Feed Forward Pre-training for Recurrent Neural Network Language Models

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
INTERSPEECH 2014 15th Annual Conference of the International Speech Communication Association

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Feed Forward Pre-training for Recurrent Neural Network Language Models

Siva Reddy Gangireddy, Fergus McInnes and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, UK
s.gangireddy@sms.ed.ac.uk, {fergus.mcinnes, s.renals}@ed.ac.uk

Abstract

The recurrent neural network language model (RNNLM) has been demonstrated to consistently reduce perplexities and automatic speech recognition (ASR) word error rates across a variety of domains. In this paper we propose a pre-training method for the RNNLM, by sharing the output weights of the feed forward neural network language model (NNLM) with the RNNLM. This is accomplished by first fine-tuning the weights of the NNLM, which are then used to initialise the output weights of an RNNLM with the same number of hidden units. We have carried out text-based experiments on the Penn Treebank Wall Street Journal data, and ASR experiments on the TED talks data used in the International Workshop on Spoken Language Translation (IWSLT) evaluation campaigns. Across the experiments, we observe small improvements in perplexity and ASR word error rate.

Index Terms: Language Modelling, Recurrent Neural Network, Pre-training, Automatic Speech Recognition, TED talks

1. Introduction

Until recently, smoothed n-gram language models (LMs) defined the state-of-the-art in language modelling, and most large scale automatic speech recognition (ASR) systems used a 3-gram, 4-gram or 5-gram language model [1]. In an n-gram LM, the probability of a word sequence is approximated as the product of conditional probabilities of each word in the sequence conditioned on the previous n–1 words. An unseen n-gram is generalized by interpolation with a lower order n-gram model.

The n-gram LM has two principal drawbacks. First, the language model is defined in an unstructured binary space of word indices with no way to learn commonalities between words, or generalise across them. Second, the size of the model increases exponentially with respect to n. Neural network language models (NNLMs), which learn distributed continuous-valued word representations to compute the probability of a word given the previous n–1 words, have been proposed to address these problems [2, 3, 4]. In an NNLm, an unseen n-gram is generalized based on distance and similarity between the words in this continuous space. If the words in an unseen n-gram are close enough to seen words during training, then a similar probability will be assigned.

Similar to n-gram LMs, feed-forward NNLMs have a finite context of n–1 words — although the size of an NNLm grows linearly, rather than exponentially, with n. Recurrent neural network language models (RNNLMS) can learn contexts of potentially infinite length using a recurrently connected hidden layer. Experimental results using RNNLMS have shown that they consistently result in reduced perplexities and word error rates (WERs) compared with n-gram models and with feed-forward NNLMs [5, 6].

Recurrent neural networks (RNNs) are typically trained using back-propagation through time (BPTT) [7] in which the network is “unrolled” through time, and the error signal is back-propagated through multiple time steps. The basic form of this algorithm can be problematic to use in practice, owing to the problem of vanishing or exploding gradients [8, 9]. This problem was also identified, and addressed, by Robinson [10, 11] who introduced a modified version of BPTT, in which the gradient signals were used to increase or decrease a step-size for each weight, rather than being used directly in the weight updates. This training method proved to be a reliable way to train RNN acoustic models [12], with order 10^3 parameters. An alternative technique for training RNNs, known as real-time recurrent learning (RTRL) [13], does not require transmission of an error signal back through multiple time frames, at the cost of greatly increased space and computational requirements — O(n^2) and O(n^3) respectively, for an n-unit recurrent network, compared with O(nT) and O(n^2) for BPTT, back-propagating through T time steps. Both BPTT and RTRL correspond to special cases of the technique known as algorithmic differentiation (AD): BPTT is an example of reverse mode (or adjoint mode) AD, and RTRL is an example of forward tangent linear AD [14]. The problem of training RNNs using BPTT has been further addressed by Sutskever et al [15], who highlight the importance of initialisation when training RNNs, setting the scale of the input weights such that the learning dynamics of the RNN neither quickly “forget” the hidden state, nor cause the error gradients to explode.

In this work we investigate the whether transferring information from a trained feed-forward NNLm can be used to initialise an RNNLM to improve the training performance in terms of time and accuracy. We have conducted text-based experiments (measured using perplexity) using the Penn Tree Bank (PTB) Wall Street Journal data, and ASR-based experiments using the TED talks task that we have previously investigated as part of the International Workshop on Spoken Language Translation (IWSLT) evaluation campaigns [16, 17].

