



Towards better decoding and language model integration in sequence to sequence models

Jan Chorowski¹, Navdeep Jaitly^{2*}

¹Google Brain, USA

²Nvidia, USA

jan.chorowski@cs.uni.wroc.pl, njaitly@nvidia.com

Abstract

The recently proposed Sequence-to-Sequence (seq2seq) framework advocates replacing complex data processing pipelines, such as an entire automatic speech recognition system, with a single neural network trained in an end-to-end fashion. In this contribution, we analyse an attention-based seq2seq speech recognition system that directly transcribes recordings into characters. We observe two shortcomings: overconfidence in its predictions and a tendency to produce incomplete transcriptions when language models are used. We propose practical solutions to both problems achieving competitive speaker independent word error rates on the Wall Street Journal dataset: without separate language models we reach 10.6% WER, while together with a trigram language model, we reach 6.7% WER, a state-of-the-art result for HMM-free methods.

Index Terms: attention mechanism, recurrent neural networks, LSTM

1. Introduction

Deep learning [1] has led to many breakthroughs including speech and image recognition [2, 3, 4, 5, 6, 7]. Especially, Sequence-to-Sequence (seq2seq) neural networks have proved to be very successful on complex transduction tasks, such as machine translation [8, 9, 10], speech recognition [11, 12, 13], or lip-reading [14]. Seq2seq models replace a full ASR pipeline and are directly trained to maximize the probability of observing desired outputs conditioned on the inputs. This discriminative training mode is fundamentally different from the generative "noisy channel" formulation used to build classical state-of-the-art speech recognition systems. As such, it has benefits and limitations that are different from classical ASR systems.

Understanding and preventing limitations specific to seq2seq models is crucial for their successful development. Discriminative training allows seq2seq models to focus on the most informative features. However, it also increases the risk of overfitting to those few distinguishing characteristics. We have observed that seq2seq models often yield very sharp predictions, and only a few hypotheses need to be considered to find the most likely transcription of a given utterance. However, high confidence reduces the diversity of transcripts obtained using beam search.

During typical training the models are conditioned on ground truth transcripts and are scored on one-step ahead predictions. By itself, this training criterion does not ensure that all relevant fragments of the input utterance are transcribed. Subsequently, mistakes that are introduced during decoding may cause the model to skip some words and jump to another place in the recording. The problem of incomplete transcripts is especially apparent when external language models are used.

* Work done at Google.

2. Model Description

Our speech recognition system builds on the recently proposed Listen, Attend and Spell network [13]. It is an attention-based seq2seq model that is able to directly transcribe an audio recording \mathbf{x} into a space-delimited sequence of characters \mathbf{y} . Similarly to other seq2seq neural networks, it uses an encoder-decoder architecture composed of three parts: a *listener* module tasked with acoustic modeling, a *speller* module tasked with emitting characters and an *attention* module serving as the intermediary between the speller and the listener:

$$\mathbf{h} = \text{Listen}(\mathbf{x}) \quad (1)$$

$$p(\mathbf{y}|\mathbf{x}) = \text{AttendAndSpell}(\mathbf{y}, \mathbf{h}) \quad (2)$$

2.1. The Listener

The listener is a multilayer Bi-LSTM network that transforms a sequence of N frames of acoustic features $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ into a possibly shorter sequence of hidden activations $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{N/k}$, where k is a time reduction constant [12, 13].

2.2. The Speller and the Attention Mechanism

The speller computes the probability of a sequence of characters conditioned on the activations of the listener. The probability is computed one character at a time, using the chain rule:

$$p(\mathbf{y}|\mathbf{h}) = \prod_i p(\mathbf{y}_i|\mathbf{y}_{<i}, \mathbf{h}). \quad (3)$$

To emit a character the speller uses the attention mechanism to find a set of relevant activations of the listener α and summarize them into a context \mathbf{c} . The history of previously emitted characters is encapsulated in a recurrent state \mathbf{s} :

$$\mathbf{s}_i = \text{RecurrentStep}(\mathbf{y}_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_{i-1}), \quad (4)$$

$$\mathbf{c}_i, \alpha_i = \text{Attend}(\mathbf{h}, \mathbf{s}_i, \alpha_{i-1}), \quad (5)$$

$$p(\mathbf{y}_i|\mathbf{y}_{<i}, \mathbf{h}) = \text{CharacterDistribution}(\mathbf{s}_i, \mathbf{c}_i). \quad (6)$$

We implement the recurrent step using a single LSTM layer. The attention mechanism is sensitive to the location of frames selected during the previous step and employs the convolutional filters over the previous attention weights [11]. The output character distribution is computed using a SoftMax function.

