TEMPORAL MODELING USING DILATED CONVOLUTION AND GATING FOR VOICE-ACTIVITY-DETECTION

Shuo-Yiin Chang, Bo Li, Gabor Simko, Tara N. Sainath, Anshuman Tripathi

Google Inc., U.S.A

shuoyiin,boboli,gsimko, tsainath,anshumant@google.com

Aäron van den Oord, Oriol Vinyals

Google DeepMind, London, U.K.

vdnoord,vinyals@google.com

ABSTRACT

Voice activity detection (VAD) is the task of predicting which parts of an utterance contains speech versus background noise. It is an important first step to determine which samples to send to the decoder and when to close the microphone. The long short-term memory neural network (LSTM) is a popular architecture for sequential modeling of acoustic signals, and has been successfully used in several VAD applications. However, it has been observed that LSTMs suffer from state saturation problems when the utterance is long (i.e., for voice dictation tasks), and thus requires the LSTM state to be periodically reset. In this paper, we propose an alternative architecture that does not suffer from saturation problems by modeling temporal variations through a stateless dilated convolution neural network (CNN). The proposed architecture differs from conventional CNNs in three respects: it uses dilated causal convolution, gated activations and residual connections. Results on a Google Voice Typing task shows that the proposed architecture achieves 14% relative FA improvement at a FR of 1% over state-of-the-art LSTMs for VAD task. We also include detailed experiments investigating the factors that distinguish the proposed architecture from conventional convolution.

Index Terms— CNN, voice activity detection, LSTM

1. INTRODUCTION

In many automatic speech recognition (ASR) applications the VAD is an essential component that identifies speech and filters out background noise. Such a task is often an important pre-processing stage of an ASR system to determine when to close the microphone. In short the VAD reduces computation and latency and also guides the user interface.

A typical VAD system uses a frame-level classifier with acoustic features to make speech/non-speech decisions for each audio frame (every 10ms) [1]. In a typical ASR system, the VAD needs to work accurately in challenging environments, including noisy conditions, reverberated environments and environments with background speech. Poor VAD could either accept background noise, which makes recognition slow and expensive, or reject speech which increases deletion errors (a few milliseconds of missed audio could remove an entire word).

Significant research has been devoted to finding the optimal VAD model [2, 3, 4, 5, 6]. In the literature, LSTM is a popular architecture for sequential modeling of the VAD task showing state-of-the-art performance [2, 7]. Theoretically, LSTMs can model any arbitrary length of history as it can learn to forget the past. However, in practice, the gates learned from data may not always operate as expected. In VAD tasks, LSTMs have been observed to suffer from state saturation problems in very long utterances. One solution to address this is to periodically reset the LSTM states. However, this approach is a bit ad-hoc because the time of where to reset the state is empirically chosen.

In this paper we propose a modeling alternative to LSTMs to address the saturation issue. Architectures which perform convolution in time [8, 9, 10, 11] have been explored as alternatives to LSTMs for general acoustic modeling tasks. In this paper, we adopt the WaveNet architecture [9] that models temporal patterns with dilated convolution and gated activation as shown in Fig. 1.

Comparing to conventional time convolution, dilated convolutions allow broader receptive field with fewer layers by skipping some inputs. In the proposed architecture, gated convolution activations are used to precisely control information flows. Also, residual connection is added to ease the training of very deep neural network.

We compared the proposed architecture against LSTM for a VAD task on Google Voice Typing [12] (using your voice to dictate a message on phone). Results show that the proposed architecture achieves 14% relative improvement in terms of false alarm (FA) when fixing false reject rate (FR) at 1%.

The rest of this paper is as follows. In Section 2, we describe the proposed neural network architecture. The experimental setup is described in Section 3 and results and analyses are presented in Section 4. Finally, Section 5 concludes the paper.

