

DISCRIMINATIVE ACOUSTIC WORD EMBEDDINGS: RECURRENT NEURAL NETWORK-BASED APPROACHES

Shane Settle, Karen Livescu

Toyota Technological Institute at Chicago

{settle.shane, klivescu}@ttic.edu

ABSTRACT

Acoustic word embeddings — fixed-dimensional vector representations of variable-length spoken word segments — have begun to be considered for tasks such as speech recognition and query-by-example search. Such embeddings can be learned discriminatively so that they are similar for speech segments corresponding to the same word, while being dissimilar for segments corresponding to different words. Recent work has found that acoustic word embeddings can outperform dynamic time warping on query-by-example search and related word discrimination tasks. However, the space of embedding models and training approaches is still relatively unexplored. In this paper we present new discriminative embedding models based on recurrent neural networks (RNNs). We consider training losses that have been successful in prior work, in particular a cross entropy loss for word classification and a contrastive loss that explicitly aims to separate same-word and different-word pairs in a “Siamese network” training setting. We find that both classifier-based and Siamese RNN embeddings improve over previously reported results on a word discrimination task, with Siamese RNNs outperforming classification models. In addition, we present analyses of the learned embeddings and the effects of variables such as dimensionality and network structure.

Index Terms— acoustic word embeddings, recurrent neural networks, Siamese networks

1. INTRODUCTION

Many speech processing tasks — such as automatic speech recognition or spoken term detection — hinge on associating segments of speech signals with word labels. In most systems developed for such tasks, words are broken down into sub-word units such as phones, and models are built for the individual units. An alternative, which has been considered by some researchers, is to consider each entire word segment as a single unit, without assigning parts of it to sub-word units. One motivation for the use of whole-word approaches is that

they avoid the need for sub-word models. This is helpful since, despite decades of work on sub-word modeling [1, 2], it still poses significant challenges. For example, speech processing systems are still hampered by differences in conversational pronunciations [3]. A second motivation is that considering whole words at once allows us to consider a more flexible set of features and reason over longer time spans.

Whole-word approaches typically involve, at some level, template matching. For example, in template-based speech recognition [4, 5], word scores are computed from dynamic time warping (DTW) distances between an observed segment and training segments of the hypothesized word. In query-by-example search, putative matches are typically found by measuring the DTW distance between the query and segments of the search database [6, 7, 8, 9]. In other words, whole-word approaches often boil down to making decisions about whether two segments are examples of the same word or not.

An alternative to DTW that has begun to be explored is the use of acoustic word embeddings (AWEs), or vector representations of spoken word segments. AWEs are representations that can be learned from data, ideally such that the embeddings of two segments corresponding to the same word are close, while embeddings of segments corresponding to different words are far apart. Once word segments are represented via fixed-dimensional embeddings, computing distances is as simple as measuring a cosine or Euclidean distance between two vectors.

There has been some, thus far limited, work on acoustic word embeddings, focused on a number of embedding models, training approaches, and tasks [10, 11, 12, 13, 14, 15, 16, 17]. In this paper we explore new embedding models based on recurrent neural networks (RNNs), applied to a word discrimination task related to query-by-example search. RNNs are a natural model class for acoustic word embeddings, since they can handle arbitrary-length sequences. We compare several types of RNN-based embeddings and analyze their properties. Compared to prior embeddings tested on the same task, our best models achieve sizable improvements in average precision.

This research was supported by a Google faculty research award and NSF grant IIS-1321015. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

2. RELATED WORK

We next briefly describe the most closely related prior work.

Maas *et al.* [10] and Bengio and Heigold [11] used acoustic word embeddings, based on convolutional neural networks (CNNs), to generate scores for word segments in automatic speech recognition. Maas *et al.* trained CNNs to predict (continuous-valued) embeddings of the word labels, and used the resulting embeddings to define feature functions in a segmental conditional random field [18] rescoring system. Bengio and Heigold also developed CNN-based embeddings for lattice rescoring, but with a contrastive loss to separate embeddings of a given word from embeddings of other words.

Levin *et al.* [12] developed unsupervised embeddings based on representing each word as a vector of DTW distances to a collection of reference word segments. This representation was subsequently used in several applications: a segmental approach for query-by-example search [13], lexical clustering [19], and unsupervised speech recognition [20]. Voinea *et al.* [16] developed a representation also based on templates, in their case phone templates, designed to be invariant to specific transformations, and showed their robustness on digit classification.

