Generating Descriptions with Grounded and Co-Referenced People
Anna Rohrbach1, Marcus Rohrbach2,3, Siyu Tang1,4, Seong Joon Oh1, Bernt Schiele1

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

Prior work: Current clip
This work: Previous clip Current clip

Someone strides to the window.1
Sophia and Jacob walk down the corridor.

He walks away.


Overview

Step 0
• Head detection
• R-CNN [Girshick ICCV'15] trained on heads
• Head tracking

Step 1
• Automatic character-name linking
• Provides supervision for step 2

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

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

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

We apply Grounder [Rohrbach ICCV’16]
At training time:
• Consider all sentence/clip pairs where at least one name is mentioned.
• If only one track & 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.

At test time:
• Given a name X and a set of tracks choose a correct track.

Approach to step 2: Joint description + grounding model

Description dataset
Human Eval helpful for the blind criteria, where 2-3 out of 3 judges agree.

Grounding evaluation
• Gender specific label wi (M=Name, F=Name, He, She)
• Character grounding
• Co-reference resolution
• Baseline: [GCPR’15] + heuristic grounding

Grounding + Co-reference
Grounder = 0.7

Our approach
• Description, gender, grounding, co-reference

Qualitative results
Comparison with state-of-the-art
• Description, gender, grounding

Our Sophia walks to the car.
[Rohrbach GCPR’15] Sophia looks at someone, then turns to someone.
[Venugopalan ICCV’15]: Someone is standing in the front of the car.

References
• [Girshick ICCV’15] Fast r-CNN.
• [Rohrbach CVPR’15][GCPR]: Consensus-based image description evaluation.
• [Venugopalan ICCV’15] Sequence to sequence – video to text.