2. NEURAL NETWORK ARCHITECTURE

Temporal modeling using CNNs has been explored before in the literature [8, 10, 13, 14]. In this work, we explore using the WaveNet style architecture for temporal modeling. This architecture has been explored previously for text-to-speech (TTS) applications [9] but not for acoustic modeling on ASR or VAD. The proposed architecture differs from regular CNN in three respects: it uses (1) dilated causal convolutions, (2) gated activations and (3) residual connections, all of which will be described in subsequent sections.

2.1. Dilated Convolution

Temporal modeling with non-recurrent networks normally relies on an input context window. For example, in CNNs the convolution filters take in a window of adjacent frames as inputs to capture the

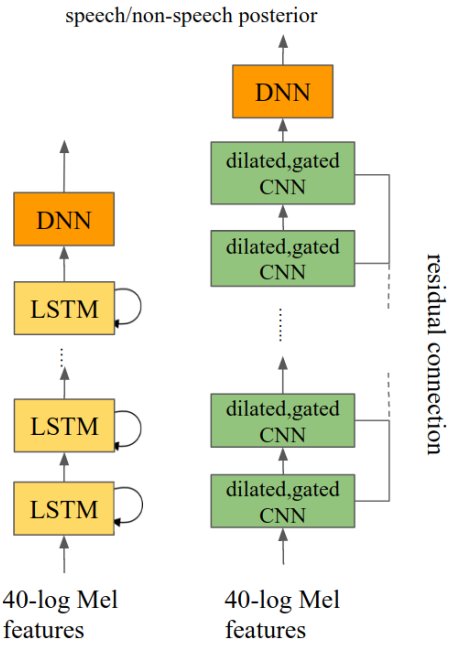

Fig. 1: *Overview of LSTM (left) and the proposed architecture (right) for voice activity detection.*

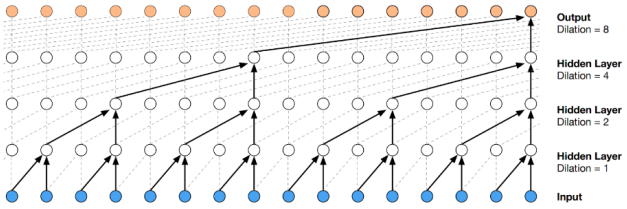

Fig. 2: *Dilated convolution with dilations 1, 2, 4 and 8.*

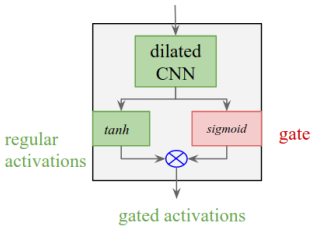

Fig. 3: *Gated convolution.*

acoustic context. Using dense input context windows increases the amount of model parameters, which is particularly a problem when doing long temporal modeling with large context windows. To address this problem, a dilated convolution (also called atrous, or convolution with holes) [9, 15, 16, 17, 18] is adopted, where the convolution filter is applied over an area larger than its length by skipping certain input values. It is equivalent to a convolution with a larger filter derived from the original filter by dilating it with zeros, but is significantly more efficient. A dilated convolution effectively allows the network to operate on a coarser scale than with a normal convolution. This is similar to pooling or strided convolutions, but here the output has the same size as the input. As a special case, dilated convolution with dilation is equivalent to the standard convolution. A stack of dilated convolutions enable networks to have very large receptive fields with just a few layers, while preserving the input resolution throughout the network as well as computational efficiency.

For speech tasks, especially the endpointing task, latency is also a crucial criterion in addition to accuracy. Normally, CNN filters takes in both left and right contexts to use both the history and future information for accurate prediction. Since latency is a concern, we need to limit the use of right context. In [9], the use of only the left context, i.e. causal convolution, is sufficient for good prediction performance for the TTS task. In this work, we adopt the same dilated causal convolution (shown in Fig. 2) for VAD.

