VISUAL FEATURES FOR CONTEXT-AWARE SPEECH RECOGNITION

Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze

Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA; U.S.A.
abhinavgupta94@gmail.com, {ymiao|lneves|fmetze}@cs.cmu.edu

ABSTRACT

Transcribing consumer generated multi-media content such as "Youtube" videos is still one of the hardest tasks for automatic speech recognition, with error rates well above 25%. Such data typically occupies a very broad domain, has been recorded in challenging conditions, with cheap hardware and a focus on the visual modality, and may have been post-processed or edited.

In this paper, we extend our earlier work on adapting the acoustic model of a DNN-based speech recognition system to an RNN language model, and show how both can be adapted to the objects and scenes that can be automatically detected in the video. We are working on a corpus of "how-to" videos from the web, and the idea is that an object that can be seen ("car"), or a scene that is being detected ("kitchen") can be used to condition both models on the "context" of the recording, thereby reducing perplexity and improving transcription. We achieve good improvements in both cases, and compare and analyze the respective reductions in word error rate.

We expect that our results can be useful for any type of speech processing in which "context" information is available, for example in robotics, man machine interaction, or when indexing large audio-visual archives, and should ultimately help to bring together the "video-to-text" and "speech-to-text" communities.

Index Terms— audio-visual speech recognition, multimodal processing, deep learning

1. INTRODUCTION

For automatic speech recognition (ASR) systems to become universally useful will require robustness to almost all types of input and signal variability. One way in which this could be achieved is to adapt both the acoustic model and the language model to the broad "context" of the input. By "context", we mean essentially anything that is known about the input speech.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

State-of-the-art recognition accuracy on a wide range of acoustic modeling tasks is defined by DNNs [**?**, **?**, **?**], or variants thereof. DNNs display generally better recognition accuracy than traditional Gaussian mixture models (GMMs) [**?**]. Robustness however remains a challenge for DNN models [**?**]. [**?**] for example shows that the performance of simple DNNs degrades significantly as the signal-to-noise ratio (SNR) drops. An effective strategy to deal with variability is to incorporate additional, longer-term knowledge explicitly into DNN models: [**?**, **?**, **?**, **?**, **?**] study the incorporation of speaker-level i-vectors to smooth out the effect of speaker variability. [] uses wide temporal input windows to improve robustness dynamically. Similarly, in [**?**], we learn a DNN-based extractor to model the speaker-microphone distance information dynamically on the frame level. Then distance-aware DNNs are built by appending these descriptors to the acoustic features as the DNN inputs.

It is an important distinction that our work *does not require* localization of lip regions and/ or extraction of frame-synchronous visual features (lip contours, mouth shape, SIFT, landmarks, etc.), as is the case in "traditional" audio-visual ASR [**?**, **?**, **?**, **?**], which has been developed mostly with a focus on noise robustness. For the majority of our data, this information is not available at all, or the quality is poor.

For example, in the automatic lip-reading literature [**?**, **?**], areas of interest (AOI) centering around the lip are extracted to form the image features. In [**?**], coefficients of lip shape and intensity, together with their temporal dynamics, are generated as the visual descriptors. A more straightforward feature type is the gray-scale pixel values of the (downsampled) image covering the speaker's mouth [**?**]. In [**?**], visual features are obtained from pixel color using raster scan, i.e., 30-dim RGB features with 10 dimensions for each color. The second problem lies in the fusion of the audio and visual modalities into the bimodal system. In practice, this fusion can be conducted either on the feature level [**?**, **?**, **?**] where a bimodal front-end is constructed with the two feature streams, or on the decision level [**?**, **?**] where outputs from classifiers are combined during the recognition stage. A major challenge for this fusion is the asynchrony of the audio and video streams. To solve this issue, attempts have been made to combine the classifier outputs (e.g., state likelihoods) at a coarser level, for

example the phoneme or even the word level [**?**, **?**]. Another useful tool to deal with this asynchrony is dynamic Bayesian networks (DBNs) which allow for different levels of information fusion and have shown effectiveness in audio-visual ASR [**?**, **?**]. Despite these advancement, audio-visual ASR still has limitations that prevent its deployment to real-world video data. For example, mouth/ lip related features are not always available in open-domain videos.

