods. The pictures are sampled video frames, the output sentence generated is by 1) GT: Human generated ground truth, 2) CCA baseline, 3) SVM baseline 4) CRF and 5) Our joint model.

Table 2: Video Retrieval and Text Retrieval, evaluated by Mean Rank (mRank)

Method	Video Retrieval mRank	Text Retrieval mRank
CCA	245.33	251.27
Ours	236.27	224.10



Query Video	Ours	CCA
	1. A young woman is putting stickers all over her face. 2. A woman puts stickers on her face. 3. A girl is putting stickers on her face. 4. A woman puts a sticker on her head.	1. A woman is applying eye liner. 2. A girl is applying eye shadow on her left eyelid. 3. A woman is painting her eyebrow. 4. A girl is making eyelid.
	1. Someone is pouring tomato sauce from a can into a saucepan containing meat pieces. 2. Someone is pouring a can of tomato sauce into a pan of sausages. 3. A man is pouring some dish from one bowl to another. 4. A woman pours sauce in a bowl.	1. A man is pouring some dish from one bowl to another. 2. A man is pouring a can of tomato sauce into a pan of sausages. 3. A man spoons sauce into a bowl of spaghetti. 4. A man pours something from a can into a white bowl.
	1. A cat is eating a slice of watermelon. 2. A cat is cleaning itself. 3. A kitten waves while knawing it's hip. 4. A cat is cleaning itself and waving its paw.	1. A cat is playing. 2. A cat is knocking down wooden figures with a ball. 3. A cat is getting in a toy. 4. A gray rabbit is playing with a stuffed white rabbit.
	1. A person folds a piece of paper. 2. A man is slicing a white onion with a mandolin slicer. 3. A banana is being peeled. 4. A banana is being peeled.	1. A boy is going down a water slide. 2. a person skiing down a mountain. 3. A car stops to pick up a hitchhiker. 4. A woman is chopping a red bell pepper into small pieces.
	1. A baby and a dog are playing with a ball. 2. A baby and a dog play in the yard. 3. A baby and a dog play with a ball. 4. A dog is licking a baby.	1. A squirrel is hopping back and forth in the grass on it's hind legs. 2. A dog is licking a baby. 3. A baby and a dog are playing with a ball. 4. A horse defecated on a woman's head.
	1. A woman is dancing with other women. 2. A group of children are singing. 3. Two women are fighting. 4. Two women are wrestling.	1. A woman applies makeup to her eye lids. 2. A woman is making an eye-shadow with a brush. 3. A woman is dancing slowly. 4. A young Asian girl is applying eyeliner.

Figure 5: Text retrieval examples. For each query video, the left and right columns show top four retrieved sentences using our model and CCA.

generated by different methods and human annotation. The qualitative results are basically consistent with quantitative results shown in Table. 1. We observe that our model and CCA tend to return similar results, meanwhile, SVM and CRF also tends to return similar results, which represents the distinct differences of those two types of methods. In the 3rd and 4th row, our model and CCA performs better in finding objects. We also find the CRF does capture some pairwise relationship among triplets, e.g. in 6th row, SVM returns in "rides a station" while CRF returns "rides a horse".

Video Retrieval In this task, we focus on evaluating how well a sentence can retrieve a video with corresponding semantic meaning. Firstly, for each testing video we select 5 sentences, so totally we have 3350 sentences and 670 videos. We map a query sentence into the joint space and find nearest neighbor videos (also in joint space) based on Euclidean distance. We record the first correct video position for each query, and then we calculate mean position, or mean rank over all sentences to measure the video retrieval performance. Note that random assignment will return mean rank of 335. From Table. 2 we find our model is better than CCA results. Qualitatively, Fig. 3 shows top retrieved videos with our method and CCA. Both methods return videos with significant human motion in first query example and return videos of "cooking" in third query example. On one hand, our method performs better in terms of specific object, such as a "guitar" in the second query and "wood" in the third query. On the other hand, our method also performs better in capturing action than CCA, e.g., in the first query, the top four retrieved videos are all related to the action of "tumble down" from the sentence, either "fell down" from motorcycle, in a dance or "be drawn" into the water.

Text Retrieval In this task, we evaluate how well a video can find suitable language descriptions. We map each query video to joint space, and find nearest neighbor sentences in same space. Similar to video retrieval, we record rank of first correct sentence that describe query video and use mean rank to measure the overall performance. In this experiments, we also observe a relatively large improvement of our method over CCA baseline. Qualitatively, Fig. 5 shows the retrieved sentences using our joint model outperforms CCA. Both methods perform quite well in recovering common object such as person, cat and dog, while our method is more stable, e.g., all top four retrieved sentences from first and the 5th query are accurate with our method while there're some error using CCA. Beside, for less common video such as the 4the query, our method get one accurate sentence, and also retrieve "banana is being peeled", which corresponds to the action and the color of the paper, while CCA dose not retrieve any meaningful sentences.

