

IMPROVING END-TO-END SPEECH RECOGNITION WITH POLICY LEARNING

Yingbo Zhou, Caiming Xiong, Richard Socher

Salesforce Research

ABSTRACT

Connectionist temporal classification (CTC) is widely used for maximum likelihood learning in end-to-end speech recognition models. However, there is usually a disparity between the negative maximum likelihood and the performance metric used in speech recognition, *e.g.*, word error rate (WER). This results in a mismatch between the objective function and metric during training. We show that the above problem can be mitigated by jointly training with maximum likelihood and policy gradient. In particular, with policy learning we are able to directly optimize on the (otherwise non-differentiable) performance metric. We show that joint training improves relative performance by 4% to 13% for our end-to-end model as compared to the same model learned through maximum likelihood. The model achieves 5.53% WER on Wall Street Journal dataset, and 5.42% and 14.70% on Librispeech test-clean and test-other set, respectively.

Index Terms— end-to-end speech recognition, LVCSR, policy gradient, deep neural networks

1. INTRODUCTION

Deep neural networks are the basis for some of the most accurate speech recognition systems in research and production [1, 2, 3]. Neural network based acoustic models are commonly used as a sub-component in a Gaussian mixture model (GMM) and hidden Markov model (HMM) based hybrid system. Alignment is necessary to train the acoustic model, and a two-stage (*i.e.* alignment and frame prediction) training process is required for a typical hybrid system. A drawback of such setting is that there is a disconnect between the acoustic model training and the final objective, which makes the system level optimization difficult.

The end-to-end neural network based speech models bypass this two-stage training process by directly maximizing the likelihood of the data. More recently, the end-to-end models have also shown promising results on various datasets [4, 5, 6, 7]. While the end-to-end models are commonly trained with maximum likelihood, the final performance metric for a speech recognition system is typically word error rate (WER) or character error rate (CER). This results a mismatch between the objective that is optimized and the evaluation metric. In an ideal setting the model should be trained to optimize

the final metric. However, since the metrics are commonly discrete and non-differentiable, it is very difficult to optimize them in practice.

Lately, reinforcement learning (RL) has shown to be effective on improving performance for problems that have non-differentiable metric through policy gradient. Promising results are obtained in machine translation [8, 9], image captioning [8, 10], summarization [8, 11], *etc.*. In particular, REINFORCE algorithm [12] enables one to estimate the gradient of the expected reward by sampling from the model. It has also been applied for online speech recognition [13]. Graves and Jaitly [4] propose expected transcription loss that can be used to optimize on WER. However, it is more computationally expensive. For example, for a sequence of length T with vocabulary size K , at least T samples and K metric calculations are required for estimating the loss.

We show that jointly training end-to-end models with self critical sequence training (SCST) [10] and maximum likelihood improves performance significantly. SCST is also efficient during training, as only one sampling process and two metric calculations are necessary. Our model achieves 5.53% WER on Wall Street Journal dataset, and 5.42% and 14.70% WER on Librispeech test-clean and test-other sets.

2. MODEL STRUCTURE

The end-to-end model structure used in this work is very similar to that of Deep Speech 2 (DS2) [6]. It is mainly composed of 1) a stack of convolution layers in the front-end for feature extraction, and 2) a stack of recurrent layers for sequence modeling. The structure of recurrent layers is the same as in DS2, and we illustrate the modifications in convolution layers in this section.

We choose to use time and frequency convolution (*i.e.* 2-D convolution) as the front-end of our model, since it is able to model both the temporal transitions and spectral variations in speech utterances. We use depth-wise separable convolution [14, 15] for all the convolution layers, due to its computational efficiency and performance advantage [15]. The depth-wise separable convolution is implemented by first convolving over the input channel-wise, and then convolve with 1×1 filters with the desired number of output channels. Stride size only influences the channel-wise convolution; the following 1×1 convolutions always have stride size of one. More pre-

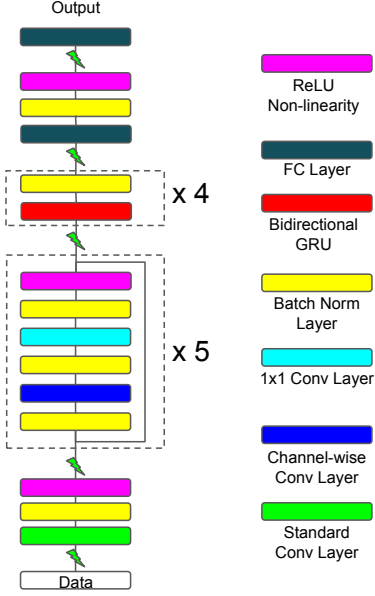


Fig. 1. Model architecture of our end-to-end speech model. Different colored blocks represent different layers as shown on the right, the lightning symbol indicates dropout happens between the two layers.