Kamper *et al.* [14] compared several types of acoustic word embeddings for a word discrimination task related to query-by-example search, finding that embeddings based on convolutional neural networks (CNNs) trained with a contrastive loss outperformed the reference vector approach of Levin *et al.* [12] as well as several other CNN and DNN embeddings and DTW using several feature types. There have now been a number of approaches compared on this same task and data [12, 21, 22, 23]. For a direct comparison with this prior work, in this paper we use the same task and some of the same training losses as Kamper *et al.*, but develop new embedding models based on RNNs.

The only prior work of which we are aware using RNNs for acoustic word embeddings is that of Chen *et al.* [17] and Chung *et al.* [15]. Chen *et al.* learned a long short-term memory (LSTM) RNN for word classification and used the resulting hidden state vectors as a word embedding in a query-by-example task. The setting was quite specific, however, with a small number of queries and speaker-dependent training. Chung *et al.* [15] worked in an unsupervised setting and trained single-layer RNN autoencoders to produce embeddings for a word discrimination task. In this paper we focus on the supervised setting, and compare a variety of RNN-based structures trained with different losses.

3. APPROACH

An acoustic word embedding is a function that takes as input a speech segment corresponding to a word, $X = \{x_t\}_{t=1}^T$, where each x_t is a vector of frame-level acoustic features, and outputs a fixed-dimensional vector representing the segment, $g(X)$. The basic embedding model structure we use is

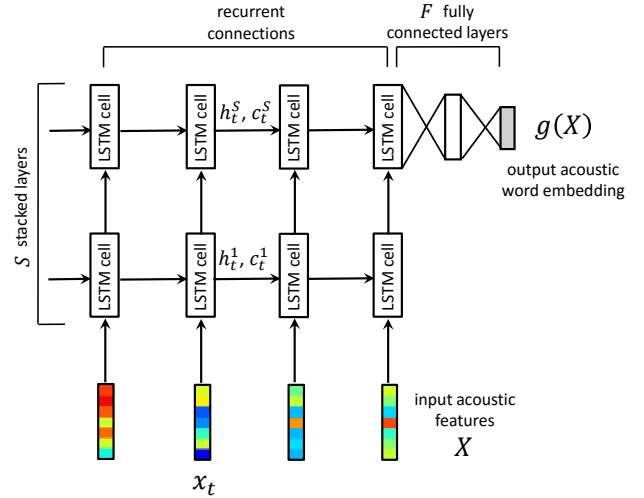


Fig. 1: LSTM-based acoustic word embedding model. For GRU-based models, the structure is the same, but the LSTM cells are replaced with GRU cells, and there is no cell activation vector; the recurrent connections only carry the hidden state vector h_t^l .

shown in Fig. 1. The model consists of a deep RNN with some number S of stacked layers, whose final hidden state vector is passed as input to a set of F of fully connected layers; the output of the final fully connected layer is the embedding $g(X)$.

The RNN hidden state at each time frame can be viewed as a representation of the input seen thus far, and its value in the last time frame T could itself serve as the final word embedding. The fully connected layers are added to account for the fact that some additional transformation may improve the representation. For example, the hidden state may need to be larger than the desired word embedding dimension, in order to be able to “remember” all of the needed intermediate information. Some of that information may not be needed in the final embedding. In addition, the information maintained in the hidden state may not necessarily be discriminative; some additional linear or non-linear transformation may help to learn a discriminative embedding.

Within this class of embedding models, we focus on Long Short-Term Memory (LSTM) networks [24] and Gated Recurrent Unit (GRU) networks [25]. These are both types of RNNs that include a mechanism for selectively retaining or discarding information at each time frame when updating the hidden state, in order to better utilize long-term context. Both of these RNN variants have been used successfully in speech recognition [26, 27, 28, 29].

