Enriching Video Captions With Contextual Text

Philipp Rimle, Pelin Dogan-Schönberger, Markus Gross
ETH Zurich, Media Technology Center

1 Motivation

Someone speaks to **people**.

Elrond addresses the **council**.

Someone walks in front of **people**.

Frodo walks towards the **stone plinth**.

Transforming a video content to a textual domain helps applications like video indexing, navigation and retrieval, automatic video search or human-robot interaction.

Generated video captions/descriptions, based on the visual input, tend to be generic and background knowledge about who, where, when and why is missing.

In this research project, we investigated how to improve automatically generated video captions by providing additional context.

2 Model Overview

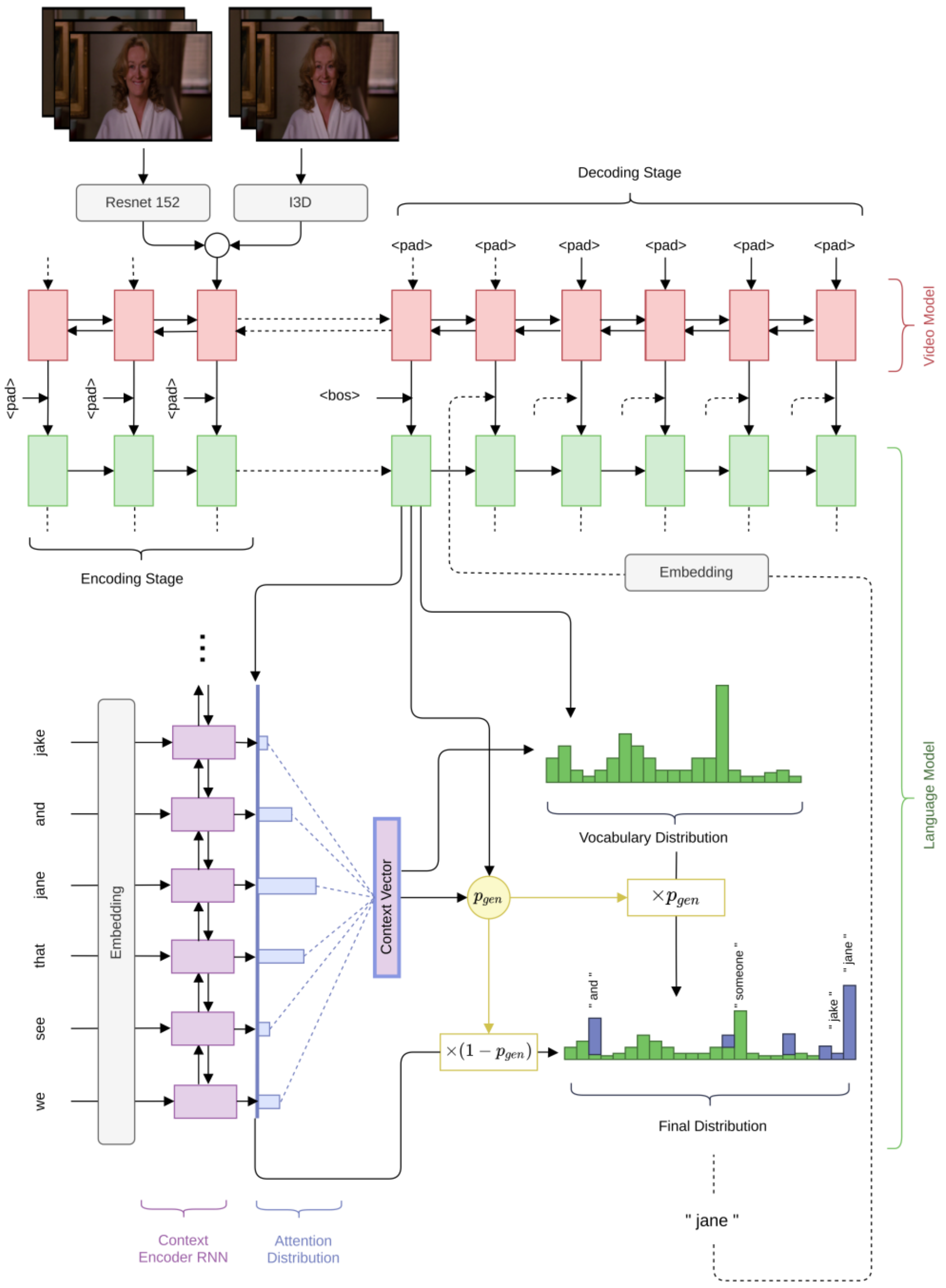


Fig. 1. Overview of our model. A CNN architecture is used to extract visual features per video frame. These are passed to a end-to-end, sequence-to-sequence recurrent neural network, modeling the video captioning. A pointer generator network allows the model to attend over additional contextual data and mine relevant background knowledge. As a result, the model is able to generate more specific video captions, including names and locations.

3 Challenges

Video Captioning

- Open domain
- Diverse set of objects, actions and scenes
- Interaction of different objects and the fine motion details

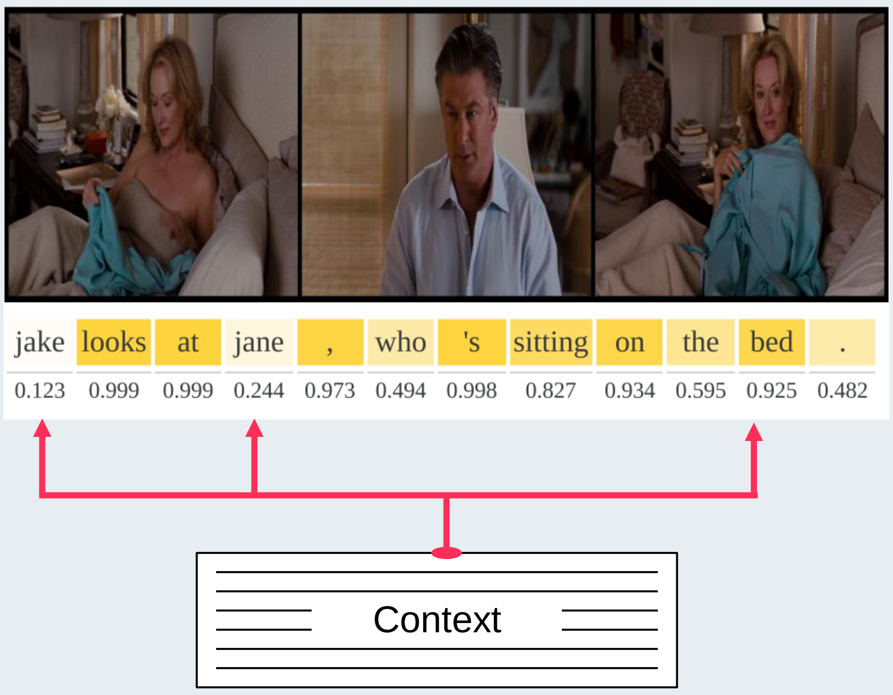
Context Mining

- Generate unknown words (not seen while training)
- Understand the contextual text and learn how to attend over it

Dataset

- No large dataset with visual and contextual input
- Augment existing video dataset with background knowledge in form of text input

4 Results



jake	looks	at	jane	,	who	's	sitting	on	the	bed	.
0.123	0.999	0.999	0.244	0.973	0.494	0.998	0.827	0.934	0.595	0.925	0.482

Fig. 4. An example where the model is able to generate a more specific video caption, including the names "jake" and "jane", based on contextual text in form of a movie script scene.