Empirical Exploration of Novel Architectures and Objectives for Language Models

Gakuto Kurata\textsuperscript{1}, Abhinav Sethy\textsuperscript{2}, Bhuvana Ramabhadran\textsuperscript{2}, George Saon\textsuperscript{2}

\textsuperscript{1}IBM Research - Tokyo, Japan
\textsuperscript{2}IBM T. J. Watson Research Center, USA
\textsuperscript{1}gakuto@jp.ibm.com, \textsuperscript{2}\{asethy, bhuvana, gsaon\}@us.ibm.com

Abstract

While recurrent neural network language models based on Long Short Term Memory (LSTM) have shown good gains in many automatic speech recognition tasks, Convolutional Neural Network (CNN) language models are relatively new and have not been studied in-depth. In this paper we present an empirical comparison of LSTM and CNN language models on English broadcast news and various conversational telephone speech transcription tasks. We also present a new type of CNN language model that leverages dilated causal convolution to efficiently exploit long range history. We propose a novel criterion for training language models that combines word and class prediction in a multi-task learning framework. We apply this criterion to train word and character based LSTM language models and CNN language models and show that it improves performance. Our results also show that CNN and LSTM language models are complementary and can be combined to obtain further gains.

Index Terms: Language model, LSTM, CNN, dilated causal convolution, multi-task learning

1. Introduction

Recurrent architectures and deep learning based approaches have been the mainstay of language modeling research in the past few years. Long Short Term Memory (LSTM) based recurrent language models (LMs) have shown significant perplexity gains on well established benchmarks such as the Penn Tree Bank \cite{1} and the more recent one billion corpus \cite{2}. These results validate the potential of deep learning and recurrent models as key to further progress in the field of language modeling. As LMs are one of the core components of natural language processing (NLP) technologies such as automatic speech recognition (ASR) and Machine Translation (MT), improved language modeling techniques have translated to improvements in overall system performance for these technologies \cite{3, 4, 5}.

While recurrent models have been studied in-depth, other deep learning architectures such as convolutional models have received less attention. Recently, convolutional layers have been used in character level LMs \cite{6} as an input layer which transforms a sequence of characters to a continuous representation for an LSTM to process. CNN models have been shown to be competitive in other NLP tasks such as sentence classification \cite{7}. However, CNN based language modeling has not received much attention in ASR. To the best of our knowledge, this is the first research that draws comparisons between CNNs and LSTMs in the context of a state-of-the-art speech recognition task.

In this paper, we present extensive empirical results comparing CNN and LSTM based LMs on state-of-the-art broadcast news and conversational ASR systems built on publicly available data. The novel CNN LM architecture proposed in this paper leverages dilated causal convolution to efficiently exploit long range history. We also propose a multi-task learning framework for training LMs by combining word and class prediction and demonstrate the improved performance on both, CNN and LSTM LMs. Our results also show that CNN and LSTM LMs are complementary and can be combined to obtain further gains. Fully utilizing the above proposed methods contributed to achieving the current best reported accuracy in the widely-studied Switchboard (SWB) and CallHome (CH) subsets of the NIST Hub5 2000 evaluation testset \cite{8}.

This paper has three main contributions:

\begin{itemize}
  \item a novel CNN LM using dilated causal convolutions,
  \item a multi-task learning for neural network LMs, and
  \item impact of the above proposed methods in state-of-the-art ASR tasks with publicly available broadcast news and conversational telephone speech data.
\end{itemize}

2. Dilated causal convolution for language modeling

In this section, we describe our proposed convolutional LM architecture based on dilated causal convolutions and its extension.

Recurrent neural networks and their variants can capture long range dependencies and have been shown strong performance in language modeling \cite{9}. Figure 1 and Figure 2 show word based (\textit{Word-LSTM}) and character based LSTM (\textit{Char-LSTM}) architectures used in this paper. In speech synthesis, WAVE\textsuperscript{NET} \cite{10} introduced a dilated causal convolution architecture to build long range receptive fields. We build upon this dilated convolution architecture to efficiently exploit long range history for language modeling, as shown in Figure 3, henceforth referred to as \textit{Word-DCC}.

In order to understand the behavior of \textit{Word-DCC} models, we begin with a toy example shown in Figure 4(a). The input to the convolutional layer is a word-embeddings matrix of dimension \(4 \times 4\) comprising of 4-dimensional word embeddings over 4 context words. These are transformed into a \(4 \times 2\) matrix by dilated causal convolution with 4 filters whose sizes are \(4 \times 2\). That is subsequently fed into a second convolution layer, followed by a fully-connected layer and a final softmax layer. Convolution and fully connected layers are wrapped by residual connections as shown in \cite{10, 11, 12}. In addition to \textit{Word-DCC}, we also propose an extension, \textit{Word-DCC+C}, as shown in Figure 4(b), that includes an additional convolution layer before...
the final convolution layer. This convolution has a smaller filter size and can operate in either dimension. We can interpret this additional convolution operation as simply adding more non-linearity, transforming feature maps, or weighting of representations from different time indices.

