Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos

Bingbin Liu[1][0000-0002-6621-0356], Serena Yeung[1,2][0000-0003-0529-0628], Edward Chou[1][0000-0002-0670-459X], De-An Huang[1][0000-0002-6945-7768], Li Fei-Fei[1,2], and Juan Carlos Niebles[1,2][0000-0001-8225-9793]

[1] Stanford University, Stanford CA 94305, USA
[2] Google Cloud AI, Mountain View CA 94043, USA

Abstract. A major challenge in computer vision is scaling activity understanding to the long tail of complex activities without requiring collecting large quantities of data for new actions. The task of video retrieval using natural language descriptions seeks to address this through rich, unconstrained supervision about complex activities. However, while this formulation offers hope of leveraging underlying compositional structure in activity descriptions, existing approaches typically do not explicitly model compositional reasoning. In this work, we introduce an approach for explicitly and dynamically reasoning about compositional natural language descriptions of activity in videos. We take a modular neural network approach that, given a natural language query, extracts the semantic structure to assemble a compositional neural network layout and corresponding network modules. We show that this approach is able to achieve state-of-the-art results on the DiDeMo video retrieval dataset.

Keywords: Video retrieval, action recognition, modular networks

1 Introduction

A fundamental goal of computer vision is understanding rich, diverse and complex activities occurring over time in a dynamic visual world. While there has been significant progress in activity recognition, it is often restricted to a constrained setting with a fixed number of action classes for each particular dataset [1, 6, 22, 25, 26, 28, 33, 51, 66]. Scaling these recognition models to the long tail of complex activities is still an open problem in this paradigm, as it requires collecting large quantities of data for new action classes and does not explicitly exploit similarity between activities.

To address this problem, a natural solution is to describe complex activity in natural language [5, 7, 39, 44, 59]. This allows for supervised labels containing rich, unconstrained information about the activity, and motivates tasks such as video retrieval [16, 47, 52, 55]. This formulation also gives hope of leveraging the underlying structure in the activity description in order to reuse learned sub-concepts across activities. The approach we use endows models with an increasingly compositional structure. For example, a complex concept like "girl

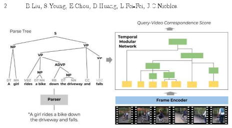

Fig. 1. Given a natural language query and video as input, Temporal Modular Networks (TMN) uses the underlying language structure of the query to dynamically assemble a corresponding modular neural network that reasons compositionally over the video to produce a query-video correspondence score.

riding a bike down a driveway then falling" can be decomposed into two subevents "riding" and "falling" which can be observed and learned in very different contexts (riding a bike vs. a skateboard, down a driveway vs. down a hill).

In this work, we focus on the *natural language video retrieval task*. Given an input in the form of natural language description, the goal is to retrieve the best matching video. The variety of the language descriptions and visual appearances makes it a challenging task beyond the classification of predefined action categories. Existing video retrieval methods typically learn embedding representations of language and video using recurrent neural networks [9, 14, 16, 61, 63] or spatio-temporal convolutions [21, 25, 54]. While simple and effective, these approaches fail to capture, and more importantly, leverage, the inherently compositional structure of the concepts and fail to properly relate each subconcept for efficient reasoning. We posit that explicitly modeling compositional structure is key for the generalizability and scalability needed for complex video understanding.

To this end, we introduce a dynamic compositional approach for reasoning about complex natural language descriptions of activity in videos. We draw inspiration from recent successes in visual question answering using compositional models [2, 3, 17, 18, 24, 58]. Given a natural language query and a video, our approach explicitly makes use of the underlying language structure of the query to dynamically (and hierarchically) assemble a corresponding modular network to reason over the video, and output the correspondence between the query and the video (Fig. 1). More specifically, we use a natural language parser to extract a structure from the description. Using this structure, we construct a hierarchical layout based on which corresponding neural network modules are assembled.

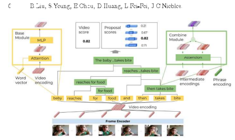

Fig. 2. Temporal modular network architecture. The compositional layout of the network is determined by structure extracted from a natural language parser (Sec. 3.1). *Base modules* (yellow) in the network reason directly over the temporal sequence of frame-level visual encodings for a video, based on node-specific word embeddings. *Combine modules* (green) combine information from child nodes, based on node-specific higher-level phrase embeddings. The output of the top-level combine module is used to produce, depending on the setting, a score corresponding to the strength of the video match with the query, or scores of temporal proposals within the video. See Sec. 3.2 for more details.

