Interspeech 2022
18-22 September 2022, Incheon, Korea

Automatic Learning of Subword Dependent Model Scales

Felix Meyer[1], *Wilfried Michel*[1,2], *Mohammad Zeineldeen*[1,2], *Ralf Schlüter*[1,2], *Hermann Ney*[1,2]

[1]Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52074 Aachen, Germany
[2]AppTek GmbH, 52062 Aachen, Germany

felix.sebastian.meyer@rwth-aachen.de, {michel,zeineldeen,schlueter,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 use individual scaling parameters per subword output token. We train these parameters via automatic differentiation and stochastic gradient decent optimization similar to the neural network model parameters.

We show on the LibriSpeech (LBS) and Switchboard (SWB) corpora that automatic learning of two scales for a combination of attention-based encoder-decoder acoustic model and language model can be done as effectively as with manual tuning. Using subword dependent model scales which could not be tuned manually we achieve 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 [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 [3]. Previous investigations have shown that shallow fusion yields good results in most learning scenarios [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 [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 [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 [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. When more models are combined, the dimension of the search grid grows exponentially and finding optimal scales by hand quickly becomes infeasibly hard.

In this work we introduce a method to learn these scale parameters via automatic differentiation [9] and stochastic gradient decent [10]. This opens up the possibility to efficiently find an arbitrary number of combination parameters and models. We use this to investigate the use of individual scale parameters per subword output token, similar to [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 $p_{\mathrm{AM}}\left(w_1^N \mid x_1^T\right)$, and the LM $p_{\mathrm{LM}}(w_1^N)$ have to be combined. To obtain 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:

$$\widetilde{p}\left(w_1^N \middle| x_1^T\right) = \frac{p_{\mathrm{AM}}^{\alpha}(w_1^N|x_1^T)\cdot p_{\mathrm{LM}}^{\beta}(w_1^N)}{\sum_{\widetilde{w}_1^N} p_{\mathrm{AM}}^{\alpha}(\widetilde{w}_1^N|x_1^T)\cdot p_{\mathrm{LM}}^{\beta}(\widetilde{w}_1^N)} \tag{1}$$

where α and $\beta \in \mathbb{R}$ 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 normalization 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:

$$\widetilde{w}_1^{\widetilde{N}} = \underset{N,w_1^N}{\arg\max}\left\{\log p_{\mathrm{AM}}\left(w_1^N \middle| x_1^T\right) + \beta\log p_{\mathrm{LM}}(w_1^N)\right\} \tag{2}$$

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

10.21437/Interspeech.2022-10136

The normalization can also be done on a per token basis:

$$\widehat{p}\left(w_1^N \middle| x_1^T\right) = \prod_{n=1}^{N} \widehat{p}_n\left(w_n \middle| w_1^{n-1}, x_1^T\right) \tag{3}$$

$$= \prod_{n=1}^{N} \frac{p_{\mathrm{AM}}^{\alpha}(w_n|w_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta}(w_n|w_1^{n-1})}{\sum_{\widetilde{w}} p_{\mathrm{AM}}^{\alpha}(\widetilde{w}|w_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta}(\widetilde{w}|w_1^{n-1})} \tag{4}$$

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 individual 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 subword units and k is the total number of subword units. The definition for the sentence level normalized prediction probability changes to

$$\widetilde{p}\left(w_1^N \middle| x_1^T\right) = \frac{\prod_{n=1}^{N} p_{\mathrm{AM}}^{\alpha_{w_n}}(w_n|w_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta_{w_n}}(w_n|w_1^{n-1})}{\sum_{\widetilde{w}_1^N} \prod_{n=1}^{N} p_{\mathrm{AM}}^{\alpha_{\widetilde{w}_n}}(\widetilde{w}_n|\widetilde{w}_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta_{\widetilde{w}_n}}(\widetilde{w}_n|\widetilde{w}_1^{n-1})} \tag{5}$$

and the token level normalized probability to

$$\widehat{p}(w_1^N|x_1^T) = \prod_{n=1}^{N} \frac{p_{\mathrm{AM}}^{\alpha_{w_n}}(w_n|w_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta_{w_n}}(w_n|w_1^{n-1})}{\sum_{\widetilde{w}} p_{\mathrm{AM}}^{\alpha_{\widetilde{w}}}(\widetilde{w}|w_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta_{\widetilde{w}}}(\widetilde{w}|w_1^{n-1})}. \tag{6}$$

3. Learning of Model Scales

Usually, the scale parameters from the decision rule of log-linear combination α and β are tuned manually. This is commonly done via grid search, running the decoding process for different scale parameters. In this work, we propose to learn these parameters automatically. We use automatic differentiation [9] of a training criterion F and a variant of stochastic gradient descent [10] to find the optimal scale values similar to how other model parameters are optimized.