cisely, let $\mathbf{x} \in \mathbb{R}^{F \times T \times D}$, $\mathbf{c} \in \mathbb{R}^{W \times H \times D}$ and $\mathbf{w} \in \mathbb{R}^{D \times N}$ denote an input sample, the channel-wise convolution and the 1×1 convolution weights respectively. The depth-wise separable convolution with D input channels and N output channels performs the following operations:

$$s(i, j, d) = \sum_{f=0}^{F-1} \sum_{t=0}^{T-1} x(f, t, d) c(i-f, j-t, d) \quad (1)$$

$$o(i, j, n) = \sum_{k=0}^{D-1} s(i, j, k) w(k, n) \quad (2)$$

where $d \in \{1, \dots, D\}$ and $n \in \{1, 2, \dots, N\}$, \mathbf{s} is the channel-wise convolution result, and \mathbf{o} is the result from depth-wise separable convolution. In addition, we add a residual connection [16] between the input and the layer output for the depth-wise separable convolution to facilitate training.

Our model is composed of six convolution layers – one standard convolution layer that has larger filter size, followed by five residual convolution blocks [16]. The convolution features are then fed to four bidirectional gated recurrent units (GRU) [17] layers, and finally two fully connected layers that make the final per-character prediction. The full end-to-end model structure is illustrated in Fig. 1.

3. MODEL OBJECTIVE

3.1. Maximum Likelihood Training

Connectionist temporal classification (CTC) [18] is a popular method for doing maximum likelihood training on sequence labeling tasks, where the alignment information is not provided in the label. The alignment is not required since CTC marginalizes over all possible alignments, and maximizes the likelihood $P(\mathbf{y}|\mathbf{x})$. It achieves this by augmenting the original label set \mathcal{L} to set $\Omega = \mathcal{L} \cup \{\text{blank}\}$ with an additional blank symbol. A mapping \mathcal{M} is then defined to map a length T sequence of label Ω^T to $\mathcal{L}^{\leq T}$ by removing all blanks and repeated symbols along the path. The likelihood can then be recovered by

$$P(\mathbf{y}'|\mathbf{x}) = \prod_t P(y'_t|\mathbf{x}), y'_t \in \Omega^T$$

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{y}' \in \mathcal{M}^{-1}(\mathbf{y})} P(\mathbf{y}'|\mathbf{x})$$

where \mathbf{x} , \mathbf{y} and \mathbf{y}' denote an input example of length T , the corresponding label of length $\leq T$ and one of the augmented label with length T .

3.2. Policy Learning

The log likelihood reflects the log probability of getting the whole transcription completely correct. What it ignores are the probabilities of the incorrect transcriptions. In other words, all incorrect transcriptions are equally bad, which is clearly not the case. Furthermore, the performance metrics typically aim to reflect the plausibility of incorrect predictions. For example, WER penalizes less for transcription that has less edit distance to the ground truth label. This results in a disparity between the optimization objective of the model and the (commonly discrete) evaluation criteria. This mismatch is mainly attributed to the inability to directly optimize the criteria.

One way to remedy this mismatch is to view the above problem in the policy learning framework. In this framework, we can view our model as an *agent* and the training samples as the *environment*. The parameters of the model θ defines a policy $P_\theta(\mathbf{y}|\mathbf{x})$, the model interacts with the environment by following this policy. The agent then performs an *action* based on its current *state*, in which case the action is the generated transcription and the state is the model hidden representation of the data. It then observes a *reward* that is defined from the evaluation metric calculated on the current sample (*e.g.* 1–WER for the current transcription). The goal of learning is to obtain a policy that minimizes the negative expected reward:

$$L_p(\theta) = -\mathbb{E}_{\mathbf{y}^s \sim P_\theta(\mathbf{y}|\mathbf{x})} [r(\mathbf{y}^s)] \quad (3)$$

where $r(\cdot)$ denotes the reward function. Gradient of eq. 3 can

be obtained through REINFORCE [12] as

$$\nabla_{\theta} L_p(\theta) = -\mathbb{E}_{\mathbf{y}^s \sim P_{\theta}(\mathbf{y}|\mathbf{x})} [r(\mathbf{y}^s) \nabla_{\theta} \log P_{\theta}(\mathbf{y}^s|\mathbf{x})] \quad (4)$$

$$\approx -r(\mathbf{y}^s) \nabla_{\theta} \log P_{\theta}(\mathbf{y}^s|\mathbf{x}) \quad (5)$$

Eq. 5 shows the Monte Carlo approximation of the gradient with a single example, which is a common practice when training model with stochastic gradient descent.