3. Multi-task learning of word and class prediction

Following the success of multi-task learning in various NLP tasks [13, 14, 15, 16], we propose to use multi-task learning in LM training. In the proposed multi-task learning framework, the main task is to predict the next word given its word history and the sub task is to predict the class of the next word given its word history. Figure 5 shows Word-LSTM-MTL, where the proposed multi-task learning is applied to Word-LSTM. All neural network layers other than the final softmax layer are shared between the two prediction tasks. There are two independent word prediction and class prediction softmax layers.

In the proposed multi-task learning framework, the classes to be predicted are obtained by clustering the words in the vocabulary using Brown clustering [17]. For both word and class prediction, the cross-entropy $XE_W$ of predicting the next word given its word history and the cross-entropy $XE_C$ of predicting the next class given its word history were used as objectives in

Figure 1: Word-LSTM: Word based LSTM language model.

Figure 2: Char-LSTM: Character based LSTM language model.

Figure 3: Word-DCC: Word based dilated causal convolution language model.

Figure 4: Detail of Word-DCC (Figure 3) and its extension Word-DCC+C. Color is different from Figure 3 for detailed explanation.

Figure 5: Word-LSTM-MTL: Word based LSTM language model with multi-task learning.
training. The training process minimizes the weighted summation of these two objectives, $X_{\text{MTL}}$ as

$$X_{\text{MTL}} = (1 - \lambda)X_W + \lambda X_C,$$

where $\lambda$ is a scaling parameter. The class prediction branch is discarded and the word prediction branch is used for perplexity calculation and $N$-best rescoring.

The proposed multi-task learning can be used with other types of neural network based LMs, such as feed forward neural network [18], other CNN [12] and gated recurrent unit (GRU) [19].

4. Experiments

We report perplexity and speech recognition experiments on two transcription tasks: English broadcast news and conversational telephone speech. For speech recognition experiments, we generated $N$-best lists from lattices produced by the baseline system for each task and rescoring them with LSTM and/or CNN LMs. LM probabilities were linearly interpolated and the interpolation weights of LMs were estimated using the heldout data.

The Word-LSTM and Char-LSTM models serve as the baseline models. The Word-LSTM model consists of one word-embeddings layer, two LSTM layers, one fully-connected layer, and one softmax layer, as described in Figure 1. The upper LSTM layer and the fully-connected layers allow residual connections [11]. Dropout is applied to the vertical dimension only and not applied to time dimension [1]. The Char-LSTM model includes an additional LSTM layer to estimate the word-embeddings from character sequence as described in Figure 2 [20]. Both models minimize the standard cross-entropy objective during training.

We compare the proposed CNN LMs with the Word-LSTM and Char-LSTM. We also study the impact of multi-task learning on each of them, using Word-LSTM-MTL, Char-LSTM-MTL, Word-DCC-MTL, and Word-DCC+C-MTL.

4.1. Network configuration and hyper-parameters

Word-LSTM model uses word embeddings of dimension 256 and 1,024 units in each hidden layer. The fully connected layer uses a gated linear unit [12] and the network is trained with a dropout rate of 0.5. Char-LSTM model uses character embeddings of dimension 32 and the LSTM layer for word embedding estimation has 256 hidden units. The upper LSTM layers and above layers are the same with the Word-LSTM.

The window sizes for the convolutional layers in Word-DCC were further tuned to minimize the perplexity on the heldout data set. For Word-DCC, we tried $\{2,3,4\}$ for convolution window size $w_1$ for the first convolution layer and $\{2,4,8,16\}$ for window size $w_2$ for the second convolution layer. For the Word-DCC+C model, the number of filters and their size were set to $w_2$ and $1 \times w_2$ so that the size of the hidden representation was not changed by this additional convolution operation. In the multi-task learning framework, the number of classes was chosen from the set of $\{32, 64, 128, 256\}$ and the scaling parameter $\lambda$ was selected from the set of $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ parameters.

The optimizer used was Adam [21] and a self-stabilization term to coordinate the layer-wise learning rates [22] was introduced.

Table 1: Perplexity on broadcast news with various LMs.