2. Model architectures

2.1. Feed forward neural network language model

The architecture of a feed-forward NNLm is shown in Figure 1. The inputs to the neural network are indices of the previous n–1 words, encoded using 1 of N (one-hot) coding. The feature vectors of each of the previous n–1 words are projected to a P-dimension continuous space using an N × P weight matrix. Since we use a one-hot representation for the words, this sim-
In this work, we use a factored output layer \[6\] to reduce the computational complexity. The architecture of an RNN is shown in Figure 2. The input to the network at time \(t\) is the index of the previous word and the state of the hidden layer at time \(t-1\). The index of the previous word is encoded using 1 of \(N\) coding. The same weight matrix is used to project other words in the context. The activations of projection and hidden layer, and the output probability distribution are computed as follows:

\[
h_j = \tanh \left( \sum_{i=1}^{(n-1)P} u_{ji} x_i + b_j \right) \quad \forall j = 1, \ldots, H
\]

\[
a_k = \sum_{j=1}^H v_{kj} h_j + b_k
\]

\[
y_k = \frac{e^{a_k}}{\sum_{k=1}^N e^{a_k}} \quad \forall k = 1, \ldots, N.
\]  

In this work, we use a factored output layer \[6\] to reduce the computational complexity.

### 2.2 Recurrent neural network language models

The architecture of a RNN is shown in Figure 2. The input to the network at time \(t\) is the index of the previous word and the state of the hidden layer at time \(t-1\). The index of the previous word is encoded using 1 of \(N\) coding.

The hidden layer at time \(t\) and the output probability distribution are computed as follows:

\[
x(t) = [w(t), s(t-1)]
\]  

where \(w(t)\) is the input vector, \(s(t)\) is the state of the hidden layer at time \(t\), and \(y(t)\) is the output probability distribution. Again, we use a factored output layer \[6\] to reduce the computational complexity.

### 2.3 Pre-trained Recurrent Neural Network Language Model (PT-RNNLM)

In Figure 3, we show an architecture in which an RNNLM and a feed-forward NNLM are combined. The feed-forward and recurrent networks have the same number of hidden and output nodes, and the output layer weights of the two networks are shared. In the first phase of training, only the feed-forward weights are trained (using back-propagation). When the feed-forward network has been trained for a number of iterations, the output layer weights are shared with the RNNLM, which may be interpreted as an initialisation or pre-training. During testing we only use the RNNLM part of PT-RNNLM to compute the probability distribution. Again, we used a factored output layer to reduce the computational requirement.

We evaluated these models using text data (Wall Street Journal) and a perplexity measure, and using speech data (TED talks) and a word error rate (WER) measure.

### 3. Text experiments

We report performed experiments on the Wall Street Journal (WSJ) data in the Penn Treebank (PTB)\(^1\). Sections 0–20 were used as training data, sections 21–22 were used as validation data, and sections 23–24 used as test data. A 10,000 (10k) word vocabulary was used, with out of vocabulary words being mapped to a special token \(<unk>\). The number of tokens

\(^1\)http://www.cis.upenn.edu/~treebank/
There is a significant reduction in PPL after just one iteration of pre-training, and a slight downward trend for a further eight iterations, before the curve flattens. From this we can conclude a few iterations of pre-training is necessary to improve the prediction accuracy of RNNLM, and the feed-forward network does not need to be trained to completion.

### Table 1: Perplexity of n-gram, NNLM, RNNLM and PT-RNNLM trained on WSJ data. 100 hidden nodes and 100 output classes are used in NNLM, RNNLM and PT-RNNLM

<table>
<thead>
<tr>
<th>Model</th>
<th>Valid. PPL</th>
<th>+KN3</th>
<th>Test PPL</th>
<th>+KN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-gram</td>
<td>162.4</td>
<td>-</td>
<td>152.9</td>
<td>-</td>
</tr>
<tr>
<td>NNLM</td>
<td>173.4</td>
<td>143.0</td>
<td>162.4</td>
<td>135.1</td>
</tr>
<tr>
<td>RNNLM</td>
<td>155.5</td>
<td>130.9</td>
<td>148.3</td>
<td>124.9</td>
</tr>
<tr>
<td>PT-RNNLM</td>
<td>150.7</td>
<td>128.7</td>
<td>143.4</td>
<td>123.0</td>
</tr>
</tbody>
</table>

Figure 4: The number of iterations of pre-training Vs perplexity of PT-RNNLM

The OOD data sources include Europarl, News Comment, News Crawl, and Gigaword. In the final ASR system KN3 smoothed 3-gram and 4-gram LMs are used for decoding and lattice rescoring. The relevant sentences to the task domain are selected by computing the cross entropy difference (CED) score \[21\] for each sentence \(s\):

\[ D_S = \{ s | H_I(s) - H_O(s) < \tau \} \]  \hspace{1cm} (6)

where \(H_I(s)\) is a cross-entropy of a sentence with a LM trained on in-domain data, \(H_O(s)\) is a cross-entropy of a sentence with a LM trained on a random subset of the OOD data of similar size to the TED in-domain data, and \(\tau\) is a threshold to control the size of \(D_S\).