3.1. Training Criteria

To train the scales we have to define a suitable training criterion. 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

$$F_{\mathrm{CE}} = \log \widehat{p}\left(w_1^N \middle| x_1^T\right) \tag{7}$$

$$= \sum_{n=1}^{N} \log \frac{p_{\mathrm{AM}}^{\alpha}(w_n|w_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta}(w_n|w_1^{n-1})}{\sum_{\widetilde{w}} p_{\mathrm{AM}}^{\alpha}(\widetilde{w}|w_1^{n-1}, x_1^T) \cdot p_{\mathrm{LM}}^{\beta}(\widetilde{w}|w_1^{n-1})}. \tag{8}$$

This criterion, however, does 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]. This training criterion uses the sentence level renormalization and is given by

$$F_{\mathrm{minWER}} = \sum_{N, w_1^N} \widetilde{p}\left(w_1^N \middle| x_1^T\right) \cdot \mathcal{A}\left(w_1^N, \widetilde{w}_1^{\widetilde{N}}\right) \tag{9}$$

where $\widetilde{w}_1^{\widetilde{N}}$ is the correct transcription of the input audio and $\mathcal{A}\left(y, y'\right)$ 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. Also $\widetilde{p}$ 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.[1] We evaluate our methods on the LibriSpeech 960h and Switchboard 300h 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 [16] 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.

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 [5] 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 or 100 epochs on the dev sets.

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

[1] https://github.com/rwth-i6/returnn-experiments/tree/master/2022-scale-learning

Table 1: *Performance of learned subword agnostic and subword dependent scales trained on LibriSpeech using different training criteria and subsets of data, where α is the (average) AM scale and β is the (average) LM scale*

	scale training		α	β	$\frac{\beta}{\alpha}$	dev WER [%]		test WER [%]	
	criterion	set				clean	other	clean	other
baseline	-	-	-	-	-	4.0	10.9	4.2	11.4
+ LM	manual	dev-other	2.77	1.00	0.36	2.9	8.3	3.2	9.0
subword agnostic scales	CE	train	1.24	0.26	0.21	3.0	8.7	3.2	9.3
		dev-clean	1.10	0.51	0.46	3.1	8.4	3.5	9.3
		dev-other	1.00	0.50	0.50	3.1	8.5	3.6	9.7
	minWER	train	3.41	1.19	0.35	2.8	8.2	3.2	8.9
		dev-clean	3.14	0.89	0.28	2.9	8.3	3.2	9.1
		dev-other	2.82	1.46	0.52	3.1	8.3	3.5	9.2
subword dependent scales	CE	train	1.49	0.63	0.45	2.8	8.0	3.1	8.5
		dev-clean	1.50	1.13	0.84	6.0	17.7	12.2	20.7
		dev-other	1.40	1.36	0.90	8.5	8.1	9.1	15.1
	minWER	train	1.50	0.63	0.45	2.7	7.8	**3.0**	**8.4**
		dev-clean	1.50	0.64	0.45	2.5	7.9	3.1	8.5
		dev-other	1.50	0.64	0.45	2.7	7.4	3.1	8.5

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 2. 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 reduced generalization of the model on other sets.

The results for Switchboard are presented in Table 2. We see that subword dependent scales achieve 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 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 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.

Table 2: *Performance of learned subword agnostic and subword dependent scales trained on Switchboard using different training criteria and subsets of data*

	scale training		WER [%]		
	criterion	set	Hub5'00	Hub5'01	RT03
baseline	-	-	12.3	11.9	14.3
+ LM	manual	Hub5'00	12.1	11.7	14.1
subword agnostic scales	CE	train	12.1	11.6	13.8
		Hub5'00	13.5	12.5	14.8
	minWER	train	12.1	11.7	13.8
		Hub5'00	12.1	11.6	13.8
subword dependent scales	CE	train	12.1	11.7	13.9
		Hub5'00	20.9	22.8	22.0
	minWER	train	**12.0**	**11.5**	**13.7**
		Hub5'00	14.0	14.5	16.5

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 3. 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.

