

Referring to Objects in Videos using Spatio-Temporal Identifying Descriptions

Peratham Wiriathamabhum[♠]◇, Abhinav Shrivastava[♠]◇,
Vlad I. Morariu[◇], Larry S. Davis[♠]◇

University of Maryland: Department of Computer Science[♠], UMIACS[◇]
peratham@cs.umd.edu, abhinav@cs.umd.edu
morariu@umd.edu, lsd@umiacs.umd.edu

Abstract

This paper presents a new task, the grounding of spatio-temporal identifying descriptions in videos. Previous work suggests potential bias in existing datasets and emphasizes the need for a new data creation schema to better model linguistic structure. We introduce a new data collection scheme based on grammatical constraints for surface realization to enable us to investigate the problem of grounding spatio-temporal identifying descriptions in videos. We then propose a two-stream modular attention network that learns and grounds spatio-temporal identifying descriptions based on appearance and motion. We show that motion modules help to ground motion-related words and also help to learn in appearance modules because modular neural networks resolve task interference between modules. Finally, we propose a future challenge and a need for a robust system arising from replacing ground truth visual annotations with automatic video object detector and temporal event localization.

1 Introduction

Localizing referring expressions in videos involves *both static and dynamic information*. A referring expression (Dale and Reiter, 1995; Roy and Reiter, 2005) is a linguistic expression that grounds its meaning to a specific referent object in the world. The input video can be very long, have unknown length, contain many objects from the same class, or contain similar actions and interactions throughout the video. A successful, grounded communication between a speaker and a listener must ensure that the sentence or discourse provides enough information such that the listener can eliminate all distractors and focus only on the referent object that acts in a specific time interval. That essential information varies from the diversity of events in the world. However, a speaker is

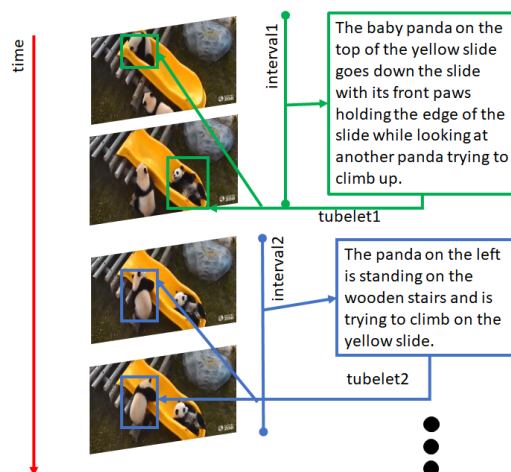


Figure 1: The first spatio-temporal identifying description in the green box grounds to the event that a panda goes down the slide. Another panda can be a context because they are interacting in the same scene. The second identifying description in the blue box grounds to the event that another panda climbs up the slide.

likely to mention salient properties and also salient differences based on the referent in comparison to other distractors. The differences can be about object category, attributes, poses, actions, changes in location, relationships and contexts in the scene.

Existing *image referring expression* datasets (Mao et al., 2016; Johnson et al., 2015; Kazemzadeh et al., 2014; Plummer et al., 2017; Krishna et al., 2017b) do not contain referring expressions that refer to dynamic properties or movements of the referent. These datasets do not require temporal understanding that would require a system to learn that “moving to the right” is different from “moving to the left” and “getting up” is different from “lying down”. Existing *video referring expression* datasets and approaches (Krishna et al., 2017a; Hendricks et al., 2017; Gao et al., 2017; Berzak et al., 2015; Li et al., 2017; Hendricks et al., 2018; Gavriyuk et al., 2018) focus only on temporal localization but referent

Two-stream Modular Attention Neural Network

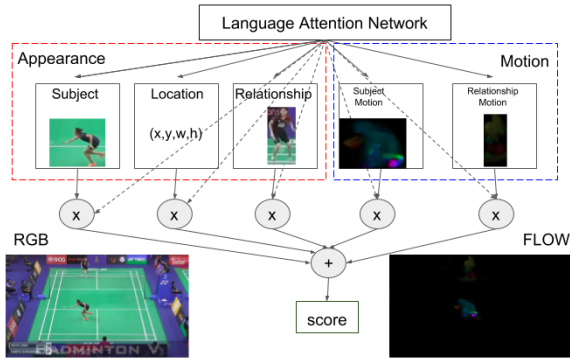


Figure 2: We add a motion stream to modular attention network. Our motion modules take optical flow input and model motion information for the subject and its relationship.

object localization. In other words, they do not ground events in both space and time. Emphasized by (Cirik et al., 2018), the data collection process for referring expressions should incorporate linguistic structure such that the model can learn more than shallow correlations between pairs of a sentence and visual features. That is, a particular dataset should not have a shortcut that only detecting nouns (object class) can perform well. Our dataset mitigates this issue by forcing instance-level recognition. We create a requirement that grounding the referring expressions must identify the target object among many distractors from the *contrast set* (same class distractors).