4.2. ASR system

The ASR system \[18, 19\] uses deep neural network-based acoustic models in both Tandem and Hybrid configurations, both of which use the Multi Level Adaptive Neural Network (MLAN) architecture for domain adaptation \[18\]. In the first step, deep neural networks were trained on out-of-domain (OOD). In the next step, the bottleneck features generated on in-domain data using DNNs trained on OOD data are concatenated with the in-domain features. These concatenated features are used to train Tandem and Hybrid MLAN HMMs.

The training data consisted of 143 hours of transcribed and aligned TED lectures, and 127 hours of transcribed AMI\(^2\) meeting data used to train the OOD networks in MLAN. To avoid the overlap between development and evaluation sets all the selected lectures for training are dated before 2010. Tandem MLANs are trained with four hidden layers, each consists of 1024 hidden neurons. Six hidden layers are used in Hybrid MLANs, each layer consists of 2048 hidden neurons. All the scripts are written in Theano \[20\] and the models were trained on NVIDIA GeForce GTX 690 GPUs.

4.3. Language models

4.3.1. n-grams

The ASR system uses Kneser-Ney (KN) smoothed n-gram language models for decoding and lattice rescoring. The n-gram language models are obtained by interpolating the language models trained on 2.4M in-domain and 312M OOD word tokens. Given the mismatch between in-domain and OOD data, the relevant sentences to the task domain are selected by computing the cross entropy difference (CED) score \[21\] for each sentence \(s\):

\[ D_S = \{ s | H_I(s) - H_O(s) < \tau \} \]  \hspace{1cm} (6)

where \(H_I(s)\) is a cross-entropy of a sentence with a LM trained on in-domain data, \(H_O(s)\) is a cross-entropy of a sentence with a LM trained on a random subset of the OOD data of similar size to the TED in-domain data, and \(\tau\) is a threshold to control the size of \(D_S\).

The OOD data sources include Europarl, News Comment, News Crawl, and Gigaword. In the final ASR system KN smoothed 3-gram and 4-gram LMs are used for decoding and lattice rescoring, respectively.

---

\(^2\)http://www.ted.com

\(^3\)http://www.iwslt2013.org

\(^4\)https://www.idiap.ch/dataset/ami
In this work we have proposed feed forward pre-training for RNNLM by sharing the output weights of NNLM with the RNNLM. We report perplexity results on WSJ/PTB data, with the PT-RNNLM reducing perplexity on both the validation and test sets. We also investigated how many iterations of feed forward pre-training is necessary to improve the prediction accuracy of RNNLM. We have observed a few iterations of pre-training is sufficient. Finally we report the WERs on by rescoring 100-best lists of Tandem and Hybrid MLAN HMMs for the IWSLT TED talk recognition task, and observed a small but significant reduction in WER from the PT-RNNLM over the RNNLM.

There are a number of further approaches which we are investigating to combine feed-forward and recurrent NNLMs. This includes co-training rather than pre-training (i.e. training the output weights is interleaved between the feed-forward and recurrent neural networks), sharing the projected continuous space representations between the two architectures and hence were not independent samples, but the speaker and topic changed between talks. Also it was the overall difference between RNNLM and PT-RNNLM that was of interest, rather than the difference on a specific training set size and with a specific acoustic model type. Therefore, a paired samples t-test (two-tailed) was run on the per-talk average word error rates with RNNLM and PT-RNNLM, where the averaging for each talk and language modelling technique was over the three training data sizes (7.4M, 12.4M and 22.4M words) and the tandem and hybrid models. This was done on the full set of 27 talks (8 in dev2010, 11 in tst2010 and 8 in tst2011). The result was a highly significant difference ($p = 0.000275$) in favour of PT-RNNLM. This is significant at the 0.001 level, as recommended in a recent analysis of significance levels with regard to strength of evidence assessed by the corresponding Bayesian tests [23].
6. References