In an LSTM RNN, at each time frame both the hidden state h_t and an associated “cell memory” vector c_t , are updated and passed on to the next time frame. In other words, each forward edge in Figure 1 can be viewed as carrying both the cell memory and hidden state vectors. The updates are modulated by the values of several gating vectors, which control the degree to which the cell memory and hidden state are

updated in light of new information in the current frame. For a single-layer LSTM network, the updates are as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_i[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_i) && \text{input gate} \\
\mathbf{f}_t &= \sigma(\mathbf{W}_f[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_f) && \text{forget gate} \\
\tilde{\mathbf{c}}_t &= \tanh(\mathbf{W}_c[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_c) && \text{candidate cell memory} \\
\mathbf{c}_t &= \mathbf{i}_t \odot \tilde{\mathbf{c}}_t + \mathbf{f}_t \odot \mathbf{c}_{t-1} && \text{cell memory} \\
\mathbf{o}_t &= \sigma(\mathbf{W}_o[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_o) && \text{output gate} \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t) && \text{hidden state}
\end{aligned}$$

where \mathbf{h}_t , \mathbf{c}_t , $\tilde{\mathbf{c}}_t$, \mathbf{i}_t , \mathbf{f}_t , and \mathbf{o}_t are all vectors of the same dimensionality, \mathbf{W}_i , \mathbf{W}_o , \mathbf{W}_f , and \mathbf{W}_c are learned weight matrices of the appropriate sizes, \mathbf{b}_i , \mathbf{b}_o , \mathbf{b}_f and \mathbf{b}_c are learned bias vectors, $\sigma(\cdot)$ is a componentwise logistic activation, and \odot refers to the Hadamard (componentwise) product.

Similarly, in a GRU network, at each time step a GRU cell determines what components of old information are retained, overwritten, or modified in light of the next step in the input sequence. The output from a GRU cell is only the hidden state vector. A GRU cell uses a reset gate \mathbf{r}_t and an update gate \mathbf{u}_t as described below for a single-layer network:

$$\begin{aligned}
\mathbf{r}_t &= \sigma(\mathbf{W}_r[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_r) && \text{reset gate} \\
\mathbf{u}_t &= \sigma(\mathbf{W}_u[\mathbf{x}_t, \mathbf{h}_{t-1}] + \mathbf{b}_u) && \text{update gate} \\
\tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h[\mathbf{x}_t, \mathbf{r}_t \odot \mathbf{h}_{t-1}] + \mathbf{b}_h) && \text{candidate hidden} \\
\mathbf{h}_t &= \mathbf{u}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{u}_t) \odot \tilde{\mathbf{h}}_t && \text{hidden state}
\end{aligned}$$

where \mathbf{r}_t , \mathbf{u}_t , $\tilde{\mathbf{h}}_t$, and \mathbf{h}_t are all the same dimensionality, \mathbf{W}_r , \mathbf{W}_u , and \mathbf{W}_h are learned weight matrices of the appropriate size, and \mathbf{b}_r , \mathbf{b}_u and \mathbf{b}_h are learned bias vectors.

All of the above equations refer to single-layer networks. In a deep network, with multiple stacked layers, the same update equations are used in each layer, with the state, cell, and gate vectors replaced by layer-specific vectors \mathbf{h}_t^l , \mathbf{c}_t^l , and so on for layer l . For all but the first layer, the input \mathbf{x}_t is replaced by the hidden state vector from the previous layer \mathbf{h}_t^{l-1} .

For the fully connected layers, we use rectified linear unit (ReLU) [30] activation, except for the final layer which depends on the form of supervision and loss used in training.

3.1. Training

We train the RNN-based embedding models using a set of pre-segmented spoken words. We use two main training approaches, inspired by prior work but with some differences in the details. As in [14, 11], our first approach is to use the word labels of the training segments and train the networks to classify the word. In this case, the final layer of $g(X)$ is a log-softmax layer. Here we are limited to the subset of the training set that has a sufficient number of segments per word to train a good classifier, and the output dimensionality is equal to the number of words (but see [14] for a study of varying the dimensionality in such a classifier-based embedding model by introducing a bottleneck layer). This model

is trained end-to-end and is optimized with a cross entropy loss. Although labeled data is necessarily limited, the hope is that the learned models will be useful even when applied to spoken examples of words not previously seen in the training data. For words not seen in training, the embeddings should correspond to some measure of similarity of the word to the training words, measured via the posterior probabilities of the previously seen words. In the experiments below, we examine this assumption by analyzing performance on words that appear in the training data compared to those that do not.

