Review 8 - 1

Unsupervised Learning of Spoken Language with Visual Context

Summary

This paper presents deep neural network architecture learning visual features and free-form spoken audio captions into same space. This paper investigates novel neural network architectures for the purpose of learning high-level semantic concepts across both audio and visual modalities. Contextually correlated streams of sensor data from multiple modalities - in this case a visual image accompanied by a spoken audio caption describing that image - are used to train networks capable of discovering patterns using otherwise unlabelled training data. For example, these networks are able to pick out instances of the spoken word “water" from within continuous speech signals and associate them with images containing bodies of water. The networks learn these associations directly from the data, without the use of conventional speech recognition, text transcriptions, or any expert linguistic knowledge whatsoever.

Main Contributions:

1. Provide a dataset
2. Present a novel neural network architecture learning visual features free-form audio.
3. This represents a new direction in training neural networks that is a step closer to human learning, where the brain must utilize parallel sensory input to reason about its environment.

Pros:

1. No need for linguistic and language knowledge.
2. In the experiment, this paper also shows which regions of the spectrogram the model believes are highly relevant to the image.

Cons:

1. Doesn’t compare with the method using textual language methods
2. The dataset is relatively small.
3. For same audio captions, this paper doesn't compare the results with audios from different people.
4. The performance of the search isn't very ideal.

Potential for Improvement:

1. Comparing data with same caption in different languages could be interesting and might prove the advantages of this model.

Review 8 - 2

Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences

Summary

The basic problem of this paper is interpreting free-form instructions (especially in unknown environments) to action sequences based only on knowledge of the local and observable environment. This task is very challenging due to their ambiguity and complexity. This paper presents an end-to-end, sequence-to-sequence approach to mapping natural language navigational instructions to action plans given the local, observable world state, using a bidirectional LSTM-RNN model with a multi-level aligner. The encoder-decoder architecture is a general sequence-to-sequence LSTM-RNN architecture. However, this paper proposed a novel multi-level aligner to represent the observable context during the learning process. Specifically, the model not only relies on the previous hidden annotation variable, but also considers the original low-level word representation. It is this improvement makes this model outperforms the previous work on the same benchmark.

Main Contributions:

1. Achieves the best results reported to-date on a benchmark single-sentence dataset and competitive results for the limited-training multi-sentence setting.
2. Ablation of different components gives more insights about why the model outperforms others. This method could be used on other problems.

Cons:

1. The performance for multi-sentence is much worse than single-sentence. This problem could be addressed by learning a paragraph (multi-sentence) by structured representations.

Potential for Improvement:

1. It’s natural to consider using Reinforcement Learning to tackle this problem.