The contributions of this paper are (i) We propose a novel vision and language data collection scheme based on grammatical constraints for surface realization to ground video referring expressions in both space and time with lexical correlations between vision and language. We collected the *Spatio-Temporal Video Identifying Description Localization (STV-IDL)* dataset consisting of 199 video sequences from Youtube and 7,569 identifying descriptions. (ii) We propose an interpretable system based on *two-stream modular attention network* that models both appearance and motion to ground referring expressions as instance-level video object detection and event localization. We also perform ablation studies to get insights and identify potential challenges for the task.

2 Spatio-Temporal Localization

Given ground truth temporal intervals ($[start, end]$) and object tubelets (a sequence

of object bounding box coordinates in a given temporal interval, $\{[x_0, y_0, x_1, y_1]_{start}, \dots, [x_0, y_0, x_1, y_1]_{end}\}$), we want to localize an identifying expression ie to the correct target tubelet tb_{target} not the distractor tubelets $tb_{distractor}$ as our predicted tubelet r . We evaluate using the accuracy measure.

For automatic localization, tubelet IoU (Rusakovskiy et al., 2015) and temporal IoU are used to evaluate the bounding box and temporal interval with the ground truth respectively. Let R_i be the region in the frame i to be detected,

$$tubelet\ IoU = \frac{\sum_i \delta(IoU(r_i, R_i) > 0.5)}{N}, \quad (1)$$

where the denominator is the number of detected frame measured by the standard Intersection over Union (IoU) in an image and N denotes the number of union frames.

$$temporal\ IoU = \frac{\cap(interval_i, interval_j)}{\cup(interval_i, interval_j)}, \quad (2)$$

where $interval_i$ and $interval_j$ are input temporal intervals and the intersection and union functions are operations over 1-D intervals.

3 Related Work

Spatio-Temporal Localization. Spatio-temporal localization (or action understanding) is a long standing challenge in computer vision. Most existing datasets like LIRIS-HARL (Wolf et al., 2014), J-HMDB (Jhuang et al., 2013), UCF-Sports (Rodriguez et al., 2008), UCF-101 (Soomro et al., 2012) or AVA (Gu et al., 2018) localize a spatio-temporal tubelet for human actions in either trimmed videos or a simple visual setting or a fixed lexicon. In contrast to action labels, our work accepts a free-form referring expression annotation which also contains a richer set of relations in the forms of prepositions, adverbs and conjunctions.

Referring Expression Comprehension. The goal of referring expression comprehension (Golland et al., 2010) is to ground phrases or sentences into the specific visual regions that the phrase refers. Prior works in the image domain have either focused on using a captioning module to generate the sentence (Mao et al., 2016; Nagaraja et al., 2016) or learning a joint embedding to comprehend the sentence by modeling the corresponding region unambiguously and localize the region

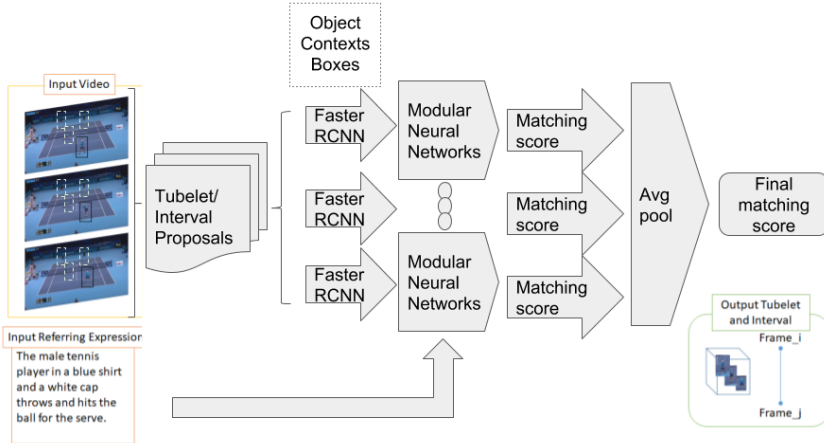


Figure 3: An overview of our system: an input video is using either ground truth annotations or is fed into both tubelet object proposal and temporal interval proposal modules. The resulting tubelet and interval proposals are then fed into an appearance and motion Faster-RCNN to extract the two-stream features. Then, a modular neural network will rank the tubelets given an input referring expression. The scores are average-pooled, and the system outputs the most likely tubelet that contains the reference object.

