AUTOMATIC LEARNING OF SUBWORD DEPENDENT MODEL SCALES

Felix Meyer\textsuperscript{1}, Wilfried Michel\textsuperscript{1,2}, Mohammad Zeineldeen\textsuperscript{1,2}, Ralf Schlüter\textsuperscript{1,2}, Hermann Ney\textsuperscript{1,2}

\textsuperscript{1}Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52074 Aachen, Germany
\textsuperscript{2}AppTek GmbH, 52062 Aachen, Germany

felix.sebastian.meyer@rwth-aachen.de, \{michel,zeineldeen,slueter,ney\}@cs.rwth-aachen.de

ABSTRACT

To improve the performance of state-of-the-art automatic speech recognition systems it is common practice to include external knowledge sources such as language models or prior corrections. This is usually done via log-linear model combination using separate scaling parameters for each model. Typically these parameters are manually optimized on some held-out data.

In this work we propose to optimize these scaling parameters via automatic differentiation and stochastic gradient decent similar to the neural network model parameters. We show on the LibriSpeech (LBS) and Switchboard (SWB) corpora that the model scales for a combination of attention-based encoder-decoder acoustic model and language model can be learned as effectively as with manual tuning. We further extend this approach to subword dependent model scales which could not be tuned manually which leads to 7\% improvement on LBS and 3\% on SWB. We also show that joint training of scales and model parameters is possible and gives additional 6\% improvement on LBS.

Index Terms— model combination, scale tuning, shallow fusion

1. INTRODUCTION

Attention-based encoder-decoder (AED) models \cite{1,2} are investigated by many researchers in the field of automatic speech recognition (ASR) due to their simple modeling approach and end-to-end nature. It is, however, yet unclear how to best make use of unpaired text-only data. A common approach to increase the performance of AED models is the inclusion of an external language model (LM). These LMs are trained on text-only data and can therefore encode information from large text corpora that otherwise cannot be used directly.

The simplest solution to integrate an external LM is to combine the scores of acoustic model (AM) and LM via a log-linear combination. This approach is also called shallow fusion \cite{3}. Previous investigations have shown that shallow fusion yields good results in most learning scenarios \cite{4}. In comparison with the usual setup, where AM and LM are trained separately and later combined, it is also possible to train both models jointly. In this case, parameters are often initialized with pretrained values. Investigations have shown that training the AM parameters while keeping an external LM fixed can yield good improvements \cite{5}.

The benefit here is likely due to the suppression of an internal language model (ILM) in the decoder of the AED model. When AED models are trained, they learn an ILM from the transcriptions of the parallel training data \cite{6,7}. Because of conflicts between internal and external LM it is worthwhile to subtract or suppress the ILM when including an additional external LM. An overview and a comparison of corresponding methods can be found in \cite{8} and includes ILM estimation via context LMs.

In most of the mentioned approaches, different models are combined with the help of scale parameters that control the influence of each component. These scale parameters have to be tuned manually, which is usually done via grid search. In this work we introduce a method to learn these scale parameters via automatic differentiation \cite{9} and stochastic gradient decent \cite{10}. This opens up the possibility to extend the log-linear combination to include more tunable parameters. We use this to investigate the use of individual scale parameters per output token, similar to \cite{11}.

2. LOG-LINEAR MODEL COMBINATION

2.1. Subword Agnostic Scales

When integrating an external LM into an ASR system, the definition of the posterior prediction probability changes. The probability distributions for the AM $p_{AM}(w^N_1 | x^T_1)$, and the LM $p_{LM}(w^N_1)$ have to be combined. To receive a valid probability distribution, we have to perform a renormalization.