While a parser provides an initial compositional structure, some POS tags neither represent nor relate visual concepts, such as *DT* (determiner) and *RP* (particle). We therefore discard these elements from the parse tree. We furthermore merge tags that differ in tenses or pluralities but belong to the same word class. For example, *VBZ* (verb, third-person singular present) and *VBD* (verb, past tense) are merged as *VB* (verb, base form). Table 1 specifies the POS tag mapping. After merging and discarding, the total number of POS tags appearing in a tree is reduced from 36 to 8.

Then, nodes in the resulting tree can be categorized into two types: *base nodes* that correspond to single words in a description, and *combine nodes* which correspond to phrases (sequences of words) and combine its child nodes.

3.2 Dynamically assembling compositional networks over video

We have described in Sec. 3.1 how we can use natural language parsing to obtain inherent compositional structures from arbitrary descriptions of complex activities. The challenge at hand then becomes how we can use this structure to perform compositional reasoning in videos. Our key insight is that we can leverage this language structure to modularize the corresponding video understanding network for modeling the structure of the activity.

Table 2: TMN outperforms MCN using RGB, flow, and fused (RGB+flow) features. The significant gain on RGB features in particular shows TMN's ability of temporal reasoning without relying on motion information in features.

Feature	Model	Rank@1	Rank@5	mean IoU
RGB	MCN	13.10	44.82	25.13
	TMN	**18.71**	**72.97**	**30.14**
Flow	MCN	18.35	56.25	31.46
	TMN	**19.90**	**75.14**	**31.95**
Fuse	MCN	19.88	62.39	33.51
	TMN	**22.92**	**76.08**	**35.17**

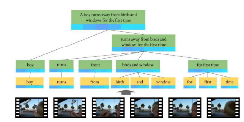

Fig. 3. Qualitative example of TMN evaluated on a query-video pair. Attention maps at base and combine nodes are visualized, where the color bars represent attention weights with darker blue indicating higher weights. Attention maps for the base nodes show the activation of a single word, while the attention map in a combine node shows how information from children are aggregated, and how the modules specifically take in the temporal quality ("for the first time") of the phrase encoding in each module

Ablation Study We perform ablation studies to investigate the variations in module design, network structures, and loss functions:

- *Type of base modules*: We experimented with two types of base module: the *POS* setting with one base module per POS tag, and the *Single* setting where a single based module is shared across all tags. The *POS* setting may ease the learning for TMN by making each module more specialized, whereas the *single* setting allows TMN to learn from larger amounts of data and may help capture patterns existing across POS tags. For example, a single shared module may be similarly parameterized by words with different POS tags but appearing in a similar context. Moreover, using a single module was shown to be more robust since it provides better tolerance towards the

Fig. 4. Example outputs of TMN, where TMN is able to recognize temporal changes such as "*switch on*" and "*turn/run away*", as well as compositional relations such as "*then*".

  parser error, which sometimes mistakenly assigns a noun tag to a singular form verb. The *single* setting was chosen based on our experimental results.
- *Attention in combine module*: In addition to having the combine module selectively attend to different temporal segments based on the phrase encoding, we also considered max pooling as a simplified alternative to combining multiple child feature maps, where the output feature map is element-wise max pooled from all children. This is inspired by the parent-child constraint in [58], where a structure loss is used to penalize the combined feature map from deviating from the union of child feature maps, which is essentially approximating a max pool layer. Formally, the combined feature map is defined such that $\forall i \in \{1 \dots n\},\ j \in \{1 \dots D_v\}$,

$$M^{out}_{i,j} = \max_{c \in C} M^c_{i,j} \quad (5)$$

  where C is the set of children and M^c is the feature map of the c_{th} child.

- *Effect of a proper compositional network structure*: We compared three network structures. The first one was without the compositional tree. Since TMN resembles a vanilla RNN when the compositional structure is taken away, the performance gap between MCN[16] and TMN corresponds to the gain of the compositional structure. The other two structures came from a dependency parser and a constituency parser. We found out that structures from both parsers were able to outperform MCN, demonstrating the importance of compositional reasoning. Further, the performance gap between the