LOOK, LISTEN, AND DECODE: MULTIMODAL SPEECH RECOGNITION WITH IMAGES

Felix Sun, David Harwath, and James Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
{felixsun, dharwath, glass }@mit.edu

ABSTRACT

In this paper, we introduce a *multimodal speech recognition* scenario, in which an image provides contextual information for a spoken caption to be decoded. We investigate a lattice rescoring algorithm that integrates information from the image at two different points: the image is used to augment the language model with the most likely words, and to rescore the top hypotheses using a word-level RNN. This rescoring mechanism decreases the word error rate by 3 absolute percentage points, compared to a baseline speech recognizer operating with only the speech recording.

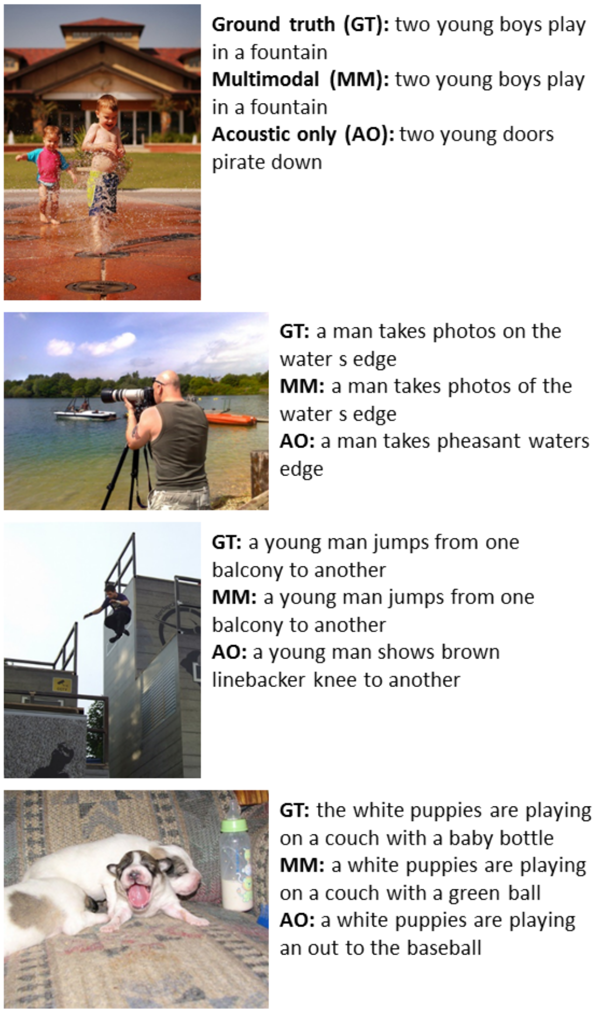
***Index Terms*—** Multimodal speech recognition, image captioning, CNN, lattices

1. INTRODUCTION

In many real-world speech recognition applications, contextual information may be available that can make the recognition problem easier. For example, an automatic captioning service for a TV show can reasonably expect that the speech is related to the images being shown in the show. In this paper, we describe an automatic speech recognition (ASR) system that uses context from an image to inform decoding.

In the computer vision community, much work has been done on generating captions for a provided image, and on the related, and more constrained, problem of scoring the relevancy of a caption for an image. For the caption scoring problem, Yan and Mikolajczyk [1] extracted features from the image using a convolutional neural network (CNN) and features from a bag of words representation of the sentence using another CNN. These features were matched using cross-correlation. Socher et al. [2] used a similar approach, except with a syntax tree-based neural network to process the sentence.

For the caption generation problem, early approaches focused on picking the best caption from a large database, or filling a caption template. Kuznetsova et al. [3] use feature matching to pick words that match a test image, from a training database of images and words. They then stitch the best words together into a sentence. More recent approaches use a recurrent neural network (RNN) to generate words of a sentence in order, from scratch. Vinyals et al. [4] as well as



Ground truth (GT): two young boys play in a fountain
Multimodal (MM): two young boys play in a fountain
Acoustic only (AO): two young doors pirate down

GT: a man takes photos on the water s edge
MM: a man takes photos of the water s edge
AO: a man takes pheasant waters edge

GT: a young man jumps from one balcony to another
MM: a young man jumps from one balcony to another
AO: a young man shows brown linebacker knee to another

GT: the white puppies are playing on a couch with a baby bottle
MM: a white puppies are playing on a couch with a green ball
AO: a white puppies are playing an out to the baseball

Fig. 1. A demonstration of how image context informs speech decoding in our multimodal recognition system. The three sentences next to each image show the ground truth utterance (GT), the decoding with image context ("multimodal", MM), and the decoding without image context ("acoustic only", AO). These examples were manually selected from the development set.