We propose two ways to achieve this renormalization. The straightforward way is to perform it on a sentence level resulting in:

$$\tilde{p}(w^N_1 | x^T_1) = \frac{\alpha p_{AM}(w^N_1 | x^T_1) \cdot \beta p_{LM}(w^N_1)}{\sum_{w^N_1} \alpha p_{AM}(w^N_1 | x^T_1) \cdot \beta p_{LM}(w^N_1)}$$

where \(\alpha, \beta \in \mathbb{R}\) are scale parameters controlling the influence of AM and LM. We note that in general, it is not possible...
to compute the above probability exactly, because of the sum
over all possible sentences. In decoding, however, because
the argmax is used to determine the best transcription and the
denominator is independent of the argument, the normaliza-
tion can be omitted. We note that in this case, only the ratio
of the two parameters matters. Therefore, one of them can be
fixed to 1. This results in the following decision rule:
\[
\hat{w}_1^N = \arg \max_{N, w_1^N} \{ \log p_{\text{AM}}(w_1^n | x_1^T) + \beta \log p_{\text{LM}}(w_1^n) \}
\]
(2)

In the literature, this way of combining AM and LM is com-
monly referred to as shallow fusion [3, 4].

The normalization can also be done on a per token basis:
\[
\tilde{p}(w_1^n | x_1^T) = \prod_{n=1}^N p_n(w_n | w_1^{n-1}, x_1^T)
\]
(3)
\[
= \prod_{n=1}^N \frac{p_{\text{AM}, n}^{\alpha}(w_n | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(w_n | w_1^{n-1})}{\sum_{w_n} p_{\text{AM}, n}^{\alpha}(\tilde{w}_{1:n} | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(\tilde{w}_{1:n} | w_1^{n-1})}
\]
(4)

We note that in this case both scale parameters matter, even
when using the argmax.

2.2. Subword Dependent Model Scales
In this work we extend the model scales by introducing in-
dividual scale parameters on a per BPE subword unit level.
That is, each subword unit gets an individual AM and LM
scale. We redefine \( \alpha := \alpha w_1, \ldots, \alpha w_k \in \mathbb{R}^k \) and \( \beta := \beta w_1, \ldots, \beta w_k \in \mathbb{R}^k \) where \( w_1, \ldots, w_k \) are all possible sub-
word units and \( k \) is the total number of subword units.
The definition for the sentence level normalized prediction proba-
bility changes to
\[
\tilde{p}(w_1^n | x_1^T) = \prod_{n=1}^N \frac{p_{\text{AM}, n}^{\alpha}(w_n | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(w_n | w_1^{n-1})}{\sum_{w_n} p_{\text{AM}, n}^{\alpha}(\tilde{w}_{1:n} | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(\tilde{w}_{1:n} | w_1^{n-1})}
\]
(5)

and the token level normalized probability to
\[
\tilde{p}(w_1^n | x_1^T) = \prod_{n=1}^N \frac{p_{\text{AM}, n}^{\alpha}(w_n | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(w_n | w_1^{n-1})}{\sum_{w_n} p_{\text{AM}, n}^{\alpha}(\tilde{w}_{1:n} | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(\tilde{w}_{1:n} | w_1^{n-1})}
\]
(6)

3.3. Training Criteria
To train the scales we have to define a suitable training cri-
terion. In analogy to the AM training we first use the cross
entropy (CE) criterion. For simplicity we chose the per token
renormalization from Equation (4) which leads to
\[
F_{\text{CE}} = \log \tilde{p}(w_1^n | h_1^T)
\]
(7)
\[
= \frac{N}{n=1} \log \frac{p_{\text{AM}, n}^{\alpha}(w_n | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(w_n | w_1^{n-1})}{\sum_{\tilde{w}_n} p_{\text{AM}, n}^{\alpha}(\tilde{w}_1^n | w_1^{n-1}, x_1^T) \cdot p_{\text{LM}, n}^{\beta}(\tilde{w}_1^n | w_1^{n-1})}
\]
(8)

This criterion does, however, not reflect the criterion that is
used in the manual tuning process where the word error rate
(WER) of the dev set is used directly. We therefore decided to
also investigate the automatic learning of model scales with a
minimum word error rate (minWER) training criterion similar
to [12] [13]. The training criterion uses the sentence level
renormalization and is given by
\[
F_{\text{minWER}} = \sum_{N, w_1^N} \tilde{p}(w_1^n | h_1^T) \cdot A(w_1^N, \hat{w}_1^N)
\]
(9)

where \( \hat{w}_1^N \) is the correct transcription of the input audio and
\( A(y, y') \) is the accuracy of a token sequence \( y \), treating \( y' \) as
the ground-truth. In practice it is not feasible to compute this
sum exactly. Therefore, we use n-best lists to approximate the
search space. The probability \( p \) is being renormalized to this
n-best list.

