Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems Supplementary Material

Yonatan Belinkov and **James Glass**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
`{belinkov, glass}@mit.edu`

A Additional experiments

A.1 Windows of features

Our main experiments used a simple frame representation by taking the output of the ASR model at frame t, $\texttt{ASR}_t(\mathbf{x})$. We also consider a window of features around the frame at time t. This improves the representation and also accounts for possible delay effects [1]. Test set results with different window widths are shown in Figure 7 (DeepSpeech model, no strides). As expected, larger windows improve the representation quality. The absolute numbers are much better than using only a single frame (+10-15%), but the overall trend for a given window size is similar: initial performance drop after the convolutional layers, then steady increase at the first recurrent layers and another drop at the top layers. The drop is somewhat more moderate than in the single frame case (compare to Figure 1b), indicating that some shifting effect may indeed be taking place, although it might be limited given that we are using bidirectional RNNs (the results in [1] are with unidirectional RNNs).

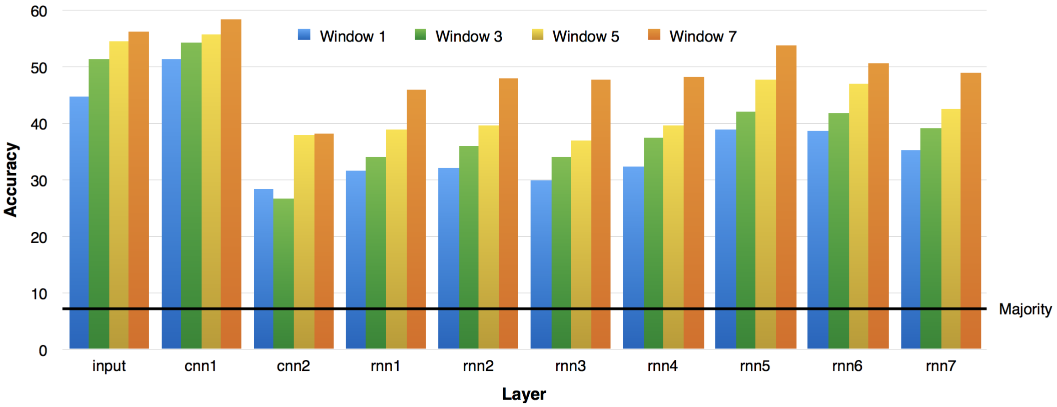

Figure 7: Frame classification accuracy using different window widths around the current frame.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

A.2 Reduced phone set

In addition to the full set of 60 phones and the coarse sound classes, we also experimented with a reduced set of 48 phones [2]. As Figure 8 shows, the trend is similar to the other phone sets. We also noticed, as with sound classes (Section 5.5), that the affricates /ch/ and /jh/ are better represented at rnn5 (F1 score of 42.5% and 34.9%, respectively) than at the input layer (7.2% and 8.3%).

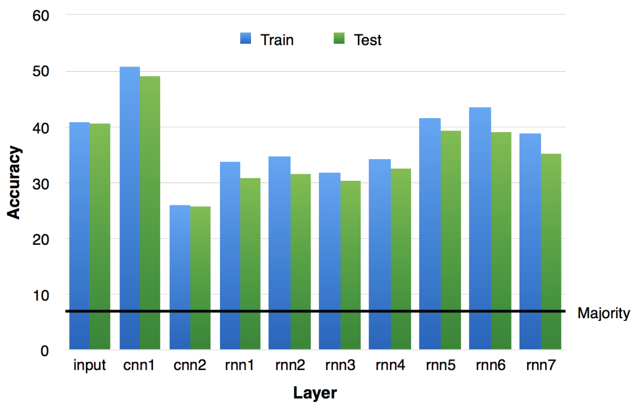

Figure 8: Frame classification accuracy with a reduced set of 48 phones.

B Visualizations of frame representations

Figure 9 shows a t-SNE [3] visualization of cluster centroids of activations from different layers in DeepSpeech2, where each cluster is assigned the phone label that had the largest number of examples in the cluster. The input layer produces clusters which show a fairly clean separation into groups of centroids with the same assigned phone. After the input layer it is less easy to detect groups, and lower layers do not show a clear structure. In layers rnn4 and rnn5 we again see some meaningful groupings, after which rnn6 and rnn7 again show less structure.

Figure 10 shows clusters that have a majority label of at least 10-20% of the examples (depending on the number of examples left in each cluster after pruning). In this case groupings are more observable in all layers, and especially in layer rnn5.

References

[1] H. Sak, F. de Chaumont Quitry, T. Sainath, K. Rao *et al.*, “Acoustic Modelling with CD-CTC-SMBR LSTM RNNs,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 604–609.