2.3. Training Criterion

Our speech recognizer computes the probability of a character conditioned on the partially emitted transcript and the whole utterance. It can thus be trained to minimize the cross-entropy between the ground-truth characters and model predictions. The

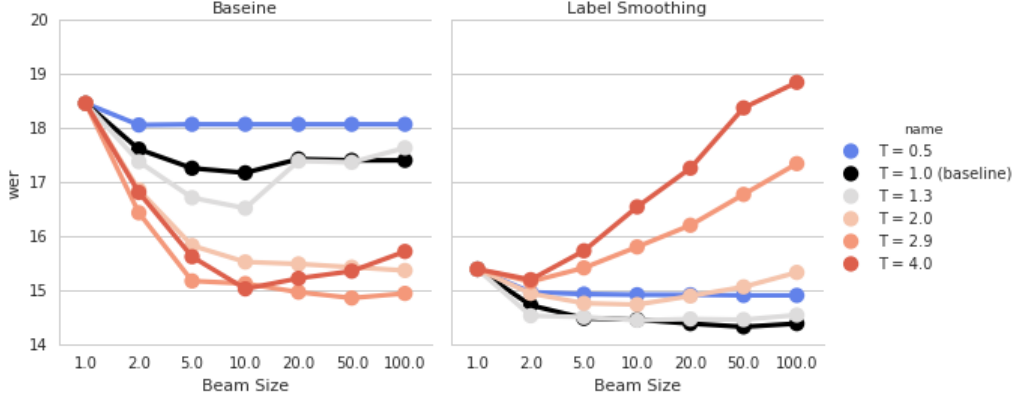


Figure 1: Influence of beam width and SoftMax temperature on decoding accuracy. In the baseline case (no label smoothing) increasing the temperature reduces the error rate. When label smoothing is used the next-character prediction improves, as witnessed by WER for beam size=1, and tuning the temperature does not bring additional benefits.

training loss over a single utterance is

$$\text{loss}(\mathbf{y}, \mathbf{x}) = - \sum_i \sum_c T(\mathbf{y}_i, c) \log p(\mathbf{y}_i | \mathbf{y}_{<i}, \mathbf{x}), \quad (7)$$

where $T(\mathbf{y}_i, c)$ denotes the target label function. In the baseline model $T(\mathbf{y}_i, c)$ is the indicator $[\mathbf{y}_i = c]$, i.e. its value is 1 for the correct character, and 0 otherwise. When label smoothing is used, T encodes a distribution over characters.

2.4. Beam Search

Basic Decoding of new utterances amounts to finding the character sequence \mathbf{y}^* that is most probable under the distribution computed by the network:

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} p(\mathbf{y} | \mathbf{x}) = \arg \min_{\mathbf{y}} - \log p(\mathbf{y} | \mathbf{x}). \quad (8)$$

Due to the recurrent formulation of the speller function, the most probable transcript cannot be found exactly using the Viterbi algorithm. Instead, approximate search methods are used. Typically, best results are obtained using beam search. The search begins with the set (beam) of hypotheses containing only the empty transcript. At every step, candidate transcripts are formed by extending hypothesis in the beam by one character. The candidates are then scored using the model, and a certain number of top-scoring candidates forms the new beam. The model indicates that a transcript is considered to be finished by emitting a special EOS (end-of-sequence) token.

Language Model Integration during decoding can be implemented by extending the beam search cost [12, 4, 15]:

$$\mathbf{y}^* = \arg \min_{\mathbf{y}} - \log p(\mathbf{y} | \mathbf{x}) - \lambda \log p_{\text{LM}}(\mathbf{y}) - \gamma \text{coverage}, \quad (9)$$

where coverage refers to a term that promotes longer transcripts described it in detail in Section 3.3.

