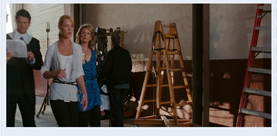
Generating Descriptions with Grounded and Co-Referenced People

Anna Rohrbach[1], Marcus Rohrbach[2,3], Siyu Tang[1,4], Seong Joon Oh[1], Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]UC Berkeley EECS [3] Facebook AI Research [4] Max Planck Institute for Intelligent Systems

Prior work:

Current clip



Someone strides to the window.

This work:

Previous clip Current clip

Sophia She and Jacob walk down the corridor.

Summary

Motivation

- Prior work ignores the person identity
- Prior work ignores the gender information
- Prior work can not localize the person
- Prior work can not resolve co-references

Contributions

- New task: video description with grounded and co-referenced people
- Joint approach which predicts: description + grounding + co-reference + gender
- Core: joint attention mechanism
- Automatically obtained attention supervision
- Extensive evaluation on the MPII Movie Description dataset [Rohrbach CVPR'15]

References

[Girshick ICCV'15] Fast r-cnn.
[Rohrbach CVPR'15] A dataset for movie description.
[Rohrbach GCPR'15] The long-short story of movie description.
[Rohrbach ECCV'16] Grounding of textual phrases in images by reconstruction.
[Tang CVPR'15] Subgraph decomposition for multi-target tracking.
[Vendatam CVPR'15] CIDEr: Consensus-based image description evaluation.
[Venugopalan ICCV'15] Sequence to sequence – video to text.

Overview

Step 0

- Head detection
  - R-CNN [Girshick ICCV'15] trained on heads
- Head tracking
  - 2 levels of clustering approach [Tang CVPR'15]

Step 1

- Automatic character-name linking
- Provides supervision for step 2

Input Output

Sophia Sophia

Step 2

- Relies on linking from step 1
- Description generation
  - With gender-specific labels (M-Name, F-Name, He, She)
- Character grounding
- Local co-reference resolution

Disclaimer! The names in the examples are only used for visualization purposes.

Input Output

Previous clip Current clip Previous clip Current clip

Sophia She and Jacob walk down the corridor.

Approach to step 1: Semi-supervised character-name linking

We apply GrounderR [Rohrbach ECCV'16]

At training time:

- Consider all sentence/clip pairs where at least one name is mentioned.
- If only one track & one name present – consider it a correct link (supervision).
- In a semi-supervised way learn to select a track, out of multiple proposals, for a given name X.

Ground-truth: Sophia enters her office.

Ground-truth: Sophia looks at him, nodding.

At test time:

- Given a name X and a set of tracks choose a correct track.

Approach to step 2: Joint description + grounding model

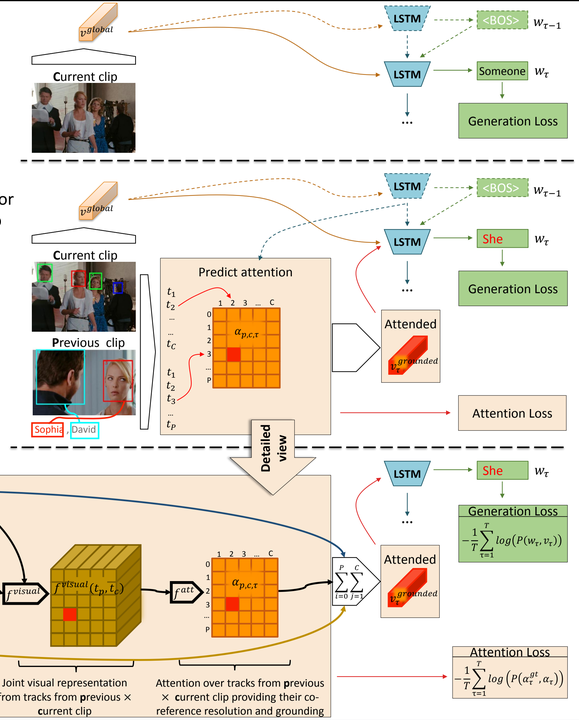
Baseline model [Rohrbach GCPR'15]

- Global video representation
  - Objects, actions, places
- Sentence generation loss

Our model overview

- Track-level representations for the current and previous clip
- Attention over current and previous tracks
- Person specific generation
- Additional attention loss

- Not visualized, but can be easily integrated:
  - Body features
  - Track statistics features



Dataset

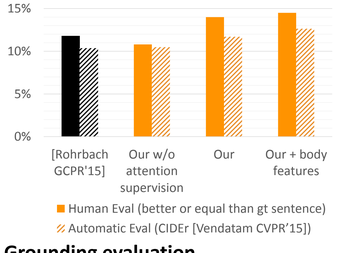
MPII Movie Description Co-ref+Gender

- Based on MPII Movie Description dataset [Rohrbach CVPR'15]
- We annotate character names and resolve he/she pronouns (match to names).
- We annotate gender for each name
- Every character mention, which appears in a previous sentence, is replaced with "He" / "She", otherwise with "MaleName" / "FemaleName".

Quantitative results

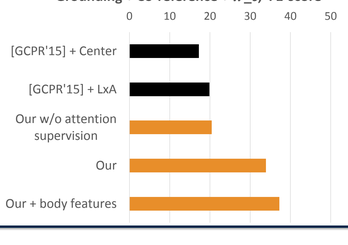
Description evaluation

- Human Eval: helpful for the blind criteria, where ≥2 our of 3 judges agree.



Grounding evaluation

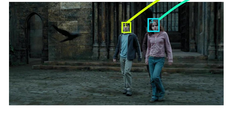
- Gender-specific label w_t (M-Name, F-Name, He, She)
- Character grounding
- Co-reference resolution
- Baseline: [GCPR'15] + heuristic grounding



Qualitative results

Comparison with state-of-the-art

- Description, gender, grounding



Our: Sophia and Jacob are walking towards the house.
[**Rohrbach GCPR'15**]: Someone is walking along the river, a large man in a long, surrounded by.
[**Venugopalan ICCV'15**]: Someone is in the back of a car.

Our: Sophia and Jacob are in the car.
[**Rohrbach GCPR'15**]: Someone looks at someone, then turns to someone.
[**Venugopalan ICCV'15**]: Someone is standing in the front garden of the car.

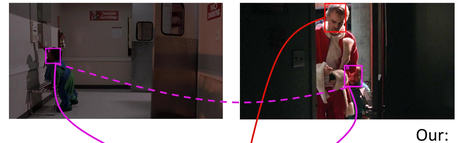
Our approach

- Description, gender, grounding, co-reference

Previous clip Current clip

Jacob Our: He walks away.

Previous clip Current clip



Jacob Our: Daniel and Jacob go to the door.

Previous clip Current clip

Jacob , Daniel Our: Daniel walks over to the man.