(Rohrbach et al., 2016; Wang et al., 2016; Hu et al., 2017; Yu et al., 2018). For the video domain, (Yamaguchi et al., 2017) further annotated the ActivityNet dataset with one referring expression per video for video retrieval with natural language query. (Li et al., 2017) uses referring expressions to help track a target object in a video sequence in a subset of OTB100 (Lu et al., 2014) and ImageNet VID (Russakovsky et al., 2015). DiDeMo (Hendricks et al., 2017) and TEMPO (Hendricks et al., 2018) focus on localizing an input sentence into the corresponding temporal interval out of a finite number of backgrounds. Importantly, these datasets do not consider distractor objects from the same class. While our work also focuses on the video domain, it focuses on localizing objects and events as spatio-temporal tubelets aligned with an input expression.

Surface Realization in Vision and Language.

Surface realization is a process for generating surface forms, like natural language sentences, based on some underlying representations. For natural language generation, the underlying representation tends to be syntactic features. In vision and language, captioning systems can use meaning representation like triplets as an input for a surface realization module to generate a sentence. (Farhadi et al., 2010) uses <Objects, Actions, Scenes>. (Yang et al., 2011) uses part-of-speech as <Nouns, Verbs, Scenes, Prepositions>. (Li et al., 2011) uses <<adj1, obj1>, prep, <adj2, obj2>> where adjectives are object attributes and prepositions are spatial relationships between ob-

jects. TEMPO (Hendricks et al., 2018) and TVQA (Lei et al., 2018) use a compositional format for words like before or after to specify temporal relationships between events during crowdsourcing.

We incorporate grammatical constraints (Linguistic prescription) based on part-of-speech into our annotation pipeline so that we can crowdsource well-formed sentences from people which contain enough meaning representations for vision systems to locate the target object with visual contexts. Instead of manually writing sentences based on context-free grammars like (Yu and Siskind, 2013), we ask the annotators to write sentences in which valid sentences contain at least a noun phrase (NP), a verb phrase (VP) and one of a prepositional phrase (PP), adverb phrase (ADVP) or conjunction phrase (CONJP). The rest of each sentence are language variations where we expect crowdsourcing to create more variations compared to manual annotations by a few annotators. We want computer vision models to learn useful and interpretable features by correlating the expressions and videos. So, we want the learned visual semantics from grounding models to be similar to structural inputs in surface realization systems. Each part-of-speech correlates with a specific visual feature.

4 STV-IDL Dataset

4.1 Dataset Construction

We develop a new data collection schema that ensures rich correspondences between referring ex-

Table 1: STV-IDL dataset statistics.

Info	Statistics
Number of Videos	199
Number of Sentences	7569
Average objects per Video	2.85
Average words per Sentence	22.65
Sentences per Video	38.04

pressions and referred objects in a video using constraints. The spatio-temporal relations that we are interested in are about state transitions, that is, what happens before and after the action and how objects move. The state transitions should be relative to other objects and background. For example, a sentence ‘*A man in a green uniform kicking the ball then running toward the net.*’ is a good video referring expression. This sentence is valid only in a spatial region that represents a noun phrase ‘*a man in a green uniform*’ and a time interval in which an action from the verb phrase ‘*hitting the ball then running toward the net*’ occurs. Also, the action ‘*hitting the ball*’ comes before his next action ‘*running toward the net*’ which shows the action steps of ‘*hitting*’ followed by ‘*running*’ and the action ‘*running*’ has a context object ‘*the net.*’

First, we ensure that all of our High Definition videos (720p) crawled from Youtube contain at least two objects similar to (Mao et al., 2016), but each video will focus on just one object class to form a contrast set. This constraint prevents a simple video object detector from resolving referential ambiguity using only nouns by just outputting based on class information as in (Cirik et al., 2018). Because a simple object detector randomly outputs one object from a combination of the target and the contrast set. Then, the output is the same as random because the object confidence scores do not correlate with the referring expression. We want more language cues to guide the system to seek additional visual contexts (Divvala et al., 2009) to focus and output *only one* unambiguous object detection. The dataset contains 13 categories of videos which are either multi-player sports or animals. Second, inspired by (Siskind, 1990; Yu and Siskind, 2013), a sentence, consists of a subject and a predicate, can be viewed as a set of structured labels based on part-of-speech and each label can be meaningfully grounded in a video. Besides, annotations can use *grammars for lexical grounding and surface real-*

Table 2: STV-IDL part-of-speech statistics. (Please see the supplementary material for more details.)

