Neural Language Modeling with Visual Features

Abstract

Multimodal language models attempt to incorporate non-linguistic features for the language modeling task. In this work, we extend a standard recurrent neural network (RNN) language model with features derived from videos. We train our models on data that is two orders-of-magnitude bigger than datasets used in prior work. We perform a thorough exploration of model architectures for combining visual and text features. Our experiments on two corpora (YouCookII and 20bn-something-something-v2) show that the best performing architecture consists of middle fusion of visual and text features, yielding over 25% relative improvement in perplexity. We report analysis that provides insights into why our multimodal language model improves upon a standard RNN language model.

Introduction

INLINEFORM0 Work performed while the author was an intern at Google.

Language models are vital components of a wide variety of systems for Natural Language Processing (NLP) including Automatic Speech Recognition, Machine Translation, Optical Character Recognition, Spelling Correction, etc. However, most language models are trained and applied in a manner that is oblivious to the environment in which human language operates BIBREF0 . These models are typically trained only on sequences of words, ignoring the physical context in which the symbolic representations are grounded, or ignoring the social context that could inform the semantics of an utterance.

For incorporating additional modalities, the NLP community has typically used datasets such as MS COCO BIBREF1 and Flickr BIBREF2 for image-based tasks, while several datasets BIBREF3 , BIBREF4

, BIBREF5 , BIBREF6 , BIBREF7 have been curated for video-based tasks. Despite the lack of big datasets, researchers have started investigating language grounding in images BIBREF8 , BIBREF9 , BIBREF10 and to lesser extent in videos BIBREF11 , BIBREF1 . However, language grounding has focused more on obtaining better word and sentence representations or other downstream tasks, and to lesser extent on language modeling.

In this paper, we examine the problem of incorporating temporal visual context into a recurrent neural language model (RNNLM). Multimodal Neural Language Models were introduced in BIBREF12 , where log-linear LMs BIBREF13 were conditioned to handle both image and text modalities. Notably, this work did not use the recurrent neural model paradigm which has now become the de facto way of implementing neural LMs.

The closest work to ours is that of BIBREF0 , who report perplexity gains of around 5–6% on three languages on the MS COCO dataset (with an English vocabulary of only 16K words).

Our work is distinguishable from previous work with respect to three dimensions:

Model

A language model assigns to a sentence INLINEFORM0 the probability: INLINEFORM1

where each word is assigned a probability given the previous word history.

For a given video segment, we assume that there is a sequence of INLINEFORM0 video frames represented by features INLINEFORM1 , and the corresponding transcription INLINEFORM2 . In practice, we assume INLINEFORM3 since we can always assign a video frame to each word by

replicating the video frames the requisite number of times. Thus, our visually-grounded language model models the probability of the next word given the history of previous words as well as video frames: INLINEFORM4

Combining the text and video modalities

There are several options for combining the text and video modalities. We opt for the simplest strategy, which concatenates the representations. For a word embedding INLINEFORM0 and corresponding visual representation INLINEFORM1 , the input to our RNNLM will be the concatenated vector INLINEFORM2 . For the examples where we were unable to compute visual features (see Section § SECREF3), we set INLINEFORM3 to be a zero-vector.

In addition to concatenating the word and visual embedding, we explore two variants of our model that allow for a finer-grained integration of the two modalities:

In this case, the RNNLM is given as input a vector INLINEFORM0 that is a weighted sum of the two embeddings: INLINEFORM1

where INLINEFORM0 are learned matrices.

Here, we apply the intuition that some words could provide information as to whether or not the visual context is helpful. In a simplistic example, if the word history is the article “the," then the visual context could provide relevant information needed for predicting the next word. For other word histories, though, the visual context might not be needed or be even irrelevant for the next word prediction: if the previous word is “carpe", the next word is very likely to be “diem", regardless of visual context. We implement a simple weighting mechanism that learns a scalar weight for the visual embedding prior to concatenation

with the word embedding. The input to the RNNLM is now INLINEFORM0 , where: INLINEFORM1

This approach does not add any new parameters to the model, but since the word representations INLINEFORM0 are learned, this mechanism has the potential to learn word embeddings that are also appropriate for weighting the visual context.

Location of combination

We explore three locations for fusing visual features in an RNNLM (Figure). Our Early Fusion strategy merges the text and the visual features at the input to the LSTM cells. This embodies the intuition that it is best to do feature combination at the earliest possible stage. The Middle Fusion merges the visual features at the output of the 1st LSTM layer while the Late Fusion strategies merges the two features after the final LSTM layer. The idea behind the Middle and Late fusion is that we would like to minimize changes to the regular RNNLM architecture at the early stages and still be able to benefit from the visual features.

Data and Experimental Setup

Our training data consist of about 64M segments from YouTube videos comprising a total of INLINEFORM0 B tokens BIBREF14 . We tokenize the training data using a vocabulary of 66K wordpieces BIBREF15 . Thus, the input to the model is a sequence of wordpieces. Using wordpieces allows us to address out-of-vocabulary (OOV) word issues that would arise from having a fixed word vocabulary. In practice, a wordpiece RNNLM gives similar performance as a word-level model BIBREF16 . For about INLINEFORM1 of the segments, we were able to obtain visual features at the frame level. The features are 1500-dimensional vectors, extracted from the video frames at 1-second intervals, similar to those used for large scale image classification tasks BIBREF17 , BIBREF18 . For a INLINEFORM2

-second video and INLINEFORM3 wordpieces, each feature is uniformly allocated to INLINEFORM4 wordpieces.

Our RNNLM models consist of 2 LSTM layers, each containing 2048 units which are linearly projected to 512 units BIBREF19 . The word-piece and video embeddings are of size 512 each. We do not use dropout. During training, the batch size per worker is set to 256, and we perform full length unrolling to a max length of 70. The INLINEFORM0 -norms of the gradients are clipped to a max norm of INLINEFORM1 for the LSTM weights and to 10,000 for all other weights. We train with Synchronous SGD with the Adafactor optimizer BIBREF20 until convergence on a development set, created by randomly selecting INLINEFORM2 of all utterances.

Experiments

For evaluation we used two datasets, YouCook2 and sth-sth, allowing us to evaluate our models in cases where the visual context is relevant to the modelled language. Note that no data from these datasets are present in the YouTube videos used for training. The perplexity of our models is shown in Table .

Conclusion

We present a simple strategy to augment a standard recurrent neural network language model with temporal visual features. Through an exploration of candidate architectures, we show that the Middle Fusion of visual and textual features leads to a 20-28% reduction in perplexity relative to a text only baseline. These experiments were performed using datasets of unprecedented scale, with more than 1.2 billion tokens – two orders of magnitude more than any previously published work. Our work is a first step towards creating and deploying large-scale multimodal systems that properly situate themselves into a given context, by taking full advantage of every available signal.