Improvements to: Hierarchical LSTMs with Adaptive Attention for Visual Captioning

CMPE544 - Pattern Recognition

Elif Yılmaz - Gizem Günbal
Phase2 Presentation

17 Feb, 2021

Summary of the Problem

The **goal** of this paper is to generate a natural language sentence automatically for a given image or video by using a Hierarchical LSTMs with adaptive attention model.
Different **data sets** are used for video and image captioning:
For videos:

- The Microsoft Video Description Corpus (MSVD) - 1,970 short video clips
- MSR Video to Text (MSR-VTT) - 10,000 web video clips
- Large Scale Movie Description Challenge (LSMDC) - 118,081 video clips from 202 movies

For images:

- COCO - 164,775 images
- Flickr30K - 31,783 images

What is a hLSTMs with adaptive attention model?

This model includes three components:

1. CNN Network as Encoder
2. Hierarchical LSTMs as Decoder
3. Maximum Likelihood Estimation as Losses

Limitations & Proposed Solution

- **Using only 2 layers of LSTM**
  The bottom LSTM layer is used to efficiently decode visual features, and the top LSTM is focusing on mining deep language context information for video captioning in our paper.

Limitations & Proposed Solution

- **Using only 2 layers of LSTM**
  The bottom LSTM layer is used to efficiently decode visual features, and the top LSTM is focusing on mining deep language context information for video captioning in our paper.
- **Solution:** the usage of more LSTM layers

Limitations & Proposed Solution

- **CNN as encoder to feature extraction**
  CNN networks are used to extract features for images and videos. The goal of this is to get compact features and then make visual data suitable for decoding part.

Limitations & Proposed Solution

- **CNN as encoder to feature extraction**
  CNN networks are used to extract features for images and videos. The goal of this is to get compact features and then make visual data suitable for decoding part.
- **Solution:** CNN with LSTM based model for feature extraction contains higher level concepts to look for visual features to text directly. This model helps to get better weight matrix for extracting features.

Limitations & Proposed Solution

- **Focus on the features extracted from the scenes**
  In video captioning, it is important to capture the relationship between scenes. Our paper focuses on the features extracted from the scenes and cannot capture the long-range temporal dependencies.

Limitations & Proposed Solution

- **Focus on the features extracted from the scenes**
  In video captioning, it is important to capture the relationship between scenes. Our paper focuses on the features extracted from the scenes and cannot capture the long-range temporal dependencies.
- **Solution:** building a graph-based reasoning framework representing videos as graph of objects $\rightarrow$ similarity graph

Hypothesis

- Our hypothesis is to use an improved hierarchical LSTMs with adaptive attention model for visual captioning.
- This model has improved parts in terms of using more layers LSTMs, spatio-temporal object interaction and CNN+LSTM as encoder while extracting features.
- This model includes three components
  1. CNN+LSTM Network as Encoder
  2. Spatio-Temporal Object Interaction as Decoder
  3. Hierarchical LSTMs as Decoder
  4. Maximum Likelihood Estimation as Losses

What did we learned during this project?

- What are RNN, CNN and LSTM?
- Connection between computer vision and natural language processing
- Literature Review