978-1-5090-4903-5/16/$31.00 ©2016 IEEE

Karpathy and Li [5] fed the output of a CNN directly into the first hidden state of a word-generating RNN. The main disdavantage of this approach is that a fixed-size representation of the image is used to generate the caption, regardless of the complexity of the image. Xu et al. [6] attempt to fix this problem using an attention-based decoder, in which each word is generated from a dynamically-chosen sub-sample of the image.

We use these advances in image caption understanding to incorporate context from a single image into a lattice-based speech recognition system. This is done in three steps: First, we extract phrases that are likely to be used to describe the image, and build a language model which puts greater emphasis on these phrases. Then, we combine this language model with an acoustic model, and extract the most likely sentences for the utterance. Finally, we rescore each likely sentence on how well it matches the image. Both the word extraction step and the sentence rescoring step are performed with the same image neural network model.

In this paper, we use the "neuraltalk2" library by Karpathy and Li [5] as our image model. In this image model, each input image is first fed through the VGG-16 CNN [7], an image preprocessing architecture commonly used in computer vision applications. The activations from the final convolutional layer of the CNN are interpreted as a feature vector for the image. This feature vector is used as the initial state of a word-generating LSTM, which takes as input the previous generated word and produces as output a probability distribution over the next word in the sentence. This architecture supports both caption scoring, by feeding the caption into the LSTM and multiplying the probabilities of each word in the caption; and caption generation, by sampling one word at a time from the LSTM.

We train our system on a spoken version of the Flickr8k dataset [8], which contains five written captions describing each of 8000 images, plus a spoken audio version of each caption. We find that our recognizer is significantly more accurate than a recognizer that uses only the spoken captions and transcripts.

2. SYSTEM DESIGN

In any speech recognition system, there are usually two major components: an acoustic model $P(S|W)$ that gives the probability that list of words W sounds like a list of speech frames S; and a language model $P(W)$ that provides a prior distribution over the word sequences in the language. The probability that an utterance S contains a sentence W is calculated using Bayes' rule:

$$P(W|S) = \frac{P(S|W) \cdot P(W)}{P(S)} \propto P(S|W) \cdot P(W) \qquad (1)$$

In our multimodal recognition system, we introduce a new variable: the image I. We assume that I is independent of S given W; in other words, the image only affects the speech through affecting the words in the speech. With this Markovian relationship, we can rewrite the decoding rule as

$$P(W|S,I) = \frac{P(W,S,I)}{P(S,I)} = \frac{P(S|W) \cdot P(W|I) \cdot P(I)}{P(S,I)} \qquad (2)$$

At decoding time, S and I are fixed, so

$$P(W|S,I) \propto P(S|W) \cdot P(W|I) \qquad (3)$$

The first term $P(S|W)$ is the same acoustic model as before, and the second term $P(W|I)$ is an image captioning model. Below, we will focus on the design of this image captioning model.

We define the image captioning model as the weighted combination of two components: a trigram language model P_{lm}, and a RNN caption-scoring model P_{rnn}. The total caption probability is

$$P(W|I) = P_{lm}(W|I)^{\alpha} \cdot P_{rnn}(W|I)^{\beta} \qquad (4)$$

The trigram model is faster but less precise than the RNN model, and is used to prune the decoding lattice in a first pass. This language model approximates the true $P(W|I)$ by sampling many sentences from the caption generation model, and then summarizing the sentences in a trigram model. As such, it is specific to each image. A large number N_c of captions are generated for the image, using the RNN caption generator by Karpathy and Li. These captions are combined with all of the real captions in the training set, and a trigram language model is trained on the entire combined corpus. The generated captions are intended to bias the language model towards words and short phrases that are more likely, given the image. The trigram model is not designed to be precise enough to reliably pick out only the correct sentence; rather, it is designed to preserve in the lattice a number of possible sentences that could be correct, so that the more precise RNN caption-scoring model can then find the best one.

The resulting trigram model can be used in the Kaldi speech recognition toolkit [9], in place of a regular language model. From the resulting lattices, the 100 most likely sentences for each utterance are extracted, and rescored using the full $P(W|S,I)$: a weighted combination of the acoustic model, the image-conditioned trigram model, and the RNN caption scoring model by Karpathy and Li. The most likely sentence at this point is returned as the final answer. The recognition process is summarized in Figure 2.