We have identified two challenges in adding the language model. First, due to model overconfidence deviations from the best guess of the network drastically changed the term $-\log p(\mathbf{y} | \mathbf{x})$, which made balancing the terms in eq. (9) difficult. Second, incomplete transcripts were produced unless a recording coverage term was added.

3. Solutions to Seq2Seq Failure Modes

We have analysed the impact of model confidence by separating its effects on model accuracy and beam search effectiveness. We also propose a practical solution to the partial transcriptions problem, relating to the coverage of the input utterance.

3.1. Impact of Model Overconfidence

The cross-entropy training criterion (7) promotes confident answers, because it is minimized when the model concentrates all of its output distribution on the correct ground-truth character. This leads to very peaked probability distributions, effectively preventing the model from indicating sensible alternatives to a given character, such as its homophones. Moreover, overconfidence can harm learning of deeper layers. The derivative of the loss backpropagated through the SoftMax function to the logit corresponding to character c equals $[\mathbf{y}_i = c] - p(\mathbf{y}_i | \mathbf{y}_{<i}, \mathbf{x})$, which approaches 0 as the network’s output becomes concentrated on the correct character. Therefore whenever the spelling RNN makes a good prediction, very little training signal is propagated through the attention mechanism to the listener.

Model overconfidence can have two consequences. First, next-step character predictions may have low accuracy due to overfitting. Second, overconfidence may impact the ability of beam search to find good solutions and to recover from errors.

We first investigate the impact of confidence on beam search by varying the temperature of the SoftMax function. Without retraining the model, we change the character probability distribution to depend on a *temperature* hyperparameter T :

$$p(y_i) = \frac{\exp(l_i/T)}{\sum_j \exp(l_j/T)}. \quad (10)$$

As temperature increases the distribution over characters becomes more uniform. However, the preferences of the model are retained and the ordering of tokens from the most to least probable is preserved. Tuning the temperature therefore allows to demonstrate the impact of model confidence on beam search, without affecting the accuracy of next step predictions.

Decoding results of a baseline model on the WSJ dev93 data set are presented in Figure 1. We haven’t used a language model. At high temperatures deletion errors dominated. We didn’t want to change the beam search cost and instead constrained the search

Table 1: Example of model failure on validation '4k0c030n'

Transcript	LM cost $\log p(y)$	Model cost $\log p(y x)$
"chase is nigeria's registrar and the society is an independent organization hired to count votes"	-108.5	-34.5
"in the society is an independent organization hired to count votes"	-64.6	-19.9
"chase is nigeria's registrar"	-40.6	-31.2
"chase's nature is registrar"	-37.8	-20.3
""	-3.5	-12.5

to emit the EOS token only when its probability was within a narrow range from the most probable token. We compare the default setting ($T = 1$), with a sharper distribution ($T = 0.5$) and smoother distributions ($T \in \{1.3, \dots, 4\}$). All strategies lead to the same greedy decoding accuracy, because temperature changes do not affect the selection of the most probable character. As temperature increases beam search finds better solutions, however care must be taken to prevent truncated transcripts.

3.2. Label Smoothing Prevents Overconfidence

A elegant solution to model overconfidence was problem proposed for the Inception image recognition architecture [16]. For the purpose of computing the training cost the ground-truth label distribution is smoothed, with some fraction of the probability mass assigned to classes other than the correct one. This in turn prevents the model from learning to concentrate all probability mass on a single token. Additionally, the model receives more training signal because the error function cannot easily saturate.

Originally *uniform* label smoothing scheme was proposed in which the model is trained to assign β probability mass to the correct label, and spread the $1 - \beta$ probability mass uniformly over all classes [16]. Better results can be obtained with *unigram* smoothing which distributes the remaining probability mass proportionally to the marginal probability of classes [17].

In this contribution we propose a *neighborhood* smoothing scheme that uses the temporal structure of the transcripts: the remaining $1 - \beta$ probability mass is assigned to tokens neighboring in the transcript. Intuitively, this is similar to training the model on insertion and deletions errors. To recover from insertion/deletion the model should emit the previous/next token.