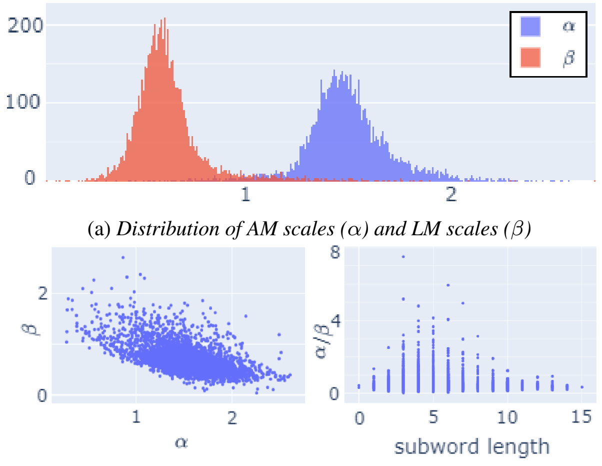
The results of subword agnostic and subword dependent scales after running the joint training phase are almost identical. This indicates that the advantage that is achieved by the subword dependent scales can also be learned by the AM.

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 with means $\bar{\alpha} = 1.50$ and $\bar{\beta} = 0.63$. 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. Accord-

Table 3: *Joint training results for subword agnostic and subword dependent scales trained on LibriSpeech with the CE criterion.*

	train		dev WER [%]		test WER [%]	
	AM	scales	clean	other	clean	other
subword agnostic	no	yes	3.0	8.7	3.2	9.3
	yes	no	**2.6**	**7.5**	**2.8**	8.0
		yes	**2.6**	7.6	**2.8**	8.2
subword dependent	no	yes	2.8	8.0	3.1	8.5
	yes	no	**2.6**	**7.5**	**2.8**	8.0
		yes	**2.6**	**7.5**	**2.8**	**7.9**



(a) *Distribution of AM scales (α) and LM scales (β)*

(b) *Correlation of α and β*

(c) *Length distribution*

Figure 1: *Analysis of subword dependent AM scales α and LM scales β trained on LibriSpeech 960h with minWER criterion*

ing to Figure 1c the relative importance of AM and LM seems to be independent of the length of the BPE token in characters.

6. Conclusion

In this work we proposed to use subword dependent model scales for the log-linear combination of an attention-based encoder-decoder acoustic model and a neural language model. To this end, we automated the tuning process for model scales by using automatic differentiation and gradient-based updates.

We conducted experiments with both cross entropy and minimum word error rate training criteria on LibriSpeech and Switchboard. Using subword agnostic scales and the minWER criterion, we recovered 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, when using subword dependent scales, 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. By training acoustic model parameters jointly with the model scales we could increase the relative improvement to 11.1% on test-other.

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] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *Int. Conf on Learning Representations (ICLR)*, May 2015.

[2] W. Chan, N. Jaitly, Q. Le, and O. 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] C. Gülcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On Using Monolingual Corpora in Neural Machine Translation," in *Computer Speech & Language*, vol. 45, September 2017, pp. 137–148.

[4] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. 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] W. Michel, R. Schlüter, and H. Ney, "Early Stage LM Integration Using Local and Global Log-Linear Combination," in *Proc. Interspeech*, October 2020, pp. 3605–3609.

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

[7] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid Autoregressive Transducer (HAT)," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6139–6143.

[8] M. Zeineldeen, A. Glushko, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Investigating Methods to improve Language Model Integration for Attention-based Encoder-Decoder ASR Models," in *Proc. Interspeech*, August 2021, pp. 2856–2860.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

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

[11] B. Hoffmeister, R. Liang, R. Schlüter, and H. Ney, "Log-linear model combination with word-dependent scaling factors," in *Proc. Interspeech*, September 2009, pp. 248–251.

[12] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C. Chiu, and A. 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] J. Guo, G. Tiwari, J. Droppo, M. V. Segbroeck, C.-W. Huang, A. Stolcke, and R. 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] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, and H. Ney, "RETURNN: 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] A. Zeyer, T. Alkhouli, and H. Ney, "RETURNN as a Generic Flexible Neural Toolkit with Application to Translation and Speech Recognition," in *Proc. Assoc. for Computational Linguistics (ACL)*, July 2018.

[16] A. Zeyer, P. Bahar, K. Irie, R. Schluter, and H. 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.