Part-of-Speech	percents
Noun, singular or mass (NN)	28.1
Determiner (DT)	15.3
Preposition or subordinating conjunction (IN)	10.9
Adjective (JJ)	9.7
Possessive pronoun (\$PRP)	5.7
Verb, 3rd person singular present (VBZ)	5.1
Adverbs (RB)	3.6
Coordinating conjunction (CC)	3.4

ization. Therefore, we ensure that every referring expression in our dataset provides grammatically relevant visual grounding based on part-of-speech such that a valid sentence must contain at least a noun phrase (NP), a verb phrase (VP) and one of a prepositional phrase (PP), adverb phrase (ADVP) or conjunction phrase (CONJP). We also found that the annotators may write relevant sentences without the constraints but the contents are random and may not be visually grounded either spatially or temporally or both in the video. Some example sentences without the constraints are “*The guy was lucky to save the tennis ball.*” and “*The sun is blocking the ball for the back player.*”.

4.2 Video Tubelet, Temporal Interval, and Expressions Annotations

We manually identify interesting events in each video and select a keyframe for that action in the presence of distractors. Then, we manually annotate the start and end of that event into an interval lasting around one second. For bounding box annotation, we use a javascript variant of Vatic (Vondrick et al., 2013; Bolkensteyn) to manually draw a tubelet of bounding boxes in all frames for each object of interest in every video. We crowdsource annotations of referring expressions from Amazon Mechanical Turk (AMT). We create a clip segment with a bounding box around the target object to fixate the annotator’s attention.

Next, we manually verify the referring expressions using another web interface that helps us evaluate if the sentence refers to the target object, is correct based on the video, is different from sentences for the distractors and is sufficient to distinguish the target object from the distractors and the

Two-stream Modular Attention Neural Network

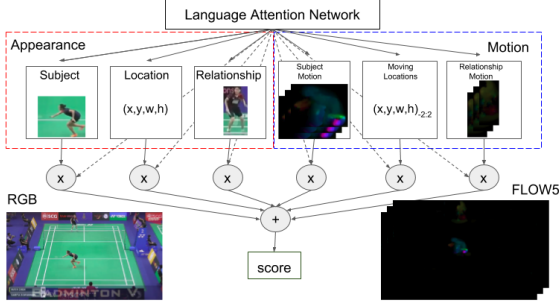


Figure 4: Stacked two-stream modular attention network based on five optical flow image input. We model the bounding box sequence is a moving location module and a relationship module. The motion Faster-RCNN is also trained using a stack of five flow images for frame index $f_i \in [t - 2, t + 2]$.

background. The annotation interfaces, payment and dataset statistics are shown in supplementary material. We refer to the resulting referring expressions as *identifying descriptions* (Mitchell et al., 2013) because our expressions are referring expressions in the verified intervals which may be overspecified but are also descriptions which may be underspecified for the whole videos. Our referring expressions are long because we want to make sure that they are clear enough to provide input cues for the system. However, it still might be not enough to localize an event from the whole video because the video has many events and can be exhaustive to be specific for a particular event.

5 Approach: Two-stream Modular Attention Network

We start by employing a state-of-the-art image referring expression localization, namely, Modular Attention Network (MAttNet) (Yu et al., 2018) for our tasks. This model fits our objective since it is a variant of modular neural networks (Auda and Kamel, 1999; Andreas et al., 2016) that is decomposed based on tasks according to Fodor’s modularity of mind (Fodor, 1985). Therefore, we can interpret the model in an ablation study on each neural module for a specific vision subtask and input type. Also, the model also provides linguistic interpretability using its language attention module that can visualize different bindings from word symbols in a referring expression to each visual module as attention $a_{m,t}$ where module $m \in \{subj, loc, rel\}$ (subject, location, rela-

tionship) and t is the index location of the word this attention weights its hidden representation the Bi-LSTM encodes.

The original MAttNet model (RGB) decomposes image referring expression grounding into three modules, a subject module, a location module, and a relationship module. The network output score for an object o_i and an expression r is,

$$S(o_i|r) = \sum_{m \in \text{modules}} w_m S(o_i|q^m), \quad (3)$$

where w_m is the weight vector from the language attention module on the visual module m . q^m is the weighted sum of attention $a_{m,t}$ over all word embedding. $S(o_i|q^m)$ is the module score from a cosine similarity in the joint embedding between the visual representation of o_i denoted as \tilde{v}_i^m and q^m .