In this paper, we aim to relax these constraints and demonstrate the effectiveness of adapting both the acoustic and the language model of a multi-modal ASR system by using "context" information from challenging, open-domain Internet videos. Our approach is based on deep learning, and can be split into two parts:

First, we extract visual features using deep convolutional neural networks (CNNs) trained for object recognition and scene labeling tasks. We extract such information from a random frame within an utterance only, rather than at the level of each frame, but other levels of granularity, or smoothing approaches are also possible. We do thus not require perfect alignment between audio and video channels, which is often almost impossible to achieve on data that has been collected "in the wild". Our "context vector" is thus an n-dimensional representation of the visuals which are present while an utterance is being spoken.

Then, we adapt the acoustic model of the recognizer using a framework in which the residual error at the feature inputs of a DNN is reduced with an adaptation network. This network is trained on the context vector, and predicts a linear shift of the main DNN's input features, as we have presented previously [**?**, **?**, **?**]. For decoding, we use an in-domain 3-gram language model. We re-rank 30-best lists with an RNN language model, which has been conditioned on the same segment-level "context vector" as the acoustic model.

Audio-visual speech recognition, [1, 2, 3, 4].

2. EXTRACTION OF VISUAL FEATURES

The extraction of visual features follows our previous work on adaptation of DNNs using speaker attributes [**?**] and visual features [**?**].

Suppose we are dealing with an utterance u, which has the acoustic features $O = o_1, o_2, ..., o_T$, where T is the total number of speech frames. On a video transcribing task, there always exists a video segment corresponding to u. This video segment is represented as $V = v_1, v_2, ..., v_N$, where N is the number of video frames. The video frames are sampled normally at a lower sampling rate than the speech frames, i.e., $N < T$. From this video segment, we select a video frame v_n which serves as the image representation for the speech utterance. Then two types of image features are extracted from v_n.

2.1. Object Features

Our first type of visual information is derived from object recognition, the task on which deep learning has accomplished tremendous success [**?**]. The intuition is that object features encapsulate information regarding the acoustic environment/condition of speech data. For example, classifying an image to the classes "computer keyboard" and "monitor" in-dicates that the speech segment has been recorded in an office.

We extract this object information using a deep CNN model which has been trained on a comprehensive object recognition dataset, a 1.2 million image subset of ImageNet [**?**] used for the 2012 ILSVR challenge, and the resulting CNN model is referred to as OBJECT-CNN. Then, on our target ASR task, the video frame v_n is fed into the CNN model, from which we get the distribution (posterior probabilities) over the object classes. These probabilities encode the object-related information that are finally incorporated into DNN acoustic models.

The OBJECT-CNN network follows the standard AlexNet architecture [25]. The network contains 5 convolution layers which use the rectifier non-linearity (ReLU) [**?**] as the activation function. In the first and second convolution layers, a local response normalization (LRN) layer is added after the ReLU activation, and a max pooling layer follows the LRN layer. In the third and fourth convolution layers, we do not apply the LRN and pooling layers. In the fifth convolution layer, we only apply the max pooling layer, without LRN applied. 3 fully-connected (FC) layers are placed on top of the convolution layers. The first and second FC layers have 4096 neurons, whereas the number of neurons in the last FC layer is equal to the number of classes, 1000 in our case. Model training optimizes the standard cross-entropy (CE) objective. The resulting OBJECT-CNN achieves a 20% top-5 error rate on the ILSVRC 2012 testing set.

2.2. Place Features

The utility of the object features comes from the "place" information that is implicitly encoded by the object classification results. It is then natural to utilize place features in a more explicit way. To achieve this, we train a deep CNN model meant for the scene labeling task. Given a video frame, the classification outputs from this PLACE-CNN encode the place information, which is then incorporated into acoustic models. For convenience of formulation, the resulting visual feature vector for this utterance u is represented as f_u.

In order to extract place information, we train the place-CNN network on the MIT Places dataset [**?**] which contains 2.5 million images belonging to 205 scene categories. Examples of the scenes include "dining room", "coast", "conference center", "courtyard", etc. We use the complete set of 2.5 million images for training, and follow the same image pre-processing as used on ImageNet (Section 2.1). The ar-

chitecture of the PLACE-CNN is almost the same as that of the OBJECT-CNN. The only difference is that in the final FC layer, the PLACE-CNN has 205 neurons corresponding to the 205 scene classes, whereas the OBJECT-CNN contains 1000 neurons.