We have repeated the analysis of SoftMax temperature on beam search accuracy on a network trained with neighborhood smoothing in Figure 1. We can observe two effects. First, the model is regularized and greedy decoding leads to nearly 3 percentage smaller error rate. Second, the entropy of network predictions is higher, allowing beam search to discover good solutions without the need for temperature control. Moreover, since model is trained and evaluated with $T = 1$ we didn't have to control the emission of EOS token.

3.3. Solutions to Partial Transcripts Problem

When a language model is used wide beam searches often yield incomplete transcripts. With narrow beams, the problem is less visible due to implicit hypothesis pruning. We illustrate a failed decoding in Table 1. The ground truth (first row) is the least probable transcript according both to the network and the language model. A width 100 beam search with a trigram language model finds the second transcript, which misses the beginning of the utterance. The last rows demonstrate severely incomplete tran-

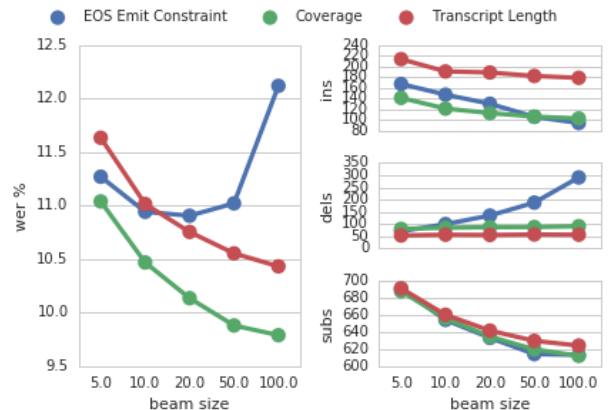


Figure 2: Impact of using techniques that prevent incomplete transcripts when a trigram language models is used on the dev93 WSJ subset. Results are averaged across two networks

scriptions that may be discovered when decoding is performed with even wider beam sizes.

We compare three strategies designed to prevent incomplete transcripts. The first strategy doesn't change the beam search criterion, but forbids emitting the EOS token unless its probability is within a set range of that of the most probable token. This strategy prevents truncations, but is inefficient against omissions in the middle of the transcript, such as the failure shown in Table 1. Alternatively, beam search criterion can be extended to promote long transcripts. A term depending on the *transcript length* was proposed for both CTC [4] and seq2seq [12] networks, but its usage was reported to be difficult because beam search was looping over parts of the recording and additional constraints were needed [12]. Finally, we propose to use a *coverage* term that counts the number of frames that have received a cumulative attention greater than τ :

$$\text{coverage} = \sum_j \left[\sum_i \alpha_{ij} > \tau \right]. \quad (11)$$

The coverage criterion prevents looping over the utterance because once the cumulative attention bypasses the threshold τ a frame is counted as selected and subsequent selections of this frame do not reduce the decoding cost. In our implementation, the coverage is recomputed at each beam search iteration using all attention weights produced up to this step.

In Figure 2 we compare the effects of the three methods when decoding a network that uses label smoothing and a trigram language model. Unlike [12] we didn't experience looping when beam search promoted transcript length. We hypothesize that label smoothing increases the cost of correct character emissions which helps balancing all terms used by beam search. We observe that at large beam widths constraining EOS emissions is not sufficient. In contrast, both promoting coverage and transcript length yield improvements with increasing beams. However, simply maximizing transcript length yields more word insertion errors and achieves an overall worse WER.

4. Experiments

We conducted all experiments on the Wall Street Journal dataset, training on si284, validating on dev93, and evaluating on eval92 set. We used 80-dimensional mel-scale filterbanks extracted

Table 2: Results without separate language model on WSJ.