When training the subword dependent scale parameters
introduced in Section 2.2 we use the same training criteria,
replacing the single scales with the subword dependent scale
parameters.

4. EXPERIMENTAL SETUP
For all of our experiments we use the RETURNN training
framework [14, 15]. Configs are available online [16]. We eval-
uate our methods on the LibriSpeech 960h [10] and Switch-
board 300h [17] corpora.

Our acoustic models are attention-based encoder-decoder
models with CNN+BLSTM encoder and a single layer LSTM
decoder that predict subword units generated by byte-pair-
encoding (BPE) (LBS:10025, SWB:534). The details of our
LibriSpeech Model can be found in [18] and our Switchboard
Model follows [8].

As our language model for LibriSpeech we use a 4 layer
LSTM based model with 140M parameters, the SWB LM is
a 6 Layer transformer with 76M parameters, both trained on
additional text data. In the experiments with joint training
(Section 5.3) we also use a single layer LSTM LM, which
has the same size as the AM decoder. This LM was trained
solely on the transcriptions of the LibriSpeech corpus.
Table 1: Performance of learned subword agnostic and subword dependent scales trained on LibriSpeech using different training criteria and subsets of data, where \( \alpha \) is the (average) AM scale and \( \beta \) is the (average) LM scale.

| scale training | dev WER [%] | test WER [%] |
|----------------|-------------|--------------|
| criterion set  | clean       | other        | clean       | other        |
| baseline + LM  | -           | -            | -           | -            |
| manual         | -           | -            | -           | -            |
| CE             | train       | dev-clean    | 2.77        | 1.00         |
|                |             | dev-other    | 1.00        | 0.50         |
|                | train       | dev-clean    | 1.24        | 0.26         |
|                |             | dev-other    | 1.10        | 0.51         |
|                | train       | dev-clean    | 3.41        | 1.19         |
|                |             | dev-other    | 3.14        | 0.89         |
|                | train       | dev-clean    | 3.82        | 1.46         |
|                |             | dev-other    | 1.50        | 0.63         |
|                | train       | dev-clean    | 1.49        | 0.63         |
|                |             | dev-other    | 1.50        | 0.64         |
| CE             | train       | dev-clean    | 1.50        | 0.64         |
|                |             | dev-other    | 1.50        | 0.64         |
| minWER          | train       | dev-clean    | 3.14        | 0.89         |
|                |             | dev-other    | 2.82        | 1.46         |
|                | train       | dev-clean    | 3.41        | 1.19         |
|                |             | dev-other    | 3.14        | 0.89         |
|                | train       | dev-clean    | 1.50        | 0.63         |
|                |             | dev-other    | 1.50        | 0.64         |
|                | train       | dev-clean    | 1.50        | 0.63         |
|                |             | dev-other    | 1.50        | 0.64         |
| minWER          | train       | dev-clean    | 1.50        | 0.63         |
|                |             | dev-other    | 1.50        | 0.64         |

4.1. Training Procedure

As in the standard shallow fusion approach, we start by first training both AM and LM separately with the CE objective function. Afterwards, we combine them and initialize the scales randomly with mean 1.0 and small variance. Then we train only the introduced scale parameters while keeping the model parameters fixed.

We investigate the joint training of the parameters by first training the scale parameters, AM and LM as described above and then running a joint training phase. Here, we decided to still keep the LM parameters fixed, as prior investigations have shown severe degradation when training the LM only on transcribed audio data.

5. RESULTS

5.1. Subword Agnostic Scale Training

We conduct experiments for subword agnostic scales, that is, one scale parameter for AM and the LM each. We train the scales for 5 epochs on the train dataset or for 100 epochs on the dev sets with a conservative learning rate. It is likely possible to reduce the number of epochs.