Given a positive pair (o_i, r_i) , the network is discriminatively trained by sampling two negative pairs (o_i, r_j) and (o_k, r_i) where r_j is the expression from other contrast object and o_k is the contrast object from the same frame. The combined hinged loss L_r is,

$$L_r = \sum_i \lambda_1 \max(0, \Delta + S(o_i, r_j) - S(o_i, r_i)) + \lambda_2 \max(0, \Delta + S(o_k, r_i) - S(o_i, r_i)). \quad (4)$$

The loss is linearly combined with other loss terms such as attribute prediction with cross-entropy loss L_{att} from the subject module in a multi-task learning setting.

We extend MAttNet to the video domain by applying two things. First, MAttNet uses Faster-RCNN (Girshick, 2015) for feature extraction so we follow a well-established actor-action detection pipeline which extends image object detection to frame-based spatio-temporal action detection (Peng and Schmid, 2016). With this, we reframe the problem by replacing action labels with referring expressions and putting MAttNet on top of Faster-RCNN. Also, we use external object and interval proposal instead of Region Proposal Network (RPN) in Faster-RCNN. Second, we add subject motion and relationship motion modules to capture temporal information in a two-streams setting (Simonyan and Zisserman, 2014). These modules have the same architecture as the subject and relationship module but are using optical flow as their input. We replace the three channel RGB

input with a stack of flow-x, flow-y and flow magnitude from the flow image. The aim of these modifications, depicted in Figure 2, is to better model attributes, motion, movements and dynamic context in a video.

Previous work (Simonyan and Zisserman, 2014) has shown that stacking many optical flow images can help recognition. So, we train another variant of two-stream modular attention network using stacked five optical flow frames shown in Figure 4. In this setting, we train the stacked motion Faster-RCNN by stacking flow images F_{idx} where frame index $idx \in [t - 2, t + 2]$. The input becomes a 15 channel stacked optical flow image. In addition, we add the moving location module to further model the movement of the location by stacking location features $l_i = [\frac{x_{min}}{W}, \frac{y_{min}}{H}, \frac{x_{max}}{W}, \frac{y_{max}}{H}, \frac{Area_{region}}{Area_{image}}]$ where W and H are width and height of the image. Then the location features are concatenated with the location difference feature of the target object with up to five context objects from the same class, $\delta_{ij} = [\frac{\Delta x_{min}}{W}, \frac{\Delta y_{min}}{H}, \frac{\Delta x_{max}}{W}, \frac{\Delta y_{max}}{H}, \frac{\Delta Area_{region}}{Area_{image}}]$ so that we have a sequence of $[l_i; \delta_{ij}]_{idx}$ where frame index $idx \in [t - 2, t + 2]$. Then, we place an LSTM on top of the sequence and we forward the concatenation of all hidden states to a fully connected layer and output the final location features. We also make a location sequence and place an LSTM on top of location in the relationship motion module in this stacked optical flow setting.

5.1 Tubelet and Temporal Interval Proposals

We employ the state-of-the-art video object detector, flow-guided feature aggregation (FGFA) (Zhu et al., 2017), finetuned on STV-IDL to generate the tubelet proposals. The per-frame detections from FGFA are post-processed by linking into tubelets using Seq-NMS (Han et al., 2016) based on the top 300 bounding boxes ranked by the confidence of the category scores.

For temporal proposals, we implemented a variant of Deep Action Proposals (DAPs) (Escorcia et al., 2016) based on multi-scale proposal. First, we use a temporal sliding window with a fixed length of L frames and a stride of s (8 in our case). This produces a set of intervals, (b_i, e_i) where b_i and e_i are the beginning and the end of the interval. Then, we extract the C3D features (Tran et al., 2015) from the image frames in that interval using the activation in the ‘fc7’ layer, pretrained

Table 3: Identifying Description Localization: mAP for each collection. (values are in percents.) The fused1 MAttNet is the proposed two-stream method and the fused5 MAttNet is the stacked version of the proposed two-stream method.

Model	mAP
random	29.68
RGB MAttNet	41.51
flow MAttNet	39.02
flow5 MAttNet	41.90
fused1 MAttNet	44.66
fused5 MAttNet	42.82

Table 4: Ablation study on fused1 MAttNet: mAP for each module combination. (values are in percents.)