<table>
<thead>
<tr>
<th>$n$-gram</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-LSTM</td>
<td>98.74</td>
</tr>
<tr>
<td>Word-LSTM-MTL</td>
<td>96.47</td>
</tr>
<tr>
<td>Char-LSTM</td>
<td>103.57</td>
</tr>
<tr>
<td>Char-LSTM-MTL</td>
<td>102.77</td>
</tr>
<tr>
<td>Word-DCC</td>
<td>119.75</td>
</tr>
<tr>
<td>Word-DCC-MTL</td>
<td>115.56</td>
</tr>
<tr>
<td>Word-DCC+C</td>
<td>112.81</td>
</tr>
<tr>
<td>Word-DCC+C-MTL</td>
<td>110.51</td>
</tr>
</tbody>
</table>

4.2. Broadcast news

Broadcast news evaluation was done on the Defense Advanced Research Projects Agency (DARPA) Effective Affordable Reusable Speech-to-Text (EARS) rt04 testset that contains approximately 4 hours of data. We used two types of acoustic models. The first model is a discriminatively-trained, speaker-adaptive Gaussian Mixture Model (GMM) acoustic model (AM) trained on 430h of broadcast news audio [23]. The second model is a Convolutional Neural Net (CNN) acoustic model trained on 1,000h of similar audio data. The CNN-based AM was first trained with cross-entropy training [24] and then with Hessian-free state-level Minimum Bayes Risk (sMBR) sequence training [25, 26].

The baseline LM is a conventional word 4-gram model trained on a total of 350M words from multiple sources [27] with a vocabulary size of 84K words. For training LSTM and CNN LMs, we used a 12M-word subset of the original 350M-word corpus, as done in [28]. The hyper-parameters were optimized on a heldout data set. The window sizes for convolution in the two layers was set to 2 and 8 respectively. This implies that 16 words were considered in the context for Word-DCC, Word-DCC+C, Word-DCC-MTL, and Word-DCC+C-MTL. The number of classes for Char-LSTM-MTL, Word-DCC-MTL, and Word-DCC+C-MTL models was chosen to be 256 and Word-LSTM-MTL uses 128 classes. The scaling parameter was set to 0.3 for Char-LSTM-MTL and 0.1 for the remaining models.

Table 1 illustrates the perplexity of these models on the heldout set. While the Word-DCC model performs better than
and decoding [8]. Lattices were generated using this acoustic model for the SWB and CH subsets of the NIST Hub5 2000 evaluation data set. The acoustic model is an LSTM and ResNet acoustic model based on the CNN-based acoustic model. This resulted in a further reduction in WER with an interpolated model comprising of all three multitask learning frameworks. LSTM and CNN LMs were built with a vocabulary of 85K words. In the first pass, the LMs were trained with the corpus of 560M words consisting of publicly available text data from LDC, including Switchboard, Fisher, Gigaword, and Broadcast News and Conversations. In a second pass, this model was refined further with just the transcripts (approximately, 24M words) corresponding to the 1,975 hour audio data used to train the acoustic models [3]. The window sizes for the convolution operations were set to 2 and 16. This implies that the context of the Word-DCC+C and Word-DCC+C-MTL models spans 32 history words. The number of classes for the Word-LSTM-MTL and Char-LSTM-MTL models was 128, while it was set at 256 for the Word-DCC+C-MTL model. The scaling parameter was 0.1 for all models. Again, these hyper-parameters were optimized on a heldout data set.

The perplexity of these LMs on the heldout data set are tabulated in Table 3. It can be seen that the Word-DCC+C model is better than the n-gram and model-M LMs. However, it is worse than the Word-LSTM model while displaying comparable performance to the Char-LSTM. Marginal reduction in perplexity was observed with all models trained with multi-task learning.

The WERs on the SWB and CH subsets are tabulated in Table 4. We demonstrate a significant reduction in WER over a strong n-gram and model-M baseline with the proposed Word-DCC+C model on both these subsets. Perplexity reduction through multi-task learning translated into WER reduction with most of the LMs on the SWB subset and with all the LMs for the CH subset. To investigate the complementarity of LSTM and CNN LMs, we rescoring as before with all three multi-task learning based models and obtained further reduction in WER. The WERs of 5.5% and 10.3% achieved by the combination of LSTM and CNN LMs are the best reported WER on SWB and CH tasks in the literature[1].

5. Conclusion

In this paper, we presented an empirical comparison of a range of LSTM and CNN architectures for language modeling on state-of-the-art ASR tasks. We proposed novel LM architectures and multi-task learning, that helped achieve the best reported result on SWB and CH subsets [8]. We conclude:

- Introduction of an additional convolution layer in the Word-DCC LM with a different convolution direction resulted in reduction in perplexity.
- The novel Word-DCC+C LM provided gains over an n-gram LM baseline in the broadcast news transcription task and over a strong baseline consisting of n-gram and model-M LMs on conversational telephone transcription tasks.
- The proposed multi-task learning framework demonstrated steady perplexity reduction, which translated into WER reduction in most configurations.
- CNN and LSTM LMs are complementary and can result in further reduction in WER.

6. Acknowledgment

We would like to thank Dr. Hong-Kwang J. Ku of IBM T.J. Watson Research Center for his valuable support.

---

1The entire system description is given in [8]
7. References