The policy gradient obtained from eq. 5 is often of high variance, and the training can get unstable. To reduce the variance, Rennie *et al.* [10] proposed self-critical sequence training (SCST). In SCST, the policy gradient is computed with a *baseline*, which is the greedy output from the model. Formally, the policy gradient is calculated using

$$\nabla_{\theta} L_p(\theta) = -\mathbb{E}_{\mathbf{y}^s \sim P_{\theta}(\mathbf{y}|\mathbf{x})} [(r(\mathbf{y}^s) - r(\hat{\mathbf{y}})) \nabla_{\theta} \log P_{\theta}(\mathbf{y}^s|\mathbf{x})] \quad (6)$$

$$\approx -(r(\mathbf{y}^s) - r(\hat{\mathbf{y}})) \nabla_{\theta} \log P_{\theta}(\mathbf{y}^s|\mathbf{x}) \quad (7)$$

where $\hat{\mathbf{y}}$ is the greedy decoding output from the model for the input sample \mathbf{x} .

3.3. Multi-objective Policy Learning

A potential problem with policy gradient methods (including SCST) is that the learning can be slow and unstable at the beginning of training. This is because it is unlikely for the model to have reasonable output at that stage, which leads to implausible samples with low rewards. Learning will be slow in case of small learning rate, and unstable otherwise. One way to remedy this problem is to incorporate maximum likelihood objective along with policy gradient, since in maximum likelihood the probability is evaluated on the ground truth targets, and hence will get large gradients when the model output is incorrect. This leads to the following objective for training our end-to-end speech model:

$$L(\theta) = -\log P_{\theta}(\mathbf{y}|\mathbf{x}) + \lambda L_{scst}(\theta) \quad \text{where} \quad (8)$$

$$L_{scst}(\theta) = -\{g(\mathbf{y}^s, \mathbf{y}) - g(\hat{\mathbf{y}}, \mathbf{y})\} \log P_{\theta}(\mathbf{y}^s|\mathbf{x})$$

where $g(\cdot, \cdot)$ is the reward function and $\lambda \in (0, +\infty)$ is the coefficient that controls the contribution from SCST. In our case we choose $g(\cdot, \mathbf{y}) = 1 - \max(1, \text{WER}(\cdot, \mathbf{y}))$. Training with eq. 8 is also efficient, since both sampling and greedy decoding is cheap. The only place that might be computationally more demanding is the reward calculation, however, we only need to compute it twice per batch of examples, which adds only a minimal overhead.

4. EXPERIMENTS

We evaluate the proposed objective by performing experiments on the Wall Street Journal (WSJ) and LibriSpeech [19] datasets. The input to the model is a spectrogram computed with a 20ms window and 10ms step size. We first normalize

each spectrogram to have zero mean and unit variance. In addition, we also normalize each feature to have zero mean and unit variance based on the training set statistics. No further preprocessing is done after these two steps of normalization.

We denote the size of the convolution layer by the tuple (C, F, T, SF, ST), where C, F, T, SF, and ST denote number of channels, filter size in frequency dimension, filter size in time dimension, stride in frequency dimension and stride in time dimension respectively. We have one convolutional layer with size (32,41,11,2,2), and five residual convolution blocks of size (32,7,3,1,1), (32,5,3,1,1), (32,3,3,1,1), (64,3,3,2,1), (64,3,3,1,1) respectively. Following the convolutional layers we have 4 layers of bidirectional GRU RNNs with 1024 hidden units per direction per layer. Finally, we have one fully connected hidden layer of size 1024 followed by the output layer. Batch normalization [20] is applied to all layers' pre-activations to facilitate training. Dropout [21] is applied to inputs of each layer, and for layers that take sequential input (*i.e.* the convolution and recurrent layers) we use the dropout variant proposed by Gal and Ghahramani [22]. The convolutional and fully connected layers are initialized uniformly following He *et al.* [23]. The recurrent layer weights are initialized with a uniform distribution $\mathcal{U}(-1/32, 1/32)$. The model is trained in an end-to-end fashion to minimize the mixed objective as illustrated in eq. 8. We use mini-batch stochastic gradient descent with batch size 64, learning rate 0.1, and with Nesterov momentum 0.95. The learning rate is reduced by half whenever the validation loss has plateaued. We set $\lambda = 0.1$ at the beginning of training, and increase it to 1 after the model has converged (*i.e.* the validation loss stops improving). The gradient is clipped [24] to have a maximum ℓ_2 norm of 1. For regularization, we use ℓ_2 weight decay of 10^{-5} for all parameters. Additionally, we apply dropout for inputs of each layer (see Fig. 1). The dropout probabilities are set as 0.1 for data, 0.2 for all convolution layers, and 0.3 for all recurrent and fully connected layers. Furthermore, we also augment the audio training data through random perturbations of tempo, pitch, volume, temporal alignment, along with adding random noise.