Model	Parameters	dev93	eval92
CTC [3]	26.5M	-	27.3
seq2seq [12]	5.7M	-	18.6
seq2seq [24]	5.9M	-	12.9
seq2seq [22]	-	-	10.5
Baseline	6.6M	17.9	14.2
Unigram LS	6.6M	13.7	10.6
Temporal LS	6.6M	14.1	10.7

every 10ms form 25ms windows, extended with their temporal first and second order differences and per-speaker mean and variance normalization. Our character set consisted of lowercase letters, the space, the apostrophe, a noise marker, and start- and end- of sequence tokens. For comparison with previously published results, experiments involving language models used an extended-vocabulary trigram language model built by the Kaldi WSJ s5 recipe [18]. We have used the FST framework to compose the language model with a "spelling lexicon" [6, 12, 19]. All models were implemented using Tensorflow [20].

Our base configuration implemented the Listener using 4 bidirectional LSTM layers of 256 units per direction (512 total), interleaved with 3 time-pooling layers which resulted in an 8-fold reduction of the input sequence length, approximately equating the length of hidden activations to the number of characters in the transcript. The Speller was a single LSTM layer with 256 units. Input characters were embedded into 30 dimensions. The attention MLP used 128 hidden units, previous attention weights were accessed using 3 convolutional filters spanning 100 frames. LSTM weights were initialized uniformly over the range ± 0.075 . Networks were trained using 8 asynchronous replica workers each employing the ADAM algorithm [21] with default parameters and the learning rate set initially to 10^{-3} , then reduced to 10^{-4} and 10^{-5} after 400k and 500k training steps, respectively. Static Gaussian weight noise with standard deviation 0.075 was applied to all weight matrices after 20k training steps. We have also used a small weight decay of 10^{-6} .

We have compared two label smoothing methods: *unigram* smoothing [17] with the probability of the correct label set to 0.95 and *neighborhood* smoothing with the probability of correct token set to 0.9 and the remaining probability mass distributed symmetrically over neighbors at distance ± 1 and ± 2 with a 5 : 2 ratio. Grid searchtuning of smoothing parameters revealed that good results can be obtained for a broad range of settings.

We have gathered results obtained without language models in Table 2. We have used a beam size of 10 and no mechanism to promote longer sequences. We report averages of two runs taken at the epoch with the lowest validation WER. Label smoothing brings a large error rate reduction, nearly matching the performance achieved with very deep and sophisticated encoders [22].

Table 3 gathers results that use the extended trigram language model. We report averages of two runs. For each run we have tuned beam search parameters on the validation set and applied them on the test set. A typical setup used beam width 200, language model weight $\lambda = 0.5$, coverage weight $\gamma = 1.5$ and coverage threshold $\tau = 0.5$. Our best result surpasses CTC-based networks [6] and matches the results of a DNN-HMM and CTC ensemble [23].

Table 3: Results with extended trigram language model on WSJ.

Model	dev93	eval92
seq2seq [12]	-	9.3
CTC [3]	-	8.2
CTC [6]	-	7.3
Baseline + Cov	12.6	8.9
Unigram LS + Cov.	9.9	7.0
Temporal LS + Cov.	9.7	6.7

5. Related Work

Label smoothing was proposed as an efficient regularizer for the Inception architecture [16]. Several improved smoothing schemes were proposed, including sampling erroneous labels instead of using a fixed distribution [25], using the marginal label probabilities [17], or using early errors of the model [26]. Smoothing techniques increase the entropy of a model's predictions, a technique that was used to promote exploration in reinforcement learning [27, 28, 29]. Label smoothing prevents saturating the SoftMax nonlinearity and results in better gradient flow to lower layers of the network [16]. A similar concept, in which training targets were set slightly below the range of the output nonlinearity was proposed in [30].

Our seq2seq networks are locally normalized, i.e. the speller produces a valid probability distribution at every step. Alternatively, normalization can be performed globally on whole transcripts. In discriminative training of classical ASR systems normalization is performed over lattices [31]. In the case of recurrent networks lattices are replaced by beam search results. Global normalization has yielded important benefits on many NLP tasks including parsing and translation [32, 33]. Global normalization is expensive, because beam search is run at each training step. It yet unsettled whether globally normalized models can be approximated by cheaper to train locally normalized models with proper regularization such as label smoothing.