The second training approach, based on earlier work of Kamper *et al.* [14], is to train “Siamese” networks [31]. In this approach, full supervision is not needed; rather, we use weak supervision in the form of pairs of segments labeled as same or different. The base model remains the same as before—an RNN followed by a set of fully connected layers—but the final layer is no longer a softmax but rather a linear activation layer of arbitrary size. In order to learn the parameters, we simultaneously feed three word segments through three copies of our model (i.e. three networks with shared weights). One input segment is an “anchor”, x_a , the second is another segment with the same word label, x_s , and the third is a segment corresponding to a different word label, x_d . Then, the network is trained using a “cos-hinge” loss:

$$l_{\text{cos hinge}} = \max\{0, m + d_{\text{cos}}(x_a, x_s) - d_{\text{cos}}(x_a, x_d)\}$$

where $d_{\text{cos}}(x_1, x_2) = 1 - \cos(x_1, x_2)$ is the cosine distance between x_1, x_2 . Unlike cross entropy training, here we directly aim to optimize relative (cosine) distance between same and different word pairs. For tasks such as query-by-example search, this training loss better respects our end objective, and can use more data since neither fully labeled data nor any minimum number of examples of each word should be needed.

4. EXPERIMENTS

Our end goal is to improve performance on downstream tasks requiring accurate word discrimination. In this paper we use an intermediate task that more directly tests whether same- and different-word pairs have the expected relationship. And that allows us to compare to a variety of prior work. Specifically, we use the word discrimination task of Carlin *et al.* [21], which is similar to a query-by-example task where the word segmentations are known. The evaluation consists of determining, for each pair of evaluation segments, whether they are examples of the same or different words, and measuring performance via the average precision (AP). We do this by measuring the cosine similarity between their acoustic word embeddings and declaring them to be the same if the distance is below a threshold. By sweeping the threshold, we obtain a precision-recall curve from which we compute the AP.

The data used for this task is drawn from the Switchboard conversational English corpus [32]. The word segments range from 50 to 200 frames in length. The acoustic features in each

frame (the input to the word embedding models x_t) are 39-dimensional MFCCs+ Δ + $\Delta\Delta$. We use the same train, development, and test partitions as in prior work [14, 12], and the same acoustic features as in [14], for as direct a comparison as possible. The train set contains approximately 10k example segments, while dev and test each contain approximately 11k segments (corresponding to about 60M pairs for computing the dev/test AP). As in [14], when training the classification-based embeddings, we use a subset of the training set containing all word types with a minimum of 3 occurrences, reducing the training set size to approximately 9k segments.¹

When training the Siamese networks, the training data consists of all of the same-word pairs in the full training set (approximately 100k pairs). For each such training pair, we randomly sample a third example belonging to a different word type, as required for the $l_{\cos \text{ hinge}}$ loss.

4.1. Classification network details

Our classifier-based embeddings use LSTM or GRU networks with 2–4 stacked layers and 1–3 fully connected layers. The final embedding dimensionality is equal to the number of unique word labels in the training set, which is 1061. The recurrent hidden state dimensionality is fixed at 512 and dropout [33] between stacked recurrent layers is used with probability $p = 0.3$. The fully connected hidden layer dimensionality is fixed at 1024. Rectified linear unit (ReLU) non-linearities and dropout with $p = 0.5$ are used between fully-connected layers. However, between the final recurrent hidden state output and the first fully-connected layer no non-linearity or dropout is applied. These settings were determined through experiments on the development set.

The classifier network is trained with a cross entropy loss and optimized using stochastic gradient descent (SGD) with Nesterov momentum [34]. The learning rate is initialized at 0.1 and is reduced by a factor of 10 according to the following heuristic: If 99% of the current epoch’s average batch loss is greater than the running average of batch losses over the last 3 epochs, this is considered a plateau; if there are 3 consecutive plateau epochs, then the learning rate is reduced. Training stops when reducing the learning rate no longer improves dev set AP. Then, the model from the epoch corresponding to the the best dev set AP is chosen. Several other optimizers—Adagrad [35], Adadelata [36], and Adam [37]—were explored in initial experiments on the dev set, but all reported results were obtained using SGD with Nesterov momentum.

4.2. Siamese network details

For experiments with Siamese networks, we initialize (warm-start) the networks with the tuned classification network, removing the final log-softmax layer and replacing it with a linear layer of size equal to the desired embedding dimensionality. We explored embeddings with dimensionalities between

8 and 2048. We use a margin of 0.4 in the cos-hinge loss.