3. DATA AND BASELINE DESCRIPTION

We chose to investigate context-aware ASR on a dataset of real-world English instructional videos, which we had downloaded from online video archives [**?**, **?**]. These videos have been uploaded by social media users to share expertise on specific tasks (e.g., oil change, sandwich making, etc.). ASR on these videos is challenging because they have been recorded in various environments (e.g., office, kitchen, baseball field, train, etc.), giving us a variety of contexts, yet they are rich in speech, making them suitable for the proposed work. After data preparation, we get 94 hours of speech, out of which 90 hours were used for training and 4 hours for testing.

We used Kaldi [**?**] and PDNN [**?**] for our experiments, training a 5-layer [**?**] DNN acoustic model using cross-entropy. For decoding, a trigram language model (LM) is trained on the training transcripts. This LM is then interpolated with another trigram LM trained on an additional set of 300 hours transcriptions of instructional videos.

4. ACOUSTIC MODEL ADAPTATION

In previous work [**?**, **?**], we presented a framework to perform speaker adaptive training (SAT) for DNN models. This approach requires an i-vector [**?**] to be extracted for each speaker. Based on the well-trained speaker-independent (SI) DNN, a separate adaptation neural network is learned to convert i-vectors into speaker-specific linear feature shifts. Adding these shifts to the original DNN inputs produces a speaker-normalized feature space. Parameters of the SI-DNN are re-updated in this new space, generating the SAT-DNN model. This framework has also been applied successfully to descriptors of speaker-microphone distance [**?**], and we find it to be more robust than straightforward feature concatenation [**?**].

We port this idea to visual input features, which enables us to conduct "context" adaptation for DNNs, simply replacing the i-vector representation with the visual features. An adaptation network is learned to take the visual features as inputs and generate an adaptive feature space with respect to the visual descriptors. Note that in this case, the linear feature shifts generated by the adaptation network are utterance-specific rather than speaker-specific. Re-updating the parameters of the DNN in the normalized feature space gives us the adaptively trained "video adaptive training" VAT-DNN model [**?**]. This VAT-DNN model takes advantage of the visual features as additional knowledge, and generalizes better to unseen variability. In our setup, the 100-dim utterance-level visual features are taken as the inputs into the adaptation network, which contains 3 hidden layers with 512 neurons per layer. The outputs are 40-dimensional shifts to lMEL features.

Baseline	Object Features	Place Features	Comb. of Visual F. [**?**]	i-vectors
23.4%	22.5%	22.5%	22.3%	22.0%

Table 1: Word error rates when applying acoustic model adaptation. Combination of all visual features with i-vectors reduces the WER to 21.5%.

5. LANGUAGE MODEL ADAPTATION

To adapt the language model, we used the same features that we also use for adapting the acoustic model as "topic" information in a context dependent Recurrent Neural Network (RNN) language model [5]. After initial experiments using 5-fold cross-validation on about 390 hours of data (same as used for the baseline language model), a two-layer LSTM with 512 cells performed best, and was adopted for the experiments below. The vocabulary size is **??**.

We provide the adaptation vector only at the beginning of the sentence, although it might make sense to provide it also at intermediate steps, as the average sentence length is 18 words.

6. EXPERIMENTS

To reduce the dimensionality of the adaptation feature and to facilitate comparison with earlier work on i-vector adaptation, we reduce the dimensionality of the place and scene features to 100 (from 1000 and 205) using PCA, estimated on the training part only of the audio-visual dataset.

Table 1 shows the result of adapting the DNN acoustic model with visual features, and i-vectors for comparison, as well as a combination of visual features and i-vectors. Gains are consistent, and quite complementary when using the concatenation of visual features and i-vectors for adaptation. Also, in all cases, the adaptation network method outperforms simply concatenating the adaptation vector to the input features.

Next, we use the same method to adapt the language model to the visual information. To find the best meta parameters for the LSTM language model, we performed 5-fold cross-validation on the entire 390 hours of training data, and averaged the results. Figure 1 shows that conditioning the LSTM LM on video features reduces perplexity from 89 to 74 on training data, which is a significant reduction.