For LibriSpeech, the results of the experiments for both presented training criteria and datasets are displayed in Table 1. Firstly, we observe that the scales learned with the cross entropy training criterion produce slightly worse results than the manually tuned ones. The results obtained from the scales learned with the minWER criterion reach the same performance as those obtained from manual tuning. For Switchboard, the results are presented in Table 3. Here, the manual results can be found both by CE and minWER training.

In both cases, using the train set to estimate the scales seems to be more stable and to generalize better to the other test sets. This shows that our procedure can be used to automatically learn the scales for the shallow fusion method.

5.2. Subword Dependent Scale Training

We also learn subword dependent scales, that is one AM scale and one LM scale per subword unit, by training for 5 epochs on the train dataset or for 100 epochs on the dev sets. Additionally, for the minWER training criterion we initialize them with the scales obtained from the CE training step.

The results on LibriSpeech for both the CE and the minWER criterion as well as the different training sets are displayed in Table 1. For the CE criterion we observe a clear improvement of 5.5 % over the shallow fusion baseline on test-other when using the training dataset. When using the minWER criterion, the results are even better. For the training set we reach a relative improvement of 6.6 % on the test-other dataset. When training on the dev sets, we see a much bigger improvement on the set we use for estimating the parameters but reduces generalization of the model.

The results for Switchboard are presented in Table 3. We see that subword dependend scales acheive the same performance as the manual baseline on Hub5, but show better generalization on RT03. No overfitting on Hub5’00 is observed when using it to tune the scales. Training with minWER criterion on the train set leads to the best results.

In both cases training on the dev sets is not stable enough to reach the baseline results.

5.3. Joint Training

For joint training we initialize the model with a pretrained AM, LM, and scale parameters as presented above. Afterwards, we continue fine tuning the AM parameters while keeping the LM parameters fixed. We run experiments for
Table 2: Joint training results for subword agnostic and subword dependent scales trained on LibriSpeech with the CE criterion.

|                      | train AM scales | dev WER [%] | test WER [%] |
|----------------------|-----------------|-------------|--------------|
|                      |                 | clean       | other        | clean       | other       |
| subword agnostic     | no              | 3.0         | 8.7          | 3.2         | 9.3         |
|                      | yes             | 2.6         | 7.5          | 2.8         | 8.0         |
|                      | yes             | 2.6         | 7.6          | 2.8         | 8.2         |
| subword dependent    | no              | 2.8         | 8.0          | 3.1         | 8.5         |
|                      | yes             | 2.6         | 7.5          | 2.8         | 8.0         |

Both fixed as well as trainable scale parameters. In the joint training phase we use the decoder sized LM mentioned in section 4 since it better matches the train set. After the joint training we replace this LM for the usual more powerful one and retune the scales again. We use the CE training criterion from Equation (8) in all training steps.

We run the training on LibriSpeech for 5 epochs in each individual training step using the cross entropy criterion. The results of the experiments are displayed in Table 2. We observe that the joint training phase yields clear improvements in WER of 13.8% relative on dev-other for subword agnostic scales and 6.3 for subword dependent scales. We further observe that fixing the scale parameters during the joint training phase or continuing to tune them has a negligible effect.

The results of subword agnostic and subword dependent scales after running the joint training phase are almost identical. This indicates that the advantage that is achieved by the subword dependent scales can also be learned by the AM.

5.4. Subword Dependent Scales Analysis

We analyse the scales found by training with the minWER training criterion on the train dataset of LibriSpeech. In Figure 1a we observe that the distribution of the scales follow a Gaussian distribution. When examining the correlation between AM and LM scales for each subword (cf. Figure 1b) we find a Pearson coefficient of $-0.52$. This implies a slight inversely proportional correlation. According to Figure 1c the relative importance of AM and LM seems to be independent of the length of the BPE token in characters.

