Video Object Grounding using Semantic Roles in Language Description (CVPR20)

Arka Sadhu¹
Kan Chen²
Ram Nevatia¹

¹USC Viterbi
School of Engineering
Department of Computer Science

²Facebook Research

CVPR
Seattle
Washington
June 16-18 2020
Query: **The man passes a ball to a group of kids**

Video Object Grounding: Localize the Objects in the Video referred in a language query
Query: The man passes a ball to a group of kids

As there is only one “Ball” in the video, it can be identified by a simple object detector

OBJECT RELATIONS ARE IGNORED!!!
Query: The man passes a ball to a group of kids

Consider object relations and detect the single object.

Current video grounding system ARE BEING IGNORED!!!
Query: **The man** passes **a ball** to **a group of kids**

- **Arg0**: The man
- **Verb**: passes
- **Arg1**: a ball
- **Arg2**: a group of kids

What if another ‘⚽️’ was present in the video? Will the model ground the correct ball?

**OBJECT RELATIONS ARE BEING IGNORED!!!**
Two-Step Process

1. Contrastive Sampling

2. Temporal and Spatial Concatenation
Contrastive Sampling

Arg0: man
Verb: petting
Arg1: dog

Q1: man petting dog
Contrastive Sampling

Arg0: man
Verb: petting
Arg1: dog

Q1: man petting dog
Contrastive Sampling

Arg0: man  
Verb: petting  
Arg1: dog

Q1: man petting dog

Q2: woman petting dog

Q3: man picking up dog

Q4: man petting cat
Method-1 TEMP: Temporal Concatenation

Method-2 SPAT: Spatial Concatenation along width
**Method-1 TEMP:** Temporal Concatenation

**Method-2 SPAT:** Spatial Concatenation along width

Merged Video contains multiple instances of the same object category!
Method-1 TEMP: Temporal Concatenation

Method-2 SPAT: Spatial Concatenation along width

Merged Video contains multiple instances of the same object category!

Forced to Utilize Object Relations to ground the correct instance!
Encode Object Relations via Self-Attention using Transformers

1. Add Multi-Modal Transformer

2. Use Relative Position Embedding
We encode Object Relations using Self-Attention via Transformer Networks.

Same objects can be related in multiple ways.
We encode Object Relations using Self-Attention via Transformer Networks.

Same objects can be related in multiple ways.

Learning object relations conditioned on the language input is helpful.
Transformers need positional embeddings
But Absolute Positions don’t matter in a video!
Transformers need positional embeddings
But Absolute Positions don’t matter in a video!

Self-Attention

\[ A(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k}) V \]
Transformers need positional embeddings
But Absolute Positions don’t matter in a video!

Self-Attention

\[ A(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k}) V \]

Use Relative Position Encoding (RPE)

\[ \text{pos}_A = [x_{tl}/W, y_{tl}/H, x_{br}/W, y_{br}/H, j/F] \]
\[ \Delta[h][A, B] = MLP(\text{pos}_A - \text{pos}_B) \]
\[ A(Q, K, V) = \text{SoftMax}((QK^T + \Delta[h]) / \sqrt{d_k}) V \]

Encodes the Relative Positions

Schematic of our Proposed VOGNet

VOGNet = Grounding Module + Object Tx + MultiModal Tx + RPE
ActivityNet SRL = Append semantic roles to ActivityNet Captions
+ Align with bounding boxes in ActivityNet Entities

ActivityNet-SRL available at https://github.com/TheShadow29/vognet-pytorch

Sentence: Person washes cups in a sink with water.

<table>
<thead>
<tr>
<th>Agent</th>
<th>Verb</th>
<th>Patient</th>
<th>Modifier</th>
<th>Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>washes</td>
<td>cups</td>
<td>in a sink</td>
<td>with water</td>
</tr>
<tr>
<td>Arg0</td>
<td>Verb</td>
<td>Arg1</td>
<td>ArgM-Loc</td>
<td>Arg2</td>
</tr>
</tbody>
</table>
GT5: Simplified Evaluation with 5 proposals per frame

86% Recall Rate +
Allowing Many Experiments +
Findings generalize to 100 Proposal per Frame as well
Training Grounding Systems on a Single Video doesn’t generalize at all!

It is very close to a simple object detector (which would get 0% in both TEMP, SPAT cases).

SVSQ: Single Video
TEMP: Temporal Concatenation
SPAT: Spatial Concatenation
Training Grounding Systems on a Single Video doesn’t generalize at all!

It is very close to a simple object detector (which would get 0% in both TEMP, SPAT cases).

SVSQ: Single Video
TEMP: Temporal Concatenation
SPAT: Spatial Concatenation
Training with TEMP, SPAT augmentations maintains performance on a single video setting, and improves generalization.
Training with TEMP, SPAT augmentations maintains performance on a single video setting, and improves generalization.
Training with TEMP, SPAT augmentations maintains performance on a single video setting, and improves generalization. There remains a considerable gap!
Augmentations with randomly sampled videos are competitive with contrastively sampled videos.
Interesting to note, augmentations with randomly sampled videos are competitive with contrastively sampled videos.

As expected, Random videos are much easier than contrastively sampled ones!
RPE improves ~1% performance
A single layer of MTx outperforms 3L OTx
1. We propose Video Object Grounding (VOG) with elevated role of Object Relations by temporal and spatial concatenation of the contrastive examples.

2. We release ActivityNet-SRL as a benchmark.

3. We also propose VOGNet which has Multi-Modal Transformer with Relative Position Encodings. Even with proposed contributions, there remains a large gap!

To foster reproducibility, we have open-sourced (on github) all models and logs to exactly reproduce the numbers reported in the paper.

Chat with us for more details!!

Email: asadhu@usc.edu
https://github.com/TheShadow29/vognet-pytorch