4.1. Effect of Policy Learning

To study the effectiveness of our multi-objective policy learning, we perform experiments on both datasets with various settings. The first set of experiments was carried out on the WSJ corpus. We use the standard *si284* set for training, *dev93* for validation and *eval92* for test evaluation. We use the provided language model and report the result in the 20K closed vocabulary setting with beam search. The beam width is set to 100. Results are shown in table 1. Both policy gradient methods improve results over baseline. In particular, the use of SCST results in 13.8% relative performance improvement on the *eval92* set over the baseline.

On LibriSpeech dataset, the model is trained using all 960

Method	dev93		eval92	
	CER	WER	CER	WER
Baseline	4.07%	9.93%	2.59%	6.42%
Policy (eq. 5)	3.71%	9.46%	2.31%	5.85%
Policy (eq. 7)	3.52%	9.21%	2.10%	5.53%

Table 1. Performance from WSJ dataset. Baseline denotes model trained without CTC only; policy indicates model trained using the multi-objective policy learning. Equation in parenthesis indicates the way used to obtain policy gradient.

Dataset		Baseline	Policy
dev-clean	CER	1.76%	1.69%
	WER	5.33%	5.10%
test-clean	CER	1.87%	1.75%
	WER	5.67%	5.42%
dev-other	CER	6.60%	6.26%
	WER	14.88%	14.26%
test-other	CER	6.58%	6.25%
	WER	15.18%	14.70%

Table 2. Performance from LibriSpeech dataset. Policy denotes model trained with multi-objective shown in eq. 8.

Method	WER
Hannun <i>et al.</i> [25]	14.10 %
Bahdanau <i>et al.</i> [7]	9.30%
Graves and Jaitly [4]	8.20%
Wu <i>et al.</i> [26]	8.20%
Miao <i>et al.</i> [5]	7.34%
Chorowski and Jaitly [27]	6.70%
Human [6]	5.03%
Amodei <i>et al.</i> [6]*	3.60%
Ours	5.53%
Ours (LibriSpeech)	4.67%

Table 3. Comparative results with other end-to-end methods on WSJ *eval92* dataset. LibriSpeech denotes model trained using LibriSpeech dataset *only*, and test on WSJ. Amodei *et al.* used more training data.

hours of training data. Both dev-clean and dev-other are used for validation and results are reported in table 2. The provided 4-gram language model is used for final beam search decoding. The beam width is also set to 100 for decoding. Overall, a relative $\approx 4\%$ performance improvement over the baseline is observed.

Method	test-clean	test-other
Collobert <i>et al.</i> [28]	7.20%	-
Amodei <i>et al.</i> [6]*	5.33%	13.25%
ours	5.42%	14.70%

Table 4. Word error rate comparison with other end-to-end methods on LibriSpeech dataset. Amodei *et al.* used more training data.

4.2. Comparison with Other Methods

We also compare our performance with other end-to-end models. Comparative results from WSJ and LibriSpeech dataset are illustrated in tables 3 and 4 respectively. Our model achieved competitive performance with other methods on both datasets. In particular, with the help of policy learning we achieved similar results as Amodei *et al.* [6] on LibriSpeech without using additional data. To see if the model generalizes, we also tested our LibriSpeech model on the WSJ dataset. The result is significantly better than the model trained on WSJ data (see table 3), which suggests that the end-to-end models benefit more when more data is available.

