Title of My Paper: 4 page max not counting references

Li Ruojun [1] **Tesia Shizume** [1] **Chao Wang** [1]

Abstract

The accurate automated captioning videos is an important task for both making video collections more accessible to human users and for making video content more accessible to the visually impaired. This task can be segmented into two subtasks: audio encoding into language and visual encoding into language. In this study we focus on the task of expressing the visual content of the videos in the English language. One of the more common frameworks for attempting this task is an encoder-decoder framework. We use this framework to implement and train from scratch a video captioning RNN (GAN) to investigate a current state of the art technique. To improve the performance of the technique we experiment with augmenting this approach with some of the newest pre-trained CNN models for our initial image feature extraction and use modern regularization methods. To demonstrate the effectiveness of our approach we test our method on the MSR-VTT dataset. Our experimental results show...

1. Introduction

Accurate automated captioning videos is an important yet computationally challenging task for both making video collections more accessible to the general public and for making video content more accessible to the visually impaired. Existing methods Why ours is good

1.1. Research contributions

Describe clearly and concisely what your paper **uniquely** contributes to the research community, i.e., that has never been done before.

1.2. Data

There are many openly available datasets of video clips with sentence descriptions. To reduce the search space for an appropriate dataset we chose to eliminate datasets with fewer than 1,000 videos. This left us with the datasets found in the Table in Appendix A: Candidate Datasets.

Then given the time constraints of the project we further filtered this subset of datasets for medium length videos that would be appropriate for shorter descriptions and that we could train well with a variety of parameters. This left us with MSVD, MSR-VTT, and Charades. We chose to use MSR-VTT since it had more videos than MSVD and had diverse videos, belonging to the open domain category, which would be a more complicated problem.

2. Related Work

Overview encoder-decoder framework; CNN-RNN, RNN-RNN, Reinforcement learning Hierarchical LSTM with Adjusted Temporal Attention for Video Captioning

Video Captioning by Adversarial LSTM

3. Proposed Method

We propose a CNN-RNN (GAN) based approach using the encoder-decoder framework. This category of approaches leverage advanced pre-trained CNNs to efficiently and accurately preprocess the image component of the video data and trains an RNN to collect data from multiple frames and then generate the sentences. We augment this approach with some of the newest pre-trained CNN models for our initial image feature extraction.

3.1. Preprocessing

We consider data cleaning an important step to facilitate training ANNs that will produce accurate results. Thus we preprocessed the dataset to remove data instances with typos to ensure compatibility with the GloVe word representation. This reduced the dataset to approximately 7,000 videos and a vocabulary of 6,248 words. GloVe vectors are

[1] Worcester Polytechnic Institute. Correspondence to: Cieua Vvvvv <c.vvvvv@wpi.edu>.

CS541 Special Topics Course Final Project, Worcester Polytechnic Institute.

3.2. Architecture

CNN -¿ RNN

3.3. Metrics

While automatic evaluations of the quality of machine-generated natural language sentences is an area of active research, there are several generally accepted metrics, METEOR, CIDEr, BLEU, ROUGE_L. For our project, we choose to focus on two of these metrics that have high correlations with human judgments, METEOR (Lavie and Agarwal 2007) and CIDEr (Vedantam, Lawrence, and Parikh 2015). METEOR has the advantage of taking into account synonyms in its n-gram matching. While CIDEr uses TF-IDF(Term Frequency and Inverse Document Frequency) weighted n-gram similarity.

4. Experiment

5. Results

6. Discussion

7. Conclusions and Future Work

References

A. Dataset Candidates

Dataset	Domain	# classes	# videos	Avg. len. (sec)	# clips	# sentences	# words	vocab.	len. (hrs)
MSVD	open	218	1,970	10	1,970	70,028	607,339	13,010	5
MSR-VTT	open	20	7,180	20	10,000	200,000	1,856,523	29,316	4
Charades	human	157	9,848	30	-	27,847	-	-	82
YouCook 2	cooking	89	2,000	316	15,400	-	2,600	176	
ActivityNet Captions	open	-	20,000	180	-	100,000	1,348,000	-	849
ActivityNet Entities	social media	-	14,281	180	52,000	-	-	-	-
Youtube Clips	open	-	1,967	9	-	80,839	-	12,766	-
TumblerGIF	open	-	100,000	3.66	100,000	128,000	-	12,228	103