3. EXPERIMENTS

3.1. Data

We train and evaluate our multimodal recognition system on the spoken Flickr8k dataset. The original Flickr8k dataset [10] consists of 8000 images of people or animals in action

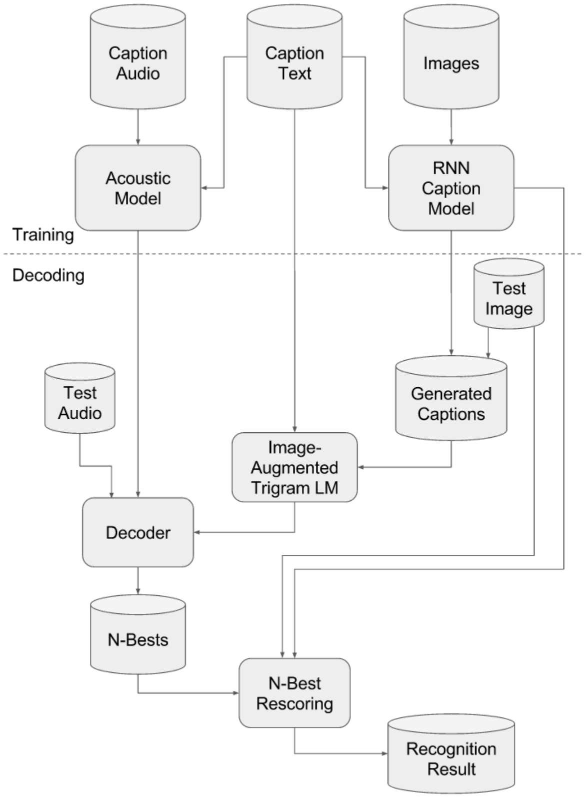


Fig. 2. Configuration of the multimodal recognition system.

from the Flickr photo community website. Each image was described by 5 human volunteers, resulting in 40,000 descriptive sentences.

The spoken Flickr8k dataset, by Harwath and Glass [8], contains spoken recordings of all 40,000 sentences. The audio was collected via Amazon Mechanical Turk, an online marketplace for small human tasks. 183 Turkers participated in this task, recording an average of just over 200 sentences/person. Due to the distributed crowdsourced collection procedure, the quality of the recordings is highly variable. As such, this dataset represents a challenging open-ended speech recognition problem.

The dataset was partitioned into training, development, and test sets using the official published Flickr8k split. 6000 images (with 30,000 sentences in all) were assigned to the training set, and 1000 images (5000 sentences) to each of the development and test sets. Note that there is speaker overlap between the three splits: some utterances in the training and testing sets are spoken by the same speaker.

We additionally use the Flickr30k dataset [11] to provide a larger training corpus for our image captioning model. The Flickr30k dataset consists of 30,000 images with 5 captions each, generated using the same procedure as the Flickr8k dataset. As such, it is a training-only dataset; all 30,000 images are intended to be used for training, and the original Flickr8k development and test sets are to be used for evaluation. It is worth reiterating that only the image captioning model can be trained using Flickr30k; the acoustic models are trained using only Flickr8k, as there is not yet a spoken version of the Flickr30k dataset.

3.2. Baseline Recognizer and Acoustic Model

We first train a Kaldi recognizer on the 30,000 spoken captions from the Flickr8k training set. Our baseline recognizer is trained using the default Kaldi recipe for the “WSJ tri2” configuration, which uses a 13-dimensional MFCC feature representation plus first and second derivatives. Feature transformation and normalization is performed via LDA and MLLT, respectively.

3.3. Building the trigram model

First, the RNN image captioning model was trained on the respective Flickr training sets, using the default parameters in the neuraltalk2 GitHub repository. The neuraltalk2 training process initializes parameters from a pre-trained VGG-16 network [7]. This pre-trained network reduces the training time, and improves performance on the Flickr datasets, which are relatively small by computer vision standards. Afterwards, stochastic gradient descent is performed on the entire model over the training data.

To make the trigram language model, we use the trained neuraltalk caption generator to sample N_c captions for each image in the dev/test set, and append these captions to the existing training set captions to create a training corpus for each image. We then optimize the discounting parameters for a trigram model with Knesser-Ney interpolation (using the kaldi_lm library). The result is a different language model for each image.

There are two parameters to adjust for the trigram model: N_c, the number of generated captions to add to the model, and T, the “temperature” of the RNN caption generator. The output of the RNN caption generator is a score for each possible vocabulary word w at the current sentence position, $s(w, i)$. The distribution over the i-th word in the sentence is defined as

$$P(w_i|w_{1:i-1}) = \frac{\exp(s(w_i, i)/T)}{\sum_w \exp(s(w, i)/T)} \quad (5)$$

Therefore, the sampling process is more likely to pick the highest-scoring words at a low temperature, and more likely to pick random words at a high temperature. The optimal temperature must strike a balance between not generalizing to the testing data at the low end, and not providing any useful information about the image at the high end.