In training the Siamese networks, each training mini-batch consists of $2B$ triplets. B triplets are of the form (x_a, x_s, x_d) where x_a and x_s are examples of the same class (a pair from the 100k same-word pair set) and x_d is a randomly sampled example from a different class. Then, for each of these B triplets (x_a, x_s, x_d) , an additional triplet (x_s, x_a, x_d) is added to the mini-batch to allow all segments to serve as anchors. This is a slight departure from earlier work [14], which we found to improve stability in training and performance on the development set.

In preliminary experiments, we compared two methods for choosing the negative examples x_d during training, a uniform sampling approach and a non-uniform one. In the case of uniform sampling, we sample x_d uniformly at random from the full set of training examples with labels different from x_a . This sampling method requires only word-pair supervision. In the case of non-uniform sampling, x_d is sampled in two steps. First, we construct a distribution $P_{y|label(x_a)}$ over word labels y and sample a different label from it. Second, we sample an example uniformly from within the subset with the chosen label. The goal of this method is to speed up training by targeting pairs that violate the margin constraint. To construct the multinomial PMF $P_{y|label(x_a)}$, we maintain an $n \times n$ matrix \mathbf{S} , where n is the number of unique word labels in training. Each word label corresponds to an integer $i \in [1, n]$ and therefore a row in \mathbf{S} . The values in a row of \mathbf{S} are considered similarity scores, and we can retrieve the desired PMF for each row by normalizing by its sum.

At the start of each epoch, we initialize \mathbf{S} with 0’s along the diagonal and 1’s elsewhere (which reduces to uniform sampling). For each training pair $(d_{\cos}(x_a, x_s), d_{\cos}(x_a, x_d))$, we update \mathbf{S} for both $(i, j) = (label(x_a), label(x_d))$ and $(i, j) = (label(x_d), label(x_a))$:

$$s_{i,j} += \begin{cases} \cos(x_a, x_d) & d_{\cos}(x_a, x_d) \leq d_{\cos}(x_a, x_s) + m^* \\ 0 & \text{otherwise} \end{cases}$$

The PMFs $P_{y|label(x_a)}$ are updated after the forward pass of an entire mini-batch. The constant m^* enforces a potentially stronger constraint than is used in the $l_{\cos \text{ hinge}}$ loss, in order to promote diverse sampling. In all experiments, we set $m^* = 0.6$. This is a heuristic approach, and it would be interesting to consider various alternatives. Preliminary experiments showed that the non-uniform sampling method outperformed uniform sampling, and in the following we report results with non-uniform sampling.

We optimize the Siamese network model using SGD with Nesterov momentum for 15 epochs. The learning rate is initialized to 0.001 and dropped every 3 epochs until no improvement is seen on the dev set. The final model is taken from the epoch with the highest dev set AP. All models were implemented in Torch [38] and used the rnn library of [39].

¹We thank Herman Kamper for assistance with the data and evaluation.

Table 1: Final test set results in terms of average precision (AP). Dimensionalities marked with * refer to dimensionality per frame for DTW-based approaches. For CNN and LSTM models, results are given as means over several training runs (5 and 10, respectively) along with their standard deviations.

Model	Dim	AP
MFCCs + DTW [14]	39*	0.214
Corr. autoencoder + DTW [22]	100*	0.469
Classifier CNN [14]	1061	0.532 ± 0.014
Siamese CNN [14]	1024	0.549 ± 0.011
Classifier LSTM	1061	0.616 ± 0.009
Siamese LSTM	1024	0.671 ± 0.011

5. RESULTS

Based on development set results, our final embedding models are LSTM networks with 3 stacked layers and 3 fully connected layers, with output dimensionality of 1024 in the case of Siamese networks. Final test set results are given in Table 1. We include a comparison with the best prior results on this task from [14], as well as the result of using standard DTW on the input MFCCs (reproduced from [14]) and the best prior result using DTW, obtained with frame features learned with correlated autoencoders [22]. Both classifier and Siamese LSTM embedding models outperform all prior results on this task of which we are aware.²

We next analyze the effects of model design choices, as well as the learned embeddings themselves.

5.1. Effect of model structure

Table 2 shows the effect on development set performance of the number of stacked layers S , the number of fully connected layers F , and LSTM vs. GRU cells, for classifier-based embeddings. The best performance in this experiment is achieved by the LSTM network with $S = F = 3$. However, performance still seems to be improving with additional layers, suggesting that we may be able to further improve performance by adding even more layers of either type. However, we fixed the model to $S = F = 3$ in order to allow for more experimentation and analysis within a reasonable time.