Model	mAP
Subject+Location	44.46
+Relationship	44.46
+Subject Motion	44.46
+Relationship Motion	44.66

on the Sports-1M dataset (Karpathy et al., 2014). The feature set $f = C3D(t_i : t_i + \delta)$, $t_i \in [b_i, e_i]$ where $\delta = 16$ from the original pretrained model. The duration of each segment L_k also increases as a power of 2, that is $L_{k+1} = 2 * L_k$. The features are fed to a 2-layered LSTM to perform $\{Event/Background\}$ sequence classification.

6 Experiments and Analysis

We want to show how and to what extent modular attention networks ground input expressions with motion information in videos. So, we perform two sets of experiments, identifying description localization and automatic video object detector and temporal event localization. Similar to (Gu et al., 2018), we split the dataset into training, validation and test sets at the video level; that is, there are no overlapping video segments for every split. There are 159 training, 13 validation, and 27 test videos. The rough ratio is 12:1:2. Implementation details are in the supplementary material.

6.1 Identifying Description Localization

Setup. We perform three experiments, localization with ground truth annotations, module ablation study, and word attention study. First, we evaluate our model by selecting the target from a pool of candidate targets plus distractors. We compare five models based on input and modules. The five models are (1) MAttNet for RGB in-

Table 5: Ablation study on fused5 MAttNet: mAP for each module combination. (values are in percents.)

Model	mAP
Subject+Location	33.97
+Relationship	35.32
+Subject Motion	35.41
+Moving Location	42.84
+Relationship Motion	42.82

put (RGB MAttNet/original model/baseline); (2) MAttNet for flow image input (flow MAttNet); (3) MAttNet for stacked five flow image input (flow5 MAttNet); (4) two-stream MAttNet for RGB and flow image input (fused1 MAttNet) and (5) two-stream MAttNet for RGB and stacked five flow image input (fused5 MAttNet). Second, we interpret the model by setting the module score weights from language attention module to zeros for the modules we want to turn off in our ablation study. Third, we collect the statistics of the attention of each word from the input expressions in the test set to explain how and which kind of words each module attends.

Results. The accuracies in Table 3 show superior performance for stacked flow5 and two-streams models. The stacked flow5 model improves over the RGB baseline by 0.39% while two-stream fused1 and fused5 models have 3.15% and 1.31% improvement respectively. Both variants of two-stream models, fused1 and fused5, outperform all one-stream models, RGB, flow, and flow5. All models perform better than randomly selecting an object from the set of tubelets.

The accuracies in Table 4 show that each module in fused1 learns better since the modules in appearance stream alone have 2.95% improvements over the RGB only baseline. We further hypothesize that *the reason is the motion stream takes care of motion grounding so the appearance modules can learn better because of the separation of unrelated information into other modules. A modular neural network avoids internal interference between features by training each module independently and each module will masters its task more precisely (Auda and Kamel, 1999)*. The additional relationship motion module also provides complementary information for the additional 0.20% improvement. The accuracies in Table 5 show that the stacked flow5 model focuses mostly on the moving location module which causes the overall improvement over the RGB baseline. *The mov-*

ing location is a predictive feature to model motion and spatial location (Yin and Ordonez, 2017), but it prevents other vision modules from becoming sufficiently tuned in this setting. We also try to combine the moving location with the fused1 setting. The results degenerate more, and the overall accuracy is only 37.12%. It is even lower than flow MAttNet model.

Figure 6 shows how the language attention network assigns weights to each module by aggregating all the weights for each word based on Penn part-of-speech tag during test set prediction of the fused1 model to explain the performance gain. The aggregated statistics show that motion words like verbs, prepositions, and conjunctions are ranked higher for flow modules on average which means more attention to motion. We also focus on just aggregating verbs in Figure 7 to further explain the modules. The statistics show that flow and location modules focus more on verbs on average compared to their corresponding appearance-based modules.

6.2 Automatic Video Object Detector and Temporal Event Localization

Because spatio-temporal detection and localization is very challenging, we want to identify potential challenges for spatio-temporal grounding when automatic computer vision systems replace the ground truth annotations. So, we replace tubelets with top 8 detections from flow-guided feature aggregation (FGFA) (Zhu et al., 2017) and temporal intervals with the proposal system described in Section 3.2. We create three scenarios: in each scenario, varying amounts of the problem are revealed via the ground truth to separate each component and measure the hardness of each sub-problem and the impact of one on another.

6.2.1 Automatic Video Object Detector

Setup. We evaluate both the tubelet object proposals and the pretrained modular attention networks. We replace the groundtruth tubelets to imperfect proposals which contains bounding box perturbations and we want to see how the model behaves.