We generated 30-best lists using the baseline acoustic models (with a WER of 23.4%), which had an oracle WER

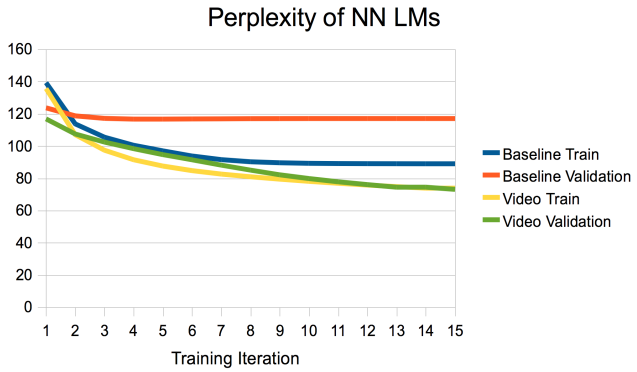

Fig. 1: Training and validation perplexity of the NN LMs trained with and without visual features, averaged across 5 folds, on 390 hours of data. Parameters were optimized for the adaptive LMs, hence the baseline LM converges quickly. Validation perplexity is lower than training perplexity initially because it is measured at the end of each training iteration, while training perplexity is being computed while processing the data and updating the model.

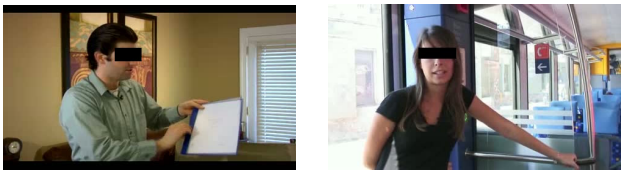

Fig. 2: Keyframes for two typical videos in our dataset. The baseline WER is 27.6% for the "home" video, and 47.7% for the "train" video. Acoustic model adaptation does not improve the "home" video, but reduces WER to 38.2% on the "train" video. Language model adaptation improves WER on both videos slightly, to 26.6% and 43.2% respectively.

of 15.6%, and re-scored them with all 5 neural network language models (NN LMs), averaging the results. Using the concatenation of object and place features as inputs to the NN LM, we achieve a word error rate of 22.6%, which is very close to the performance achieved with the adaptation of the acoustic model.

7. ANALYSIS OF RESULTS

For both acoustic and language model adaptation, we performed some more in-depth analysis to see where gains are mostly coming from.

We manually inspected those videos on which we observed more than 10% relative WER reductions, and find that they have been recorded either in outdoor environments (e.g.b̃aseball field, airport apron, street, etc.) , or in non-typical indoor conditions (e.g.k̃itchen, music studio, etc.), where music/ noise may interfere with the actual speech a lot. Adding the scene descriptors helps the DNN model normalize the acoustic characteristics of these rare conditions, and thus benefits the generalization to unseen testing speech. We then labeled all 156 testing videos as either "typical indoors" (e.g.õffice) and "other" (noisy indoors, outdoors), and analyzed the relative improvements with the of a system adapted with PLACE-CNN features only, and find that the "quiet indoors" videos get improved from 22.1% WER to 21.7%, while "other" videos get improved from 27.6% to 25.7%. "Other" videos thus get improved by 7% relative, while clean videos get improved by 2% only.

When training the NN LM on 90 hours of data only, adaptation with Object-CNN features results in a perplexity of 94.7, while adaptation with Place-CNN features gives a perplexity of 98.9. It seems intuitive that "objects" would be slightly more salient for the topic of a "how-to" video than the scene.

Figure 2 shows typical keyframes from our database, and the typical pattern of improvements: acoustic model adaptation tends to give significant improvements on "outdoor" videos only, while language model adaptation tends to give smaller improvements across the board.

8. CONCLUSION AND FUTURE WORK

In this paper, we described a system that extracts context information that is relevant for speech processing from the visual channel of the video. We showed that the information can be incorporated in both acoustic and language models, and leads to systematic and consistent improvements.

We are currently expanding the acoustic model adaptation experiments to the larger (360 hours) version of the corpus, and expect to see further performance improvements. We are also experimenting with different and better ways of incorporating the video features into the language model, and attempt more insightful analysis of the results, e.g. how much do different types of features contribute to the different models (e.g., do the scene features contribute relatively more to the acoustics, while the object features contribute more to the language model?), and what types of errors are being reduced (e.g., nouns? verbs? semantically confusing errors?).

In the long term, we plan to merge this work with fully end-to-end "video-to-text" approaches, which generate video "summaries" based on multi-modal embeddings, and reference "captions".

9. REFERENCES

[1] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach," in *Proc. INTERSPEECH*, San Francisco, CA; U.S.A., Sept. 2016, ISCA.

[2] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ; U.S.A., Dec. 2015, IEEE, https://github.com/srvk/eesen.

[3] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 1938–1949, Nov. 2015.

[4] Y. Miao and F. Metze, "Distance-aware DNNs for robust speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, Sept. 2015, ISCA.

[5] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model.," .