Image-sensitive language modeling for automatic speech recognition

Kata Naszádi[1] *, Youssef Oualil[2], and Dietrich Klakow[2]

[1] Amazon, Aachen, Germany
[2] Spoken Language Systems (LSV), Saarland Informatics Campus, Saarland University, Saarbrücken, Germany
naszadik@amazon.com, {youalil, dietrich.klakow}@lsv.uni-saarland.de

Abstract. Typically language models in a speech recognizer just use the previous words as a context. Thus they are insensitive to context from the real world. This paper explores the benefits of introducing the visual modality as context information to automatic speech recognition. We use neural multimodal language models to rescore the recognition results of utterances that describe visual scenes. We provide a comprehensive survey of how much the language model improves when adding the image to the conditioning set. The image was introduced to a purely text-based RNN-LM using three different composition methods. Our experiments show that using the visual modality helps the recognition process by a 7.8% relative improvement, but can also hurt the results because of overfitting to the visual input.

Keywords: multimodal speech recognition, multimodal language model

1 Introduction

Multimodal neural language models have been widely utilized for image captioning, but their effectiveness for other language modeling tasks is yet to be studied. The language modeling module of an automatic speech recognition pipeline could also benefit from the visual modality if the speaker refers to the visual surroundings. We implemented situated speech recognition by rescoring recognition results using multimodal language models. Natural language generation models, such as image captioning, tend to focus on the more frequent events, typically limiting the vocabulary to be smaller than 10 thousand words. For applications where the language model is used for estimating the likelihood of natural utterances, it is important to have a good probability estimate for rare events. In our language modeling experiments we aimed for high lexical coverage.

Using three different neural architectures we explore how much information image-conditioned models gain from the image. This is achieved by removing the image as an input while keeping the rest of the architecture as intact as

* This work was done while Kata Naszádi was at the Spoken Language Systems group at Saarland University.

possible. Creating purely text-based baselines also sheds light on the quality of image-captioning datasets with respect to the variability of the language used to describe the images.

2 Models

Two types of neural architectures have been implemented, both of them directly inspired by successful image captioning models. In one of them the image is only presented to the recurrent cell once. The other method only feeds textual data to the recurrent cell, then it uses the image in each time step to rescore the the output of the RNN, so the purely text based distribution $P(w|h)$ coming from the RNN becomes conditioned on the image $P(w|h, i)$. We tried two different methods for composing the output RNN-cell with the image feature vector: concatenation and compact bilinear pooling.

We used the Very Deep Convolutional Network with 16 hidden layers [1] for image feature extraction. The hidden activations of the last bottle-neck layer were used to obtain image representations of 4096 dimensions. The image feature vectors were kept fixed during training.

2.1 Text-based baseline (NI)

We trained two uni-modal language models to match the multimodal models as closely as possible. Both models are single-layer RNN-LSTM networks with vocabulary-sized softmax output layer. The word embeddings were set to be the same size as the hidden unit. The only difference between the two baseline models is the size of the hidden layer: 400 and 800 nodes. The models containing no image input will be referred to by the acronym NI.

2.2 Feeding the image to the recurrent unit (SaT)

The first architecture we implemented is a slightly adapted version of the Show and Tell (SaT) model [2]. We only changed the hidden size to be 400 units and increased the size of the output layer to match our vocabulary. The architecture builds on a standard recurrent language model with the addition of the the image as the input before the start of sentence symbol. The extra parameters introduced by this model compared to a unimodal RNN are the weights of the affine transformation $W_v \in \mathbb{R}^{d \times 4096}$ that maps the image vector v to the input size d. The input of the RNN before the start of sentence symbol x_{t-1} can be computed as:

$$x_{t-1} = W_v x_v \tag{1}$$

This architecture lends itself to being compared to a version without the image. In order to test how much perplexity reduction is due to the image, one simply needs to skip time step $t-1$ and start with the START symbol that denotes the beginning of a sentence.

2.3 Multimodal composition after the hidden unit

Concatenation (Concat) The second group of multimodal architectures compose the image with the output of the recurrent cell in each time step. The first model within this group implements the multimodal composition of the image vector and the RNN as concatenation. The concatenated layer is then directly followed by the softmax output layer. In this case, it is easy to see that the weight matrix following the activations of the concatenated layer can be split into two separate matrices.

$$r = W_v v + W_w h_t \tag{2}$$

The weight matrix is decomposed into two matrices, W_v operates on the image v, while W_w transforms the output of the LSTM unit h_t ; $W_{softmax} = [W_v, W_w]$. r contains the scores for each word in the vocabulary before normalization. The final score can be broken down into the contribution of the image and the textual input.