Results. Since all modular attention networks are not trained on tubelet proposals, the results from the automatic video object detector in Table 6 shows performance drops in all models and the performances are even lower than the object detection baseline. The object detection baseline selects the tubelet with the highest confidence score



Figure 5: A qualitative result: the first, middle and last frames from an interval in the STV-IDL dataset with an expression, ‘The male tennis player in the near court moves from right to left in order to hit the ball but his teammate outside the court reaches the ball first and just hits it.’ The fused1 MAttNet can properly refer to the object highlighted in the red box in contrast to the baseline.

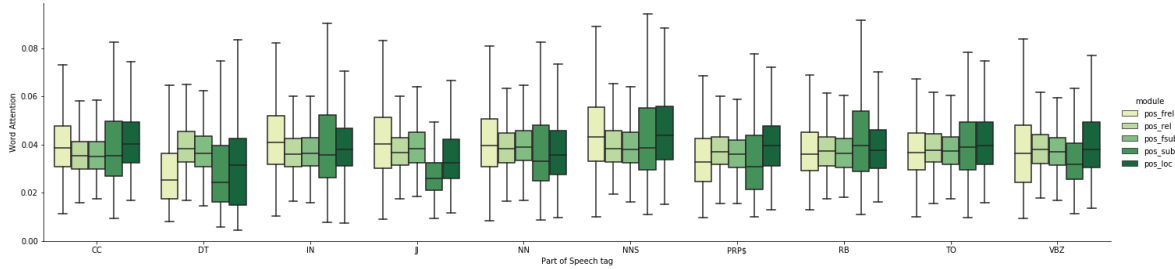


Figure 6: Aggregations of output word attention weights for each module on the STV-IDL test set. Part-of-speech tags are CC, DT, IN, JJ, NN, NNS, PRP\$, RB, TO and VBZ (left to right).

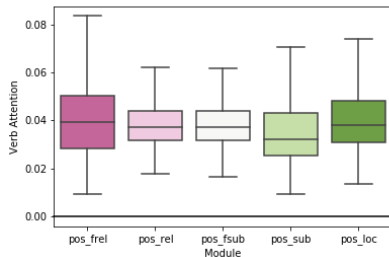


Figure 7: Aggregations on all verbs for each module. (from left to right: Relationship flow/RGB, Subject flow/RGB, Location)

Table 6: Visual Object Detection: mAP tracklet IoU@0.5 for each model. (values are in percents.)

Model	IoU@0.5
RGB MAttNet	35.02
flow MAttNet	22.63
flow5 MAttNet	28.98
fused1 MAttNet	23.93
fused5 MAttNet	24.26
<i>FGFA most conf.</i>	35.87
<i>FGFA 2nd conf.</i>	34.16

from FGFA. We hypothesize that it is from bounding box perturbation that may affect both Faster RCNN features and location features. The results also show that the performance drops are more severe in two-stream models - we think that it is from an accumulation of errors from both streams.

Table 7: Event Localization: mAP temporal IoU@0.5 for each model. (values are in percents.)

Model	tIoU@0.5
RGB MAttNet	8.72
flow MAttNet	7.28
flow5 MAttNet	8.79
fused1 MAttNet	8.07
fused5 MAttNet	7.02
speaker LSTM	7.74
speaker Bi-LSTM	10.10

6.2.2 Temporal Event Localization

Setup. We evaluate the event localization component by removing ground truth temporal intervals. All previous settings so far operate on trimmed video segments and focus on ‘where’ the sentences refer to. We want to see how the model behaves on untrimmed videos in which the system needs to answer ‘when’ the referred events occur. The system’s task is to infer the temporal intervals $[t_k, t_{k+40})$ which are likely to correspond to the input expressions. We evaluate the system via temporal mean Average Precision with temporal IoU similar to (Krishna et al., 2017b). Since our identifying descriptions are sentences for the whole videos, we compare modular attention networks to a video captioner, S2VT (Venugopalan et al., 2015), which is a speaker model (Mao et al., 2016) that output the probability of producing an expression given a video. The S2VT model is trained

Table 8: Spatio-temporal Localization: mAP temporal IoU@0.5 then tracklet IoU@0.5 for each model. (values are in percents.)

Model	tIoU@0.5
RGB MAttNet	2.75
flow MAttNet	2.04
flow5 MAttNet	2.62
fused1 MAttNet	1.70
fused5 MAttNet	1.51

on a different feature set consisting of the image features from the last layer ‘fc1000’ of ResNet-50 (He et al., 2016), the interval (b_i, e_i) and the current frame number. This S2VT model is trained on ground truth intervals and expressions, so it is likely to produce expressions with high probabilities on the ground truth event intervals compared to the background intervals which do not contain ‘interesting’ events.