5. CONCLUSION

In this work, we try to close the gap between the maximum likelihood training objective and the final performance metric for end-to-end speech models. We show this gap can be reduced by using the policy gradient method along with the negative log-likelihood. In particular, we apply a multi-objective training with SCST to reduce the expected negative reward that is defined by using the final metric. The joint training is computationally efficient. We show that the joint training is effective even with single sample approximation, which improves the relative performance on WSJ and LibriSpeech by 13% and 4% over the baseline.

6. REFERENCES

- [1] G Hinton, L Deng, D Yu, G E Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, T Sainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [2] G Saon, HK J Kuo, S Rennie, and M Picheny, “The ibm 2015 english conversational telephone speech recognition system,” *arXiv preprint arXiv:1505.05899*, 2015.
- [3] W Xiong, J Droppo, X Huang, F Seide, M Seltzer, A Stolcke, D Yu, and G Zweig, “The microsoft 2016

- conversational speech recognition system,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5255–5259.
- [4] A Graves and N Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *ICML*, 2014, pp. 1764–1772.
- [5] Y Miao, M Gowayyed, and F Metze, “Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding,” in *ASRU*. IEEE, 2015, pp. 167–174.
- [6] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *ICML*, 2016, pp. 173–182.
- [7] D Bahdanau, J Chorowski, D Serdyuk, P Brakel, and Y Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *ICASSP*. IEEE, 2016, pp. 4945–4949.
- [8] M Ranzato, S Chopra, M Auli, and W Zaremba, “Sequence level training with recurrent neural networks,” *arXiv preprint arXiv:1511.06732*, 2015.
- [9] D Bahdanau, P Brakel, K Xu, A Goyal, R Lowe, J Pineau, A Courville, and Y Bengio, “An actor-critic algorithm for sequence prediction,” *arXiv preprint arXiv:1607.07086*, 2016.
- [10] S J Rennie, E Marcheret, Y Mroueh, J Ross, and V Goel, “Self-critical sequence training for image captioning,” *arXiv preprint arXiv:1612.00563*, 2016.
- [11] R Paulus, C Xiong, and R Socher, “A deep reinforced model for abstractive summarization,” *arXiv preprint arXiv:1705.04304*, 2017.
- [12] R J Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.
- [13] Y Luo, C Chiu, N Jaitly, and I Sutskever, “Learning online alignments with continuous rewards policy gradient,” in *ICASSP*. IEEE, 2017, pp. 2801–2805.
- [14] L Sifre and S Mallat, “Rotation, scaling and deformation invariant scattering for texture discrimination,” in *CVPR*, 2013, pp. 1233–1240.
- [15] F Chollet, “Xception: Deep learning with depthwise separable convolutions,” *arXiv preprint arXiv:1610.02357*, 2016.
- [16] K He, X Zhang, S Ren, and J Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [17] K Cho, Bart Van M, D Bahdanau, and Y Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *arXiv preprint arXiv:1409.1259*, 2014.
- [18] A Graves, S Fernández, F Gomez, and J Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML*. ACM, 2006, pp. 369–376.
- [19] V Panayotov, G Chen, D Povey, and S Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *ICASSP*. IEEE, 2015, pp. 5206–5210.
- [20] S Ioffe and C Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *ICML*, Francis R. Bach and David M. Blei, Eds. 2015, vol. 37 of *JMLR Proceedings*, pp. 448–456, JMLR.org.
- [21] N Srivastava, G Hinton, A Krizhevsky, I Sutskever, and R Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [22] Y Gal and Z Ghahramani, “A theoretically grounded application of dropout in recurrent neural networks,” in *NIPS*, 2016, pp. 1019–1027.
- [23] K He, X Zhang, S Ren, and J Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
- [24] R Pascanu, T Mikolov, and Y Bengio, “On the difficulty of training recurrent neural networks,” in *ICML*, 2013, pp. 1310–1318.
- [25] A Hannun, A Maas, D Jurafsky, and A Ng, “First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns,” *arXiv preprint arXiv:1408.2873*, 2014.
- [26] Y Wu, S Zhang, Y Zhang, Y Bengio, and R Salakhutdinov, “On multiplicative integration with recurrent neural networks,” in *NIPS*, 2016, pp. 2856–2864.
- [27] J Chorowski and N Jaitly, “Towards better decoding and language model integration in sequence to sequence models,” *arXiv preprint arXiv:1612.02695*, 2016.
- [28] R Collobert, C Puhersch, and G Synnaeve, “Wav2letter: an end-to-end convnet-based speech recognition system,” *arXiv preprint arXiv:1609.03193*, 2016.