Introduction to the Semantic-Assisted Video Captioning Model Trained with Scheduled Sampling

HAORAN CHEN and HUIRAN YU*, Tsinghua University, China

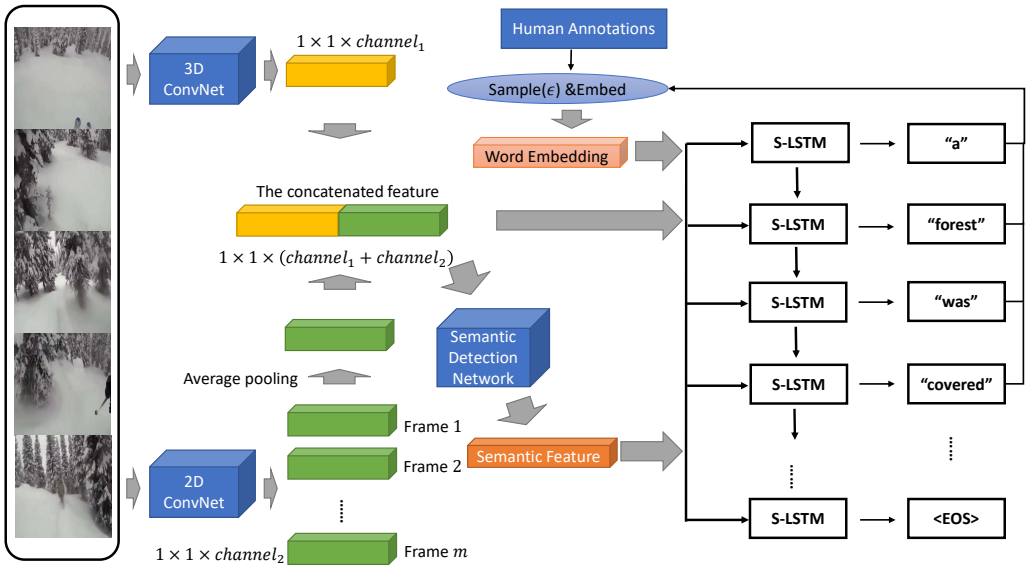

Fig. 1. Model framework for semantic-assisted video captioning model with scheduled sampling.

Recurrent neural networks (RNN) or Transformer-based models are applied to generate a caption for a video clip. In the Pre-training for Video Captioning Challenge of ACM Multimedia 2020, we apply a semantic-assisted long short-term memory (LSTM) model with scheduled sampling to solve the challenging.

Additional Key Words and Phrases: datasets, neural networks, video captioning, semantic tagging

ACM Reference Format:
Haoran Chen and Huiran Yu. 2020. Introduction to the Semantic-Assisted Video Captioning Model Trained with Scheduled Sampling. 1, 1 (July 2020), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' address: Haoran Chen, chenhaoran.earth@live.com; Huiran Yu, yuhuiran@gmail.com, Tsinghua University, Beijing, China, 100084.

XXXX-XXXX/2020/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Our model is based on the encoder-decoder-framework (Figure 1). The encoder is composed of ResNeXt, Efficient Convolutional Network (ECO) and a semantic detection network. The decoder consists of a semantic-assisted LSTM. The teacher forcing with scheduled sampling is exploited to train the model.

The video feature extracted by ResNeXt is of 2048-dim and the visual feature extracted by ECO is of 1536-dim. We conduct the experiments with 1024-dim LSTM or 2048-dim LSTM, separately. We train the model for 50 epochs and the sampling probability is determined dynamically as $\epsilon = 0.008 \times epoch$.

Four sets of data are used as the training dataset, separately: MSR-VTT, MSR-VTT + MSVD, MSR-VTT + MSVD + Pretraining and MSR-VTT + MSVD + Pretraining-mini, where Pretraining denotes the dataset provided by the Pre-training for Video Captioning Challenge and Pretraining-mini denotes a subset of the Pretraining. The overall performance of the models trained by those sets of data is shown as follows: $MSR-VTT+MSVD > MSR-VTT+MSVD+Pretraining-mini > MSR-VTT > MSR-VTT+MSVD+Pretraining$. The difference between the underlying visual-text relations in video clips and gif pictures deteriorates the model performance when those two datasets are combined together as the training data.

Unfortunately, we fail to find the suitable pretraining tasks to solve the video captioning challenge, since we are short in time and energy.