2.2. Gated Activation

Gates play an important part in the LSTM modeling [19, 20], as they control the flow of information between time steps and layers. In WaveNet, a simple gating mechanism is adopted to control the information flow through each layer. Similarly, in this work, we use gating by first applying the hyperbolic tangent nonlinearity to the output of dilated convolution and then attenuating it with sigmoid gates (shown in Fig. 3). Hence, learnable convolution filters W are splitted into W_f for filter and W_g for gate. Final activation of layer k, h_k, is obtained by Eq. (1):

$$h_k = tanh(W_{f,k-1} * x_{k-1}) \odot \sigma(W_{g,k-1} * x_{k-1}) \qquad (1)$$

where σ is the logistic sigmoid non-linearity while $\odot$ stands for element-wise dot product.

2.3. Residual Connection

The depth of the model is important for learning robust representations, but also comes with a challenge of vanishing gradients. Residual training [21] has been found to be an effective way to address this issue and build very deep networks. For speech tasks, LSTMs have not shown improvements beyond ten layers [22], but CNNs with residual connections have shown improvements with many more layers [21]. Following a similar configuration to WaveNet [9], we also use residual connections between each layer, allowing us to train a network with 36 layers. Bypassing paths are created by accumulating outputs of each layer as shown in Fig. 4. Note that small dimensional matching layers are used to accumulate outputs of the same size. These bypassing paths are presumed to be the key factor that eases the training of very deep networks.

3. EXPERIMENTAL DETAILS

3.1. Data

We conduct experiments on about 18,000 hours of noisy training data consisting of around 6.5 million English utterances This data set is created by artificially corrupting clean utterances using a simulator to add varying degrees of noise and reverberation [23]. The clean utterances are anonymized and hand-transcribed voice typing recordings, and are representative of Google voice typing traffic on Android. Noise signals, which include music and ambient noise sampled from YouTube and recordings of daily life environments, are added to the clean utterances at SNRs ranging from 0 to 30 dB,

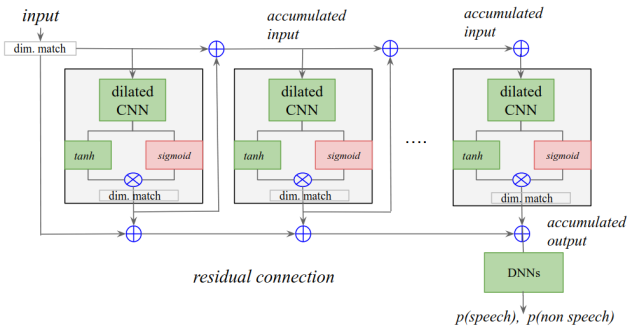

Fig. 4: *Residual connection: bypassing paths to accumulate inputs/outputs of each layer.*

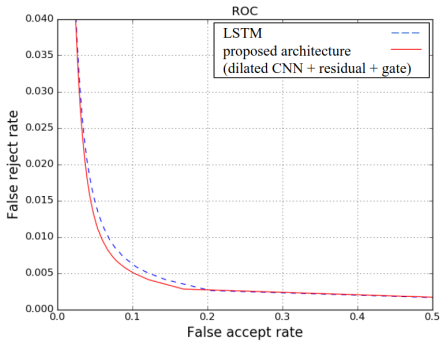

Fig. 5: *FA-FR curves for the proposed architecture and LSTM.*

with an average SNR of 11 dB. We evaluate our models using simulated noisy data. Around 15 hours (13K utterances of anonymized Android voice typing utterances were used. Noise is added using the simulator with a configuration distribution that approximately matches the training configurations. The noise snippets do not overlap with training.

3.2. Model configuration

The acoustic features used for all experiments are 40-dimensional log-mel filterbank energies, produced using a 25ms long sliding window computed every 10ms. Table 1 shows the configuration used in our experiments. Specifically, we used 64 filters for each convolution layer. The filter size was 64-d input activations with 3 frames. We used 36 convolution layers with a dilation rate repeating 1, 2, 4 and 8. A total left context of 270 frames is used given the above design parameters. In our experiments adding extra hidden layers beyond 36 did not improve the performance.

For the baseline LSTM configuration, we used 10 layers of LSTMs. Each layer consists of 64 memory cells. We also applied residual connection to LSTMs to ease the training of deeper models. Again, adding extra LSTM layers did not improve performance. The total number of parameters is roughly 400k for both LSTM and convolution models. The outputs of convolution layers or LSTMs are fed into one DNN layer with 64 hidden units, and finally a softmax layer with 2 output targets, speech and non-speech.