Compact Bilinear Pooling (CBP) In order to exploit more interactions between the two modalities, we also implemented bilinear pooling for the multimodal composition. Bilinear pooling is an unbiased estimator of the outer product. The upper bound of the variance of the estimation is inversely proportional to the size of the lower order estimation d. We implemented two bilinear pooling models, the first has $d = 400$ as the size of the hidden unit and the lower order estimation, the second has d set to 800 in order to decrease the variance. For the details of the algorithm please see [3]. Note that compact bilinear pooling does not add any trainable variables to the model compared to the text-based baseline model.

3 Experiments

For our experiments, we considered two famous image-captioning datasets: MSCOCO [4] and Flickr30k [5]. The vocabulary has been determined independently of the captions; it contains the 100 thousand most frequent words from the 1 Billion Word Language Model Benchmark by Chelba et al. [6].

In order to illustrate the quality of the captions from a language modeling perspective, a 3-gram Kneser-Ney smoothed language model has been trained on the captions. The estimation of the language models was carried out by the KenLM toolkit [7]. All punctuation symbols were removed and pruning was disabled.

3.1 Perplexity results

The 3-gram baseline perplexities in 1 illustrate that the language of the captions is very simplistic. As a comparison, the perplexity on a $25M$-word held-out portion of the Gigaword text corpus [8] is 144.6. The predictability of the language

Table 1. Perplexity results on different datasets. The number after the name of the model indicates the size of the hidden layer.

Model	Flickr30k	MSCOCO
KN-smoothed 3-gram	63.5	24.7
NI-400	33.2	17
SaT-400	23	12
CBP-400	46.5	16
Concat-400	23.1	11.3
NI-800	32.9	16.8
CBP-800	27.1	14.1
CONCAT-800	23.9	11.2

holds especially true for the MSCOCO even though the size of this dataset is more twice than that of the Flickr30k dataset.

The results show that on Flickr30k dataset Compact Bilinear Pooling only reduces perplexity if the size of the lower order estimate is big enough. With 400 hidden units the model without the picture (NI-400) outperforms pooling (CBP-400). We can see the benefit of using the image once the hidden size is large enough as in (CBP-800).

On the MSCOCO dataset there is a slight improvement even when the hidden size is only 400. The reason for this may come down to the fact that there is not a lot of variance in the textual vectors to begin with. It could also be the case that the images only cover a very limited set of visual scenes, but we ran no experiments in order to prove this point.

It is also clear from the results that concatenation always outperforms compact bilinear pooling. We argued for compact bilinear pooling because it is able to exploit interactions between all dimensions of the two modalities, but the method introduces a large estimation error due to the vastly different size of the composed vectors. The results also suggest that such interactions might not play a crucial role. The SaT-400 model performs closely to the Concat-400 model, even though the former is capable of learning non-linear interactions between the modalities.

3.2 Ratio of loss per part-of-speech tag

The CNN image encoder was trained on an object recognition task, so it would be reasonable to expect that most of the perplexity reduction is due to nouns. Table 2 shows the ratio of the loss between the models SaT-400H and NI-400H broken down to different part of speech categories. The captions were tagged using the Stanford log-linear part-of-speech tagger [9]. For a specific POS-tag each row displays the following ratio:

$$\frac{\sum_{w:POS(w)=pos} -\log(P_{WI}(w|h))}{\sum_{w:POS(w)=pos} -\log(P_{NI}(w|h))} \tag{3}$$

$P_{WI}(w|h)$ is the probability of the word according to the model that uses the image, and $P_{NI}(w|h)$ is the same probability estimate without the image. As

Table 2. Ratio of loss per part of speech tag category between the models with and without the image.

ADVERBS	96%
MODALS	100%
PARTICLES	94%
PREPOSITIONS	92%
NOUNS	94%
TO	97%
PRONOUNS	96%
ADJECTIVES	89%
VERBS	92%
DETERMINERS	96%

the results show, the performance is improved across almost all part-of-speech categories. It may only be the content words that get detected from the image, but predicting these words correctly will help the language model to make more accurate predictions for the other word categories too. Given the list of strings, for example *"dog, frisbee"*, there is only a limited way to combine these words into a fully formed sentence. It is also clear to see that the modality of a sentence can not be decided based on visual input.