To tune these parameters, and to assess whether the addition of generated captions improves the performance of our language model, we measure the perplexity of our language model on the development set, consisting of 1000 images with

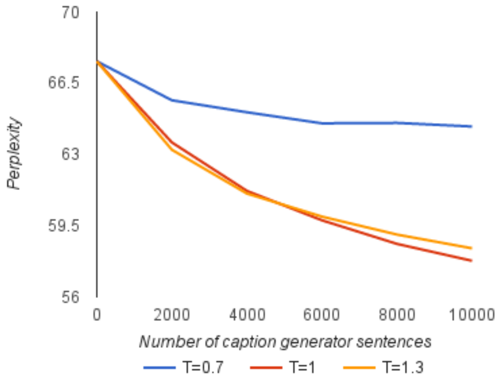


Fig. 3. Perplexity on the development set, for various image-augmented language model parameters. The image captioning model used in this experiment was trained on Flickr8k.

5 captions each. The results are shown in Figure 3. In general, our image-augmented language model is significantly better at modeling the held-out development set than a standard trigram model trained on only the given captions. To verify that the better performance is not just due to more training data, we train a trigram model on the training corpus, plus 10 sentences chosen randomly from each of the 1000 development images. The perplexity of this trigram model was 66.06, which is essentially the same as the trigram model trained on only the training corpus. In our subsequent experiments, we used the best configuration found in cross-validation, which was 10000 captions with a temperature of 1.0.

3.4. Rescoring using the RNN model

Using the trained acoustic model and the image-augmented trigram model, we build decoding lattices using the standard Kaldi procedure. We then generate the 100 most likely sentences according to the lattice, for each test utterance. We rescore each sentence using the acoustic, trigram, and RNN models together. Grid search over the weights of each model was performed on the development set, to find the best linear combination.

4. RESULTS AND ANALYSIS

Table 1 shows the word error rate (WER) of the multimodal recognition system in various configurations on the 5000 utterance development set. The full multimodal system decreases the WER by about 0.8 percentage points when the image captioning model is trained on Flickr8k, and 2.8 percentage points when the image captioning model is trained on Flickr30k. The image-augmented language model became relatively more effective when the training corpus was enlarged. It accounts for barely 0.1 percentage points of improvement when trained with Flickr8k, but a full 1.5 percentage points of improvement when trained with Flickr30k.

System	WER (%)	
Training set	Flickr8k	Flickr30k
Acoustic + LM *(baseline)*	15.27	
Acoustic + LM + RNN	14.53	13.76
Acoustic + Image-LM	15.15	13.74
Acoustic + Image-LM + RNN	14.43	12.51

Table 1. Word error rates on the Flickr8k development set. LM refers to the trigram language model trained on only the training captions; Image-LM refers to the trigram model trained on the training captions plus the captions generated by neuraltalk to describe the image. RNN refers to the caption-scoring neuraltalk model. The columns represent the data used to train the neuraltalk model; the acoustic model is trained on only Flickr8k throughout.

4.1. Oracle experiments

Next, we perform some experiments to explore the performance of each of the components of the model in more detail. The trigram model was designed to ensure that the decoding lattice contains sentences with the correct key words and phrases. To measure the extent to which this is happening, we compute the accuracy of the *best* hypothesis in the top 100 hypotheses in each lattice. This is equivalent to assuming that the RNN model is perfect, and can pick out the most likely sentence, as long as that sentence is one of the choices. (We therefore call this the *oracle RNN* model.) If the image-augmented trigram is working correctly, it should have a higher accuracy, compared to a standard trigram, when used in the oracle RNN model.

In Table 2, we see that the oracle RNN model is modestly more accurate when image-augmented trigrams are used, as opposed to corpus-only trigrams. In particular, the Flickr30k image-augmented trigrams improve the best hypotheses in the lattices by almost a whole percentage point. This provides additional evidence that adding image context at the lattice stage can be complimentary to rescoring top hypotheses using image context. At the same time, using an oracle RNN improves the WER dramatically, compared to a regular RNN. Therefore, the rescoring step is not optimal: the existing combination of acoustic, language, and image context models is not consistently identifying the best hypothesis out of the top 100.