Summary From the above three experiments, we find the mean rank of both video retrieval and text retrieval are quite high, but the SVO prediction accuracies have outperformed SVM or CRF models, which means exploiting better video-text space has great potential for natural language generation. Besides, for both our model and CCA, we find when the training error decreases to a certain point the testing error increases. It is understandable since we only have 1300 training videos while the SVO class number is as large as 504, so it is necessary that we collect larger data set or explore the ontology of classes structure.

Conclusions

In this paper, we propose a unified framework to jointly model video and language. Specifically, a compositional language model is proposed to capture high-level human language expression, and a deep video model is proposed to represent the low-level video signal. Our model can capture the compositionality of subject, verb and object leveraging on continuous word representations rather than word cooccurrence. Experiments demonstrate the advantage of our model over SVM baseline, a CRF model and a CCA baseline for video-language space. Our future work includes 1) exploring more complex sentence compositionality beyond SVO, 2) exploring better deep video model to capture motion 3) exploring more complex model that consider scene, spatial-temporal prepositions and other interesting details in video.

Acknowledgements This work was in part supported by NSF CAREER IIS-0845282, DARPA/ARL W911NF-10-2-0062, and ARO W911NF-11-1-0090.

References

Barbu, A.; Bridge, A.; Burchill, Z.; Coroian, D.; Dickinson, S.; Fidler, S.; Michaux, A.; Mussman, S.; Narayanaswamy, S.; Salvi, D.; Schmidt, L.; Shangguan, J.; rey Mark Siskind, J.; Waggoner, J.; Wang, S.; Wei, J.; Yin, Y.; and Zhang, Z. 2012. Video in sentences out. In *UAI*.

Chen, D. L., and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*.

Das, P.; Xu, C.; Doell, R. F.; and Corso, J. J. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*.

Das, P.; Srihari, R. K.; and Corso, J. J. 2013. Translating related words to videos and back through latent topics. In *WSDM*.

Deng, J.; Li, K.; Do, M.; Su, H.; and Fei-Fei, L. 2009. Construction and analysis of a large scale image ontology. In *Vision Science Society*.

Donahua, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2013. Decaf: A deep convolutional activation feature for generic visual recognition. In *arXiv:1310.1531*.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part based models. *TPAMI*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Goller, C., and Kuchler, A. 1996. Learning task-dependent distributed representations by backpropagation through structure. In *International Conference on Neural Networks*.

Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; and Saenko, K. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*.

Gupta, A.; Srinivasan, P.; Shi, J.; and Davis, L. S. 2009. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

Klein, D., and Manning, C. D. 2013. Accurate unlexicalized parsing. In *ACL*.

Krishnamoorthy, N.; Malkarnenkar, G.; Mooney, R. J.; Saenko, K.; and Guadarrama, S. 2013. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR*.

Li, L.-J.; Su, H.; Xing, E. P.; and Fei-Fei, L. 2011. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Miller, G. A. 1995. Wordnet: A lexical database for english. In *Communications of the ACM*, 39–41.

Motwani, T., and Mooney, R. 2012. improving video activity recognition using object recognition and text mining. In *ECAL*.

Moxley, E.; Mei, T.; and Manjunath, B. S. 2010. Video annotation through search and graph reinforcement mining. In *IEEE Transactions on Multimedia*.

Nakayama, H.; Harada, T.; and Kuniyoshi, Y. 2010. Evaluation of dimensionality reduction methods for image auto-annotation. In *BMVC*.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *HLT-NAACL*.

Pereira, F.; Tishby, N.; and Lee, L. 1993. Distributional clustering of english words. In *ACL*.

Ramanathan, V.; Liang, P.; and Fei-Fei, L. 2013. Video event understanding using natural language description. In *ICCV*.

Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B. 2013. Translating video content to natural language descriptions. In *ICCV*.

Sadanand, S., and Corso, J. J. 2012. Action bank: A high-level representation of activity in video. In *CVPR*.

Socher, R., and Fei-Fei, L. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*.

Socher, R.; Karpathy, A.; Le, Q. V.; Manning, C. D.; and Ng, A. Y. 2013. Grounded compositional semantics for finding and describing images with sentences. In *Transactions of the ACL*.

Song, J.; Yang, Y.; and Huang, Z. 2011. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM International Conference on Multimedia*.

Thomason, J.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; and Mooney, R. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*.

Tian, Y.; Sukthankar, R.; and Shah, M. 2013. Spatiotemporal deformable part models for action detection. In *CVPR*.

Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *CVPR*.

Yao, T.; Mei, T.; Ngo, C.-W.; and Li, S. 2013. Annotation for free: Video tagging by mining user search behavior. In *ACM International Conference on Multimedia*.

Yu, H., and Siskind, J. M. 2013. Grounded language learning from video described with sentences. In *ACL*.

Zhang, W.; Zhu, M.; and Derpanis, K. G. 2013. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*.