All networks were trained with the cross-entropy criterion using asynchronous stochastic gradient descent (ASGD) [24]. The weights for DNN layers were initialized using the Glorot-Bengio strategy described in [25], while all LSTM parameters were uniformly initialized to lie between -0.02 and 0.02. We used a constant learning rate of 2e-5.

model parameters	
number of filters per convolution layer	64
filter size	64×3
number of convolution layers	36
dilation	1,2,4,8,1,2,4,8...
context	270 left frames
number of units per DNN layer	64
parameters	~400k

Table 1: Model configuration used in the experiments.

4. RESULTS

In this section, we present our experimental results on building a VAD using dilated convolution and gating.

4.1. Results using proposed architecture vs LSTM

The FA-FR (False Accept against False Rejection) curve is frequently used to describe binary classification tasks. Here, we report the FA-FR curve for speech classification, where the speech posterior is thresholded to obtain the VAD decision. In this case, false accepts (FA) are incorrect predictions that classify the audio frames as speech when they are actually non-speech. Similarly, false rejections (FR) are speech frames misclassified as non-speech. Lower is better for both metrics. As shown in Figure 5, the proposed architecture provides better FA/FR than LSTMs. Specifically, a 14% relative improvement in FA is achieved at the operating point of 1% FR as shown in Table 2.

Table 2: FA at fixing 1% FR

model	proposed architecture	LSTM
FA	5.67	6.66

4.2. Results using proposed architecture vs conventional CNN

Next, we further analyze the importance of key factors from the proposed architecture that distinguishes it from conventional CNNs, which do not include dilation, gated activation and residual connection. Figure 6 shows the FA-FR curves for conventional CNNs, CNNs with residual connection, CNNs with gating and residual connections and dilated CNNs with gating and residual connections. Without the residual connection, performance degrades rapidly when adding more than 20 convolution layers. Hence, the conventional CNN reported in this work uses only 20 convolution layers while others have 36 convolution layers. As shown in Table 3, 16.6% improvement achieved by the residual connection, while the gate control reduced FA by another 14% relatively and dilated convolution by further 19% relatively.

Table 3: FA at fixing 1% FR

model	**FA**
conventional CNN	9.76%
+ residual connection	8.14%
+ gating	7.10%
+ dilation	5.67%

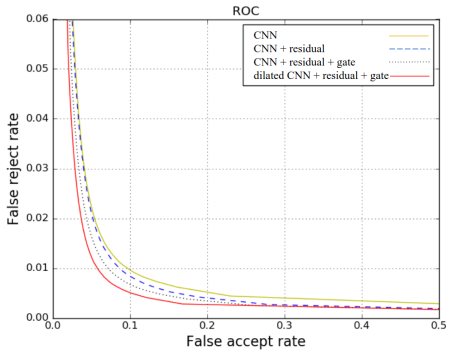

Fig. 6: *FA-FR curves for the proposed architecture and LSTM.*

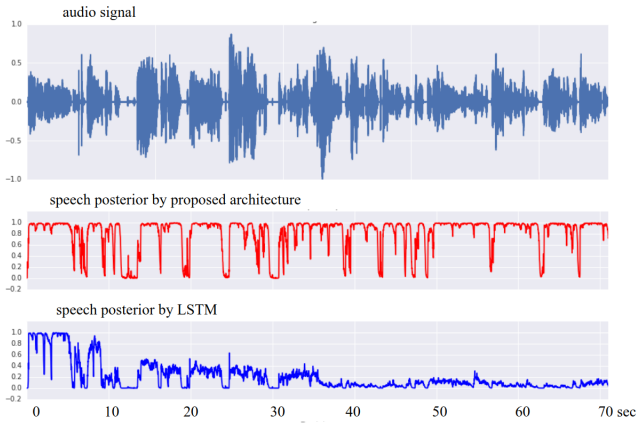

Fig. 7: *An example of an over 1 minute long audio. Posteriors of speech generated with LSTM (bottom) and the proposed architecture (above).*