[2] K.-F. Lee and H.-W. Hon, “Speaker-independent phone recognition using hidden markov models,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[3] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

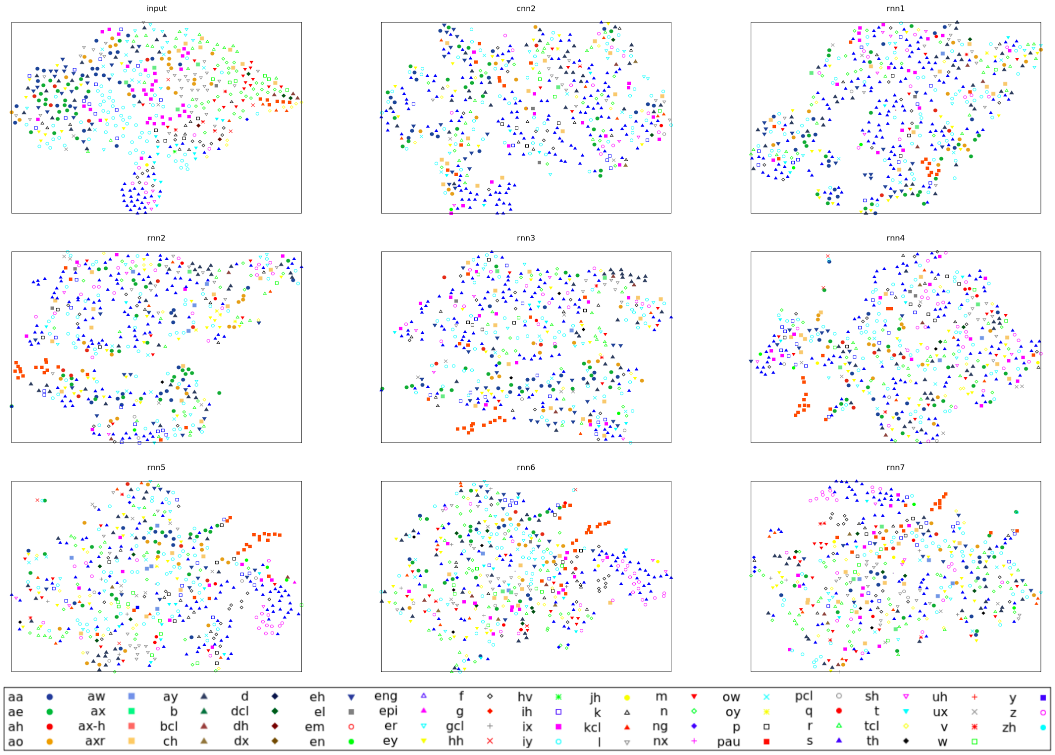

Figure 9: Centroids of all frame representation clusters using features from different layers.

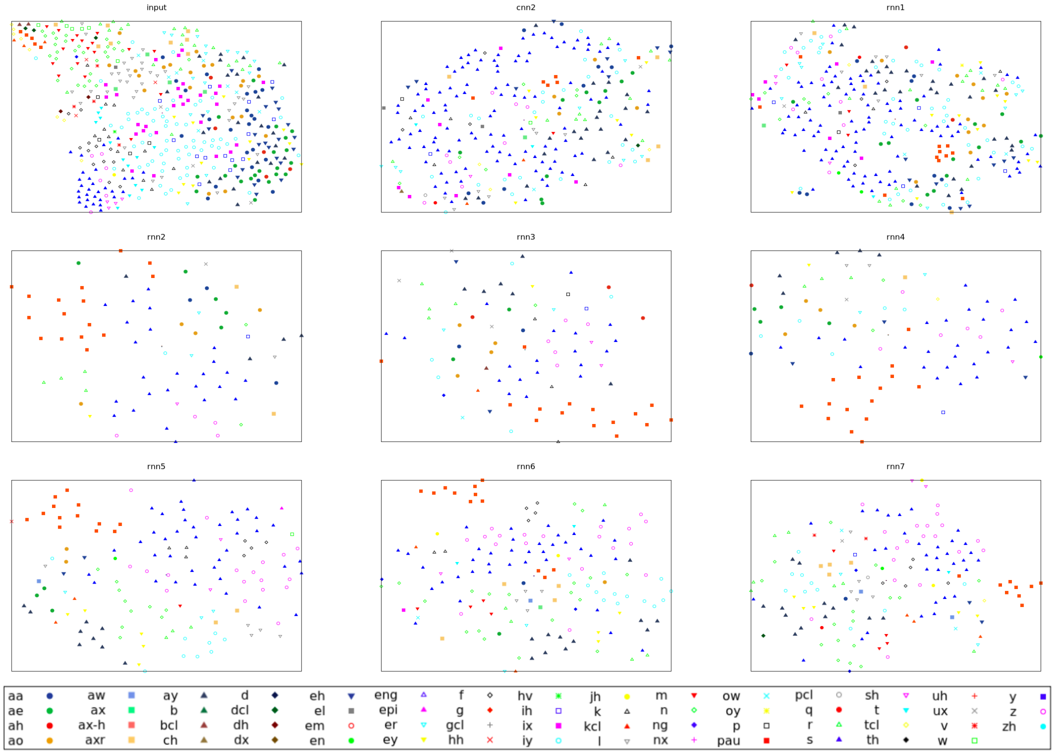

Figure 10: Centroids of frame representation clusters using features from different layers, showing only clusters where the majority label covers at least 10-20% of the cluster members.