Table 2: Average precision on the dev set, using classifier-based embeddings. $S = \#$ stacked layers, $F = \#$ fully connected layers.

S	F	GRU AP	LSTM AP
2	1	0.213	0.240
3	1	0.252	0.244
4	1	0.303	0.267
3	2	0.412	0.418
3	3	0.445	0.519

²Yuan *et al.* [40] have recently been able to improve AP on this test set even further with CNN embeddings, by using a large set of additional (cross-lingual) training data. We do not consider these results to be comparable because of their reliance on additional data.

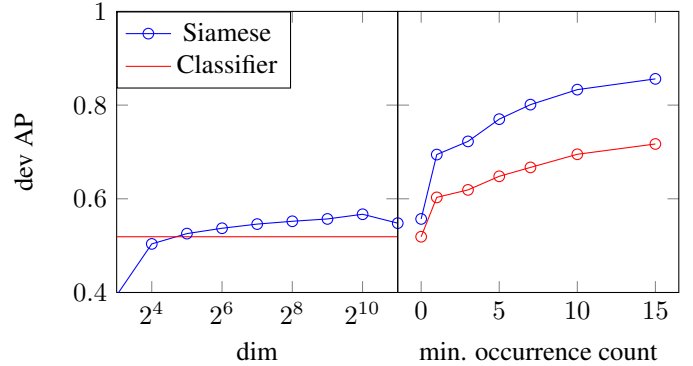


Fig. 2: Effect of embedding dimensionality (left) and occurrences in training set (right).

Table 2 reveals an interesting trend. When only one fully connected layer is used, the GRU networks outperform the LSTMs given a sufficient number of stacked layers. On the other hand, once we add more fully connected layers, the LSTMs outperform the GRUs. In the first few lines of Table 2, we use 2, 3, and 4 layer stacks of LSTMs and GRUs while holding fixed the number of fully-connected layers at $F = 1$. There is clear utility in stacking additional layers; however, even with 4 stacked layers the RNNs still underperform the CNN-based embeddings of [14] until we begin adding fully connected layers.

After exploring a variety of stacked RNNs, we fixed the stack to 3 layers and varied the number of fully connected layers. The value of each additional fully connected layer is clearly greater than that of adding stacked layers. All networks trained with 2 or 3 fully connected layers obtain more than 0.4 AP on the development set, while stacked RNNs with 1 fully connected layer are at around 0.3 AP or less. This may raise the question of whether some simple fully connected model may be all that is needed; however, previous work has shown that this approach is not competitive [14], and convolutional or recurrent layers are needed to summarize arbitrary-length segments into a fixed-dimensional representation.

5.2. Effect of embedding dimensionality

For the Siamese networks, we varied the output embedding dimensionality, as shown in Fig. 2. This analysis shows that the embeddings learned by the Siamese RNN network are quite robust to reduced dimensionality, outperforming the classifier model for all dimensionalities 32 or higher and outperforming previously reported dev set performance with CNN-based embeddings [14] for all dimensionalities ≥ 16 .

5.3. Effect of training vocabulary

We might expect the learned embeddings to be more accurate for words that are seen in training than for ones that are not. Fig. 2 measures this effect by showing performance as a function of the number of occurrences of the dev words in the