Results. The results in Table 7 shows that speaker Bi-LSTM performs the best and even better than all modular attention networks. We suspect that the reason is from the discriminative training scheme of the modular attention networks is not suitable for temporal localization. Training with only negative pairs from the same frame takes a week, so it is computationally expensive to train with all negative pairs from all frames in the whole video. The top-5 prediction for Bi-LSTM increases to 26.23% but it is still far from the upper bound of 71.02%, the recall of the proposal system.

6.2.3 Spatio-temporal Localization

Setup. We evaluate our event interval proposals, tubelet object proposals, and modular attention networks. We fix tubelet Intersection over Union (tubelet IoU) to 0.5. The evaluation is a two-step process, temporal IoU then tubelet IoU. We allow tubelet IoU over all frames of the proposal interval instead of ground truth interval to show that the system refers to the right object in an event interval and the tubelet IoU does not depend on temporal IoU.

Results. The results in Table 8 show that the performance further decreases from Table 7. We suspect that the reason is also from the discriminative training scheme because the models are not trained on some background frames.

7 Summary

We discussed the problem of grounding spatio-temporal identifying descriptions to spatio-temporal object-event tubelets in videos. The critical challenge in this dataset is to ground verbs and motion words in both space and time, and we show that this is possible by our proposed two-stream modular neural network models which have complimentary optical flow inputs to ground verbs and motion words. We validate this by collecting aggregated statistics on word attention and found that the two-stream models ground verbs better. The motion stream also helps the appearance stream learn better because it abstracts away motion noise from appearance. We further inspected the components in the system and revealed potential challenges. A better training scheme such as improved loss functions or hard example mining for future spatio-temporal grounding systems should consider both efficiency and effectiveness.

8 Acknowledgement

The authors thank the anonymous reviewers for their insightful comments and suggestions. We thank Dr. Hal Daumé III, Nelson Padua-Perez and Dr. Ryan Farrell for very useful advises and discussions. We also thank members of UMD Computer Vision Lab (CVL), UMD Human-Computer Interaction Lab (HCIL) and UMD Computational Linguistics and Information Processing Lab (CLIP) for useful insights and support. We also thank academic twitter users for useful knowledge and discussions.

References

- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48.
- Gasser Auda and Mohamed Kamel. 1999. Modular neural networks: a survey. *International Journal of Neural Systems*, 9(02):129–151.
- Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2015. Do you see what i mean? visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1487.

- Dinesh Bolkensteyn. vatic.js: A pure javascript video annotation tool. <https://dbolkensteyn.github.io/vatic.js/>.
- Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 781–787.
- Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.
- Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. 2009. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1271–1278. IEEE.
- Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer.
- Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.
- Jerry A Fodor. 1985. Precis of the modularity of mind. *Behavioral and brain sciences*, 8(1):1–5.
- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275.
- Kirill Gavriluk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966.
- Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.
- Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics.
- Chunhui Gu, Chen Sun, David Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR 2018*.
- Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. 2016. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1380–1390.
- Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4418–4427. IEEE.
- Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. 2013. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199.
- Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678.
- Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma,

- et al. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.
- Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics.
- Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, Arnold WM Smeulders, et al. 2017. Tracking by natural language specification. In *CVPR*, volume 1, page 5.
- Yang Lu, Tianfu Wu, and Song Chun Zhu. 2014. Online object tracking, learning and parsing with and-or graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3462–3469.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.
- Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).
- Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer.
- Xiaojiang Peng and Cordelia Schmid. 2016. Multi-region two-stream r-cnn for action detection. In *European Conference on Computer Vision*, pages 744–759. Springer.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.
- Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. 2008. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.
- Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer.
- Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. [ImageNet Large Scale Visual Recognition Challenge](#). *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.
- Jeffrey Mark Siskind. 1990. Acquiring core meanings of words, represented as jackendoff-style conceptual structures, from correlated streams of linguistic and non-linguistic input. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 143–156. Association for Computational Linguistics.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.
- Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.
- Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204.
- Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013.
- Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al. 2014. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30.

- Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Spatio-temporal person retrieval via natural language queries. *arXiv preprint arXiv:1704.07945*.
- Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yian-nis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics.
- Xuwang Yin and Vicente Ordonez. 2017. Obj2text: Generating visually descriptive language from object layouts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 177–187.
- Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 53–63.
- Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 3.