3.3 Automatic Speech Recognition rescoring experiments

The automatic speech recognition experiments were carried out using the MIT Flickr Audio Caption Corpus [10]. 5000 spoken captions were used to tune the acoustic scale and the interpolation weights between the original background language model and the recurrent language models trained on the captions. We report the final results on a test set of 5000 spoken captions.

The first-pass decoding was performed using the HUB4 trigram language model [11]. As a baseline, the 300 best hypotheses were rescored with the neural language model that was only trained on the captions, without using the image (NI-400). This is necessary to account for the effect of the domain-specific language. For image-sensitive rescoring we used the SaT-400 model.

Table 3. Word error rates using the model trained exclusively on the captions (NI-400) and the image-sensitive language model (SaT-400).

	WER	acoustic-weight	RNN-weight
NI-400	37.04%	0.08	1
SaT-400	36.31%	0.1	1
NI-400	34.80%	0.08	0.9
SaT-400	32.08%	0.1	0.8

Both the image sensitive and the regular language model perform better when linearly interpolated with the 3-gram broadcast news language model that was originally used to generate the 300-best list. The performance of the first-pass

language model sets limitations for the final word error-rate. The HUB4 trigram model is trained predominantly on news data, which is not similar enough to the domain of the captions. One could achieve even better performance with a stronger first-pass decoder.

Qualitative analyses of the decoding results Figure 1 shows positive examples where the image-sensitive model seems to have successfully recognized objects from the picture, thus helping the recognition process.

True transcript: Mechanics preparing a plane for departure. **Image-sensitive**: Mechanics preparing a plane for the past. **No image**: Mechanics preparing a clean for the pasture.

True transcript: Two people making their way between rocks. **Image-sensitive**: Two people making their way between rocks. **No image**: Two people making their way between walks.

True transcript: A man fishing under an umbrella. **Image-sensitive**: A man fishing under an umbrella. **No image**: A man fishing under a number.

Fig. 1. Recognizing objects from the images helps the decoding.

On the downside, the image model is prone to overfitting to the image. Figure 2 shows a clear example of this effect. The visual setting depicting a dog gets associated with the word "pizza" during training and rescoring gives too high probability to the sentence containing this word. This effect can be reduced

A dog is standing in front of a pizza parlor and the man is smiling looking at him. The pizza maker is laughing at the white dog watching him through the window. A small brown and white dog looks in the window of Driggs Pizza. A lost dog trying to find some food in a pizza restaurant. A dog is looking in the window of a pizza parlor .

True transcript: A man on a bench feeds a dog. **Image-sensitive**: A man on a bench pizza. **No image**: A man on a bench feeds a dog.

Fig. 2. The image sensitive language model overfits to the training image.

by interpolating with a purely text-based language model that was trained on considerably more data. As illustrated by the optimal interpolation weights, the image model benefits more from the 3-gram language model. Supporting the image-sensitive language model with a richer, text-based model reduces, but does not eliminate this problem.

4 Conclusions

In this paper we set out to explore the possible benefits of introducing the visual modality to language modeling. We showed that adding the image to the conditioning set helps reduce perplexity up to 30% relative to the baseline.

Conditioning on the image helped decrease word error rate from 34.8% to 32.08%. In some cases the image-sensitive model fails to identify the participants

and actions in the visual setting and copies training sentences based on superficial visual similarity. One reason for this is that the datasets are not large enough, and the model is not presented with a sufficient variety of scene and description combinations. We also believe that profound image understanding cannot be achieved by using a global image descriptor and optimizing on maximizing the likelihood of the descriptions.

Future work could further explore the benefit of using the visual modality by using a better first-pass decoder and exploring multimodal language models that achieve a more effective grounding in the visual modality.

References

1. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. (2015)
2. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3156–3164
3. Pagh, R.: Compressed matrix multiplication. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ACM (2012) 442–451
4. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
5. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2641–2649
6. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P.: One billion word benchmark for measuring progress in statistical language modeling. CoRR **abs/1312.3005** (2013)
7. Heafield, K., Pouzyrevsky, I., Clark, J.H., Koehn, P.: Scalable modified Kneser-Ney language model estimation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria (August 2013) 690–696
8. Graff, D., Cieri, C.: English gigaword, ldc catalog no. LDC2003T05. Linguistic Data Consortium, University of Pennsylvania (2003)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
10. Harwath, D., Glass, J.: Deep multimodal semantic embeddings for speech and images. arXiv preprint arXiv:1511.03690 (2015)
11. Sankar, F.W.A.S.A.: Hub4 language modeling using domain interpolation and data clustering. In: DARPA Speech Recognition Workshop, Citeseer (1997) 147