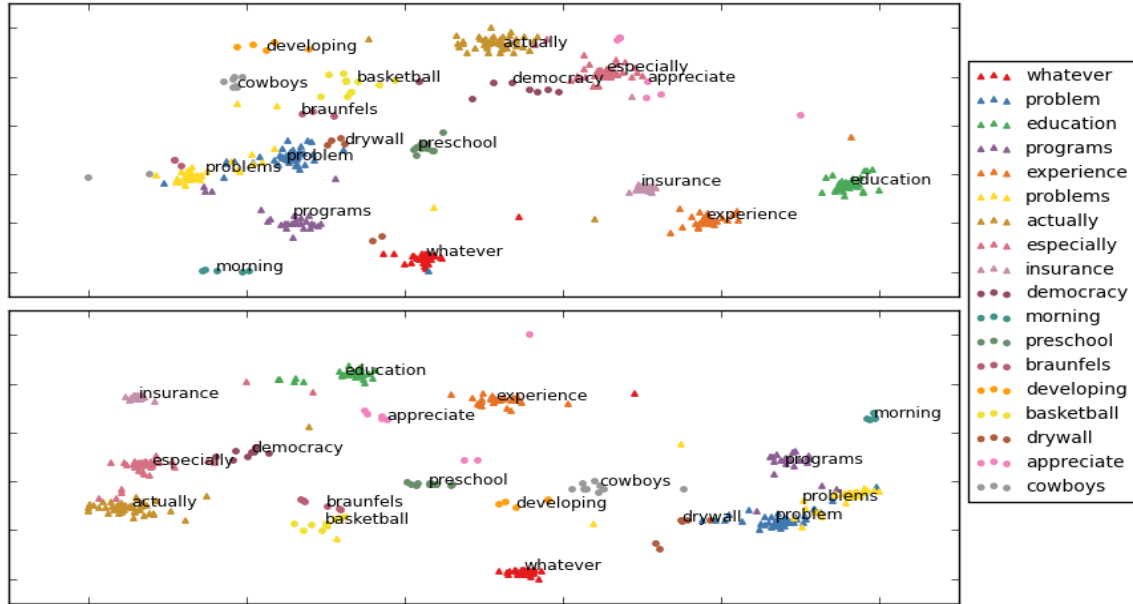


Fig. 3: t-SNE visualization of word embeddings from the dev set produced by the classifier (top) vs. Siamese (bottom) models. Word labels seen at training time are denoted by triangles and word labels unseen at training time are denoted by circles.

training set. Indeed, both model types are much more successful for in-vocabulary words, and their performance improves the higher the training frequency of the words. However, performance increases more quickly for the Siamese network than for the classifier as training frequency increases. This may be due to the fact that, if a word type occurs at least k times in the classifier training set, then it occurs at least $2 \times \binom{k}{2}$ times in the Siamese paired training data.

5.4. Visualization of embeddings

In order to gain a better qualitative understanding of the differences between classifier and Siamese-based embeddings, and of the learned embedding space more generally, we plot a two-dimensional visualization of some of our learned embeddings via t-SNE [41] in Fig. 3. For both classifier and Siamese embeddings, there is a marked difference in the quality of clusters formed by embeddings of words that were previously seen vs. previously unseen in training. However, the Siamese network embeddings appear to have better relative distances between word clusters with similar and dissimilar pronunciations. For example, the word `programs` appears equidistant from `problems` and `problem` in the classifier-based embedding space, but in the Siamese embedding space `problems` falls between `problem` and `programs`. Similarly, the cluster for `democracy` shifts with respect to `actually` and `especially` to better respect differences in pronunciation. More study of learned embeddings, using more data and word types, is needed to confirm such patterns in general. Improvements in unseen word embeddings from the classifier embedding space to the Siamese embedding space (such as for `democracy`, `morning`, and

`basketball`) are a likely result of optimizing the model for relative distances between words.

6. CONCLUSION

Our main finding is that RNN-based acoustic word embeddings outperform prior approaches, as measured via a word discrimination task related to query-by-example search. Our best results are obtained with deep LSTM RNNs with a combination of several stacked layers and several fully connected layers, optimized with a contrastive Siamese loss. Siamese networks have the benefit that, for any given training data set, they are effectively trained on a much larger set, in the sense that they measure a loss and gradient for every possible pair of data points. Our experiments suggest that the models could still be improved with additional layers. In addition, we have found that, for the purposes of acoustic word embeddings, fully connected layers are very important and have a more significant effect per layer than stacked layers, particularly when trained with the cross entropy loss function.

These experiments represent an initial exploration of sequential neural models for acoustic word embeddings. There are a number of directions for further work. For example, while our analyses suggest that Siamese networks are better than classifier-based models at embedding previously unseen words, our best embeddings are still much poorer for unseen words. Improvements in this direction may come from larger training sets, or may require new models that better model the shared structure between words. Other directions for future work include additional forms of supervision and training, as well as application to downstream tasks.