Using source coverage vectors has been investigated in neural machine translation models. Past attentions vectors were used as auxiliary inputs in the emitting RNN either directly [34], or as cumulative coverage information [35]. Coverage embeddings vectors associated with source words end modified during training were proposed in [36]. Our solution that employs a coverage penalty at decode time only is most similar to the one used by the Google Translation system [10].

6. Conclusions

We have analysed and solved two problems of seq2seq neural networks: output overconfidence and emissions of incomplete transcripts. We have demonstrated that with efficient regularization and careful decoding the sequence-to-sequence approach to speech recognition can be competitive with other non-HMM techniques, such as CTC.

7. Acknowledgements

We have greatly benefited from discussions with Mohammad Norouzi and Patrick Nguyen.

8. References

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

- [2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. 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.
- [3] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *ICML*, vol. 14, 2014, pp. 1764–1772.
- [4] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Sathesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.
- [5] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, “Deep speech 2: End-to-end speech recognition in english and mandarin,” *arXiv preprint arXiv:1512.02595*, 2015.
- [6] Y. Miao, M. Gowayed, and F. Metzke, “Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [8] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [9] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [10] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *CoRR*, vol. abs/1609.08144, 2016.
- [11] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.
- [12] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4945–4949.
- [13] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” *arXiv preprint arXiv:1508.01211*, 2015.
- [14] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” *arXiv preprint arXiv:1611.05358*, 2016.
- [15] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio, “On using monolingual corpora in neural machine translation,” *CoRR*, vol. abs/1503.03535, 2015.
- [16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” *arXiv preprint arXiv:1512.00567*, 2015.
- [17] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. Hinton, “Regularizing neural networks by penalizing confident output distributions,” in *Submitted to ICLR 2017*, 2017, <https://openreview.net/forum?id=HkCjNI5ex>.
- [18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.
- [19] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, “OpenFst: A general and efficient weighted finite-state transducer library,” in *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, ser. Lecture Notes in Computer Science, vol. 4783. Springer, 2007, pp. 11–23, <http://www.openfst.org>.
- [20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.
- [21] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [22] Y. Zhang, W. Chan, and N. Jaitly, “Very deep convolutional networks for end-to-end speech recognition,” *arXiv preprint arXiv:1610.03022*, 2016.
- [23] A. Graves and N. Jaitly, “Towards End-To-End Speech Recognition with Recurrent Neural Networks,” in *ICML ’14*, 2014, pp. 1764–1772.
- [24] W. Chan, Y. Zhang, Q. Le, and N. Jaitly, “Latent sequence decompositions,” *arXiv preprint arXiv:1610.03035*, 2016.
- [25] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian, “Disturblabel: Regularizing cnn on the loss layer,” *arXiv preprint arXiv:1605.00055*, 2016.
- [26] A. Aghajanyan, “Softtarget regularization: An effective technique to reduce over-fitting in neural networks,” *CoRR*, vol. abs/1609.06693, 2016.
- [27] R. J. Williams and J. Peng, “Function optimization using connectionist reinforcement learning algorithms,” *Connection Science*, vol. 3, no. 3, pp. 241–268, 1991.
- [28] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” *arXiv preprint arXiv:1602.01783*, 2016.
- [29] Y. Luo, C.-C. Chiu, N. Jaitly, and I. Sutskever, “Learning on-line alignments with continuous rewards policy gradient,” *arXiv preprint arXiv:1608.01281*, 2016.
- [30] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Efficient BackProp*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48.
- [31] X. He, L. Deng, and W. Chou, “Discriminative learning in sequential pattern recognition,” *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 14–36, 2008.
- [32] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, “Globally normalized transition-based neural networks,” *CoRR*, vol. abs/1603.06042, 2016.
- [33] S. Wiseman and A. M. Rush, “Sequence-to-sequence learning as beam-search optimization,” *CoRR*, vol. abs/1606.02960, 2016.
- [34] M. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *CoRR*, vol. abs/1508.04025, 2015.
- [35] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, “Modeling coverage for neural machine translation,” *CoRR*, vol. abs/1601.04811, 2016.
- [36] H. Mi, B. Sankaran, Z. Wang, and A. Ittycheriah, “Coverage embedding model for neural machine translation,” *CoRR*, vol. abs/1605.03148, 2016.