4.3. Robustness To Long Audio Signals

To further understand the gains with the dilated convolutions Figure 7 shows the comparison of the speech posteriors produced by the LSTM network and the proposed dilated convolutional network for each frame of an over 1 minute long audio signal. The audio is collected from a medical conversation between doctor and patient with consent of the patient [26]. As shown in Figure 7, after processing 40 seconds of audio the LSTM becomes trapped into a dead state and start to reject the audio (speech posterior goes closer to zero). On the other hand, the proposed architecture is robust to processing longer audio because of its stateless design.

5. CONCLUSIONS

In this study we explored a neural network architecture incorporating key parts of dilated convolution, gate control and residual connection for voice activity detection. Experiments on a Google Voice Typing task illustrated that the proposed architecture achieved a 14% FA relative improvement over the LSTM when fixing the FR at 1%. Analysis showed that gate control to convolution activations and broader receptive field using dilated convolution both contribute to the improvement. Finally, the proposed architecture is more robust to processing longer audio.

6. REFERENCES

[1] F. Xie S. Van Gerven, "A comparative study of speech detection methods.," in *Eurospeech*, 1997, vol. 97.

[2] R. Zazo, Tara N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection," in *Proc. Interspeech*, 2016.

[3] S. Thomas, G. Saon, M. V. Segbroeck, and S. Naranyanan, "Improvements to the IBM speech activity detection system for the darpa rats program," in *Proc. ICASSP*, 2015.

[4] M. Graciarena et al., "All for one: feature combination for highly channel-degraded speech activity detection," in *Proc. Interspeech*, 2013.

[5] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *Proc. ICASSP*, 2013.

[6] S. Chang, B. Li, G. Simko, Tara N. Sainath, and C. Parada, "Endpoint detection using grid long short-term memory networks for streaming speech recognition," in *Interspeech*, 2017.

[7] G. Simko, M. ahannon, S. Chang, and C. Parada, "Improved end-of-query detection for streaming speech recognition," in *Interspeech*, 2017.

[8] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," 2015.

[9] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[10] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.

[11] Tara N. Sainath, Brian Kingsbury, Abdel rahman Mohamed, George E. Dahl, George Saon, Hagen Soltau, Tomas Beran, Aleksandr Y. Aravkin, and Bhuvana Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proc. ASRU*, 2013.

[12] "Type less, talk more.," in *https://www.blog.google/products/search/type-less-talk-more*.

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

[14] K. Lang, Alex H Waibel, and G. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural networks*, vol. 3, no. 1, pp. 23–43, 1990.

[15] M Holschneider, R Kronland-Martinet, J Morlet, and P Tchamitchian, "Wavelets, time-frequency methods and phase space," *A real-time algorithm for signal analysis with the help of the wavelet transform. Springer, Berlin*, pp. 289–297, 1989.

[16] P. Dutilleux, "An implementation of the algorithme à trous to compute the wavelet transform," in *Wavelets*, pp. 298–304. Springer, 1990.

[17] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[18] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[19] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] K. Greff, R. Srivastava, J. Koutník, B. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] G. Pundak and T. N. Sainath, "Highway-lstm and recurrent highway networks for speech recognition," in *Proc. Interspeech*, 2017.

[23] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani, "Generated of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *Proc. Interspeech*, 2017.

[24] J. Dean, G.S. Corrado, R. Monga, K. Chen, M. Devin, Q.V. Le, M.Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A.Y. Ng, "Large Scale Distributed Deep Networks," in *Proc. NIPS*, 2012.

[25] X. Glorot and Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," in *Proc. AISTATS*, 2014.

[26] K. Chou C. Co N. Jaitly D. Jaunzeikare A. Kannan P. Nguyen H. Sak A. Sankar J. Tansuwan N. Wan Y. Wu X. Zhang C. Chiu, A. Tripathi, "Speech recognition for medical conversation.," in *submitted ICASSP*, 2018.