7. REFERENCES

- [1] ISCA, *Proceedings of the International Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, Estes Park, Colorado, 2002.
- [2] M. Ostendorf, “Moving Beyond the ‘Beads-on-a-String’ Model of Speech,” in *IEEE Automatic Speech Recognition & Understanding (ASRU)*, 1999.
- [3] Karen Livescu, Eric Fosler-Lussier, and Florian Metze, “Subword modeling for automatic speech recognition: Past, present, and emerging approaches,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, 2012.
- [4] Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq, Ronald Cools, and Dirk Van Compernelle, “Template-based continuous speech recognition,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1377–1390, 2007.
- [5] Georg Heigold, Patrick Nguyen, Mitchel Weintraub, and Vincent Vanhoucke, “Investigations on exemplar-based features for speech recognition towards thousands of hours of unsupervised, noisy data,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4437–4440.
- [6] Florian Metze, Xavier Anguera, Etienne Barnard, Marie Davel, and Guillaume Gravier, “The spoken web search task at MediaEval 2012,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.
- [7] Xavier Anguera, “Speaker independent discriminant feature extraction for acoustic pattern-matching,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.
- [8] Yaodong Zhang, Kiarash Adl, and James Glass, “Fast spoken query detection using lower-bound dynamic time warping on graphical processing units,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5173–5176.
- [9] Igor Szöke, Miroslav Skácel, Lukáš Burget, and Jan “Honza” Černocký, “Coping with channel mismatch in query-by-example - BUT QUESST 2014,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [10] Andrew L Maas, Stephen D Miller, Tyler M O’neil, Andrew Y Ng, and Patrick Nguyen, “Word-level acoustic modeling with convolutional vector regression,” in *International Conference on Machine Learning (ICML), Representation Learning Workshop*, 2012.
- [11] Samy Bengio and Georg Heigold, “Word embeddings for speech recognition,” in *Interspeech*, 2014.
- [12] Keith Levin, Katharine Henry, Aren Jansen, and Karen Livescu, “Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings,” in *IEEE Automatic Speech Recognition & Understanding (ASRU)*, 2013.
- [13] Keith Levin, Aren Jansen, and Benjamin Van Durme, “Segmental acoustic indexing for zero resource keyword search,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [14] Herman Kamper, Weiran Wang, and Karen Livescu, “Deep convolutional acoustic word embeddings using word-pair side information,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4950–4954.
- [15] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, and Hung-Yi Lee, “Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks,” *Interspeech*, 2016.
- [16] Stephen Voinea, Chiyuan Zhang, Georgios Evangelopoulos, Lorenzo Rosasco, and Tomaso Poggio, “Word-level invariant representations from acoustic waveforms,” in *Interspeech*, 2014, pp. 2385–2389.
- [17] Guoguo Chen, Carolina Parada, and Tara N Sainath, “Query-by-example keyword spotting using long short-term memory networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [18] Geoffrey Zweig and Patrick Nguyen, “A segmental CRF approach to large vocabulary continuous speech recognition,” in *IEEE Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 152–157.
- [19] Herman Kamper, Aren Jansen, Simon King, and Sharon Goldwater, “Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 100–105.
- [20] Herman Kamper, Aren Jansen, and Sharon Goldwater, “Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model,” in *Interspeech*, 2015.
- [21] Michael A Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky, “Rapid evaluation of speech representations for spoken term discovery,” in *Interspeech*, 2011.

- [22] H. Kamper, M. Elsner, A. Jansen, and S. J. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [23] Aren Jansen, Samuel Thomas, and Hynek Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.
- [24] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [25] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Neural Information Processing Systems (NIPS)*, 2014.
- [26] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.
- [27] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.
- [28] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Neural Information Processing Systems (NIPS)*, 2015, pp. 577–585.
- [29] Liang Lu, Xingxing Zhang, Kyunghyun Cho, and Steve Renals, "A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition," in *Interspeech*, 2015.
- [30] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.
- [31] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Rec.*, vol. 7, no. 4, pp. 669–688, 1993.
- [32] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1992, vol. 1, pp. 517–520.
- [33] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [34] Yurii Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," .
- [35] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," Tech. Rep. UCB/EECS-2010-24, EECS Department, University of California, Berkeley, Mar 2010.
- [36] Matthew D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.
- [37] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
- [38] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, Neural Information Processing (NIPS) Workshop*, 2011, number EPFL-CONF-192376.
- [39] Nicholas Léonard, Sagar Waghmare, Yang Wang, and Jin-Hwa Kim, "rnn : Recurrent library for torch," *CoRR*, vol. abs/1511.07889, 2015.
- [40] Yougen Yuan, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Learning neural network representations using cross-lingual bottleneck features with word-pair information," in *Interspeech*, 2016.
- [41] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.