Table 3: Performance of learned subword agnostic and subword dependent scales trained on Switchboard using different training criteria and subsets of data, where $\alpha$ is the (average) AM scale and $\beta$ is the (average) LM scale.

| scale training | WER [%] | WER [%] | WER [%] |
|---------------|---------|---------|---------|
| criterion     | Hub5’00 | Hub5’01 | RT03    |
| baseline + LM | -       | manual  | 12.3    | 11.9    | 14.3    |
|               |         |         | 12.1    | 11.7    | 14.1    |
| subword       |         |         |         |         |         |
| agnostic scales | CE     | train Hub5’00 | 12.1 | 11.6 | 13.8    |
|                | minWER  | train Hub5’00 | 12.1 | 12.5 | 14.8    |
| subword       |         |         |         |         |         |
| dependent scales | CE     | train Hub5’00 | 12.1 | 11.6 | 13.8    |
|                | minWER  | train Hub5’00 | 12.0 | 11.5 | 13.7    |

6. CONCLUSION

In this work we proposed the automation of the tuning process for model scales used when combining different models via log-linear combination. We conducted experiments with both cross entropy and minimum word error rate training criterion on LibriSpeech and Switchboard and could recover the result of manual scale tuning. Training the scales on the whole training data showed better generalization of the scales to other test sets.

Additionally, we also extended the simple model scales by introducing individual scale parameters for each BPE subword unit. By doing this, we achieved a clear improvement of 6.6% relative WER reduction on the LibriSpeech test-other dataset and 2.8% improvement on the RT03 test set.

7. ACKNOWLEDGEMENTS

This work has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 694537, project "SEQCLAS"). The work reflects only the authors’ views and none of the funding parties is responsible for any use that may be made of the information it contains.
8. REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *Int. Conf. on Learning Representations (ICLR)*, May 2015.

[2] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2016, pp. 4960–4964.

[3] Caglar Gülcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, “On Using Monolingual Corpora in Neural Machine Translation,” in *Computer Speech & Language*, September 2017, vol. 45, pp. 137–148.

[4] Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N. Sainath, and Karen Livescu, “A Comparison of Techniques for Language Model Integration in Encoder-Decoder Speech Recognition,” in *Proc IEEE Spoken Language Technology Workshop (SLT)*, December 2018, pp. 369–375.

[5] Wilfried Michel, Ralf Schlüter, and Hermann Ney, “Early Stage LM Integration Using Local and Global Log-Linear Combination,” in *Proc. Interspeech*, October 2020, pp. 3605–3609.

[6] Alex Graves, “Sequence Transduction with Recurrent Neural Networks,” in *Representation Learning Workshop, Int. Conf. on Machine Learning (ICML)*, June 2012.

[7] Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley, “Hybrid Autoregressive Transducer (HAT),” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6139–6143.

[8] Mohammad Zeineldeen, Aleksandr Glushko, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney, “Investigating Methods to improve Language Model Integration for Attention-based Encoder-Decoder ASR Models,” in *Proc. Interspeech*, August 2021, pp. 2856–2860.

[9] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[10] Diederik P. Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, May 2015.

[11] Björn Hoffmeister, Ruoying Liang, Ralf Schlüter, and Hermann Ney, “Log-linear model combination with word-dependent scaling factors,” in *Proc. Interspeech*, September 2009, pp. 248–251.

[12] Rohit Prabhavalkar, Tara N. Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuji Kannan, “Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4839–4843.

[13] Jinxi Guo, Gautam Tiwari, Jasha Droppo, Maarten Van Segbroeck, Che-Wei Huang, Andreas Stolcke, and Roland Maas, “Efficient minimum word error rate training of RNN-Transducer for end-to-end speech recognition,” in *Proc. Interspeech*, October 2020, pp. 2807–2811.

[14] Patrick Doetsch, Albert Zeyer, Paul Voigtlaender, Ilya Kulikov, Ralf Schlüter, and Hermann Ney, “RE-TURNN: The RWTH Extensible Training framework for Universal Recurrent Neural Networks,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5345–5349.

[15] Albert Zeyer, Tamer Alkhouri, and Hermann Ney, “RE-TURNN as a Generic Flexible Neural Toolkit with Application to Translation and Speech Recognition,” in *Proc. Assoc. for Computational Linguistics (ACL)*, July 2018.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[17] J.J. Godfrey, E.C. Holliman, and J. McDaniel, “Switchboard: Telephone Speech Corpus for Research and Development,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1992, pp. 517–520.

[18] Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schluter, and Hermann Ney, “A Comparison of Transformer and LSTM Encoder Decoder Models for ASR,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 2019, pp. 8–15.