We can also analyze the RNN model in isolation - even if the trigram model can reliably put the correct answer in the lattice, can the RNN model identify the correct answer? To do this, we add the ground truth sentence to the top 100 hypotheses from the lattice, and use the RNN model alone to rescore all 101 sentences. We compute the WER of the sentence that the RNN model marks as the best. We call this

System	WER (%)	
Training set	Flickr8k	Flickr30k
Acoustic + Image-LM + RNN	14.43	12.51
Standard trigram, oracle RNN	8.55	
Image-augmented trigram, oracle RNN	8.44	7.61
Acoustic + Image-LM lattice, rescore with RNN only	17.57	14.05
Oracle lattice, rescore with RNN only	12.71	7.46

Table 2. Word error rates on the Flickr8k development set, for models with oracle components. In the oracle RNN model, we assume that the RNN model assigns a cost of 0 to the most correct sentence, and infinity to every other sentence. In the oracle lattice model, we assume that the decoding lattice always contains the ground truth sentence.

the *oracle lattice* model, because this is equivalent to having a perfect lattice that always contains the correct hypothesis.

We would expect the oracle lattice system with RNN-only rescoring to exhibit worse performance than the actual system, because the final rescoring step does not use any acoustic or language model information. The RNN model alone must pick the correct sentence from a list of 101. However, the results in Table 2 show that, when the correct sentence is a choice, the RNN model is very good at finding it. The acoustic oracle system is in fact more accurate than any of the full-rescoring systems. Even without an oracle lattice, rescoring using the RNN alone in the Flickr30k system is more effective than rescoring using a combination of all the models in the Flickr8k system.

The oracle lattice experiment suggests that the RNN is adept at picking the best sentence from the lattice, but the lattices themselves do not contain the right sentences. However, the oracle RNN experiment shows that the lattices already contain hypotheses that are much better than the ones chosen by the rescoring. Between these two propositions, it may be that the RNN model is better at recognizing the exact sentence to describe an image, than it is at recognizing sentences that are slightly different.

4.2. Test set

Finally, we present results on the Flickr8k test set. In light of the development set experiments showing that the RNN rescoring model provided most of the word error rate improvement, we also built speaker-adapted (SAT) versions of the Acoustic + LM and Acoustic + LM + RNN models, by fitting fMLLR transforms. (In the Kaldi framework, it is difficult to perform speaker adaptation across multiple language models, so we did not apply SAT to the Image-LM models.)

Table 3 shows that adding image context improves the WER, both with and without speaker adaptation. With the speaker-adapted lattices, rescoring using an image captioning model trained on Flickr8k and Flickr30k yield roughly the same results.

System	WER (%)	
Training set	Flickr8k	Flickr30k
Acoustic + LM	14.75	
Acoustic + Image-LM + RNN	13.81	11.95
Acoustic (SAT) + LM	11.64	
Acoustic (SAT) + LM + RNN	11.08	11.05

Table 3. Word error rates on the Flickr8k test set. SAT refers to a speaker-adapted acoustic model.

5. CONCLUSIONS

If an utterance is spoken in the context of some image, we showed that the image can provide information that improves speech recognition. We found two strategies that each decrease the word error rate: rescoring the most likely sentences from a decoding lattice using a caption scoring model, and building a trigram language model that is biased towards phrases that might describe the image. Our trigram approach may be too time-consuming to be practical, but it shows that integrating image information into a lattice can improve recognition on top of simply rescoring the most likely paths in the lattice. Compared to rescoring the top hypotheses, a lattice-based approach can be used to explore a much wider range of possible decodings, because of the inherent efficiency of the lattice representation. More work is needed to determine how to efficiently integrate image context into a decoding lattice.

We found evidence that our RNN caption scoring model was not good at identifying the closest sentence, when none of the most likely sentences were exactly correct. This may be a consequence of the way the caption model is trained: during training, the caption model is used to pick the right caption from a set of captions for randomly-selected images. It is never asked to discriminate among similar captions. If this is the case, better results may be obtainable by training the caption scoring model in a way that is more similar to how we use it.

Our work shows that integrating image cues into speech recognizers is a promising approach, when appropriate visual data are available. Our system was trained on a relatively small amount of data by computer vision standards, so we expect that the recording of a larger multimodal dataset will increase the gap between multimodal and speech-only systems.

6. REFERENCES

[1] Fei Yan and Krystian Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3441–3450.

[2] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[3] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi, "Collective generation of natural image descriptions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 359–368.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition*, 2015.

[5] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv preprint arXiv:1502.03044*, 2015.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] David Harwath and James Glass, "Deep multimodal semantic embeddings for speech and images," in *Proceedings of Interspeech*, 2015.

[9] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[10] Micah Hodosh, Peter Young, and Julia Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, pp. 853–899, 2013.

[11] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.