

# LEARNING NEURAL TRANS-DIMENSIONAL RANDOM FIELD LANGUAGE MODELS WITH NOISE-CONTRASTIVE ESTIMATION

*Bin Wang, Zhijian Ou*

Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, Beijing, China.  
wangbin12@mails.tsinghua.edu.cn, ozj@tsinghua.edu.cn

## ABSTRACT

Trans-dimensional random field language models (TRF LMs) where sentences are modeled as a collection of random fields, have shown close performance with LSTM LMs in speech recognition and are computationally more efficient in inference. However, the training efficiency of neural TRF LMs is not satisfactory, which limits the scalability of TRF LMs on large training corpus. In this paper, several techniques on both model formulation and parameter estimation are proposed to improve the training efficiency and the performance of neural TRF LMs. First, TRFs are reformulated in the form of exponential tilting of a reference distribution. Second, noise-contrastive estimation (NCE) is introduced to jointly estimate the model parameters and normalization constants. Third, we extend the neural TRF LMs by marrying the deep convolutional neural network (CNN) and the bidirectional LSTM into the potential function to extract the deep hierarchical features and bidirectionally sequential features. Utilizing all the above techniques enables the successful and efficient training of neural TRF LMs on a 40x larger training set with only 1/3 training time and further reduces the WER with relative reduction of 4.7% on top of a strong LSTM LM baseline.

**Index Terms**— Language Model, Random Field, Speech Recognition, Noise-contrastive Estimation

## 1. INTRODUCTION

Statistical language models are a crucial component in many applications, such as automatic speech recognition (ASR) and machine translation (MT), by encoding the linguistic regularities in terms of the joint probability of words in a sentence. Currently, the recurrent neural network approach, which follows the directed graphical modeling approach, has shown significant perplexity reductions over the classic n-gram LMs in various benchmarks such as the Penn Tree Bank [1] and the Google One Billion corpus [2], and also achieves the state-of-the-art word error rates (WERs) in ASR [3].

In contrast, a new trans-dimensional random field (TRF) LM [4, 5, 6] has been proposed in the undirected graphical modeling approach, where sentences are modeled as a collection of random fields and the joint probability is defined in terms of potential functions. With the power of flexibly supporting rich features, neural TRF LMs [6] with a nonlinear potential function defined by a deep convolutional neural network (CNN), outperform the n-gram LMs significantly and perform close to LSTM LMs and are computational more efficient in inference.

However the training speed of neural TRF LMs is not satisfactory, which is caused by several factors. First, the estimation of TRFs

depends on the newly developed augmented stochastic approximation (AugSA) algorithm [5], where the samples generated based on the Markov chain Monte Carlo (MCMC) theory are used to update the parameters. The convergence of such approach heavily depends on the quality and quantity of the samples and usually needs a lot of training iterations. Second, the trans-dimensional mixture sampling method (TransMS) which is used to generate samples of varying dimensions is computationally expensive, even when the joint stochastic approximation (JSA) [7] strategy is introduced to improve its efficiency. This is partly due to the fact that the sampling operation depends on the current model parameters, and can not be performed beforehand or be parallelized with other training operations. Third, learning the neural TRF LMs need to optimize a non-convex objective function, which is much harder than learning the discrete TRF LMs which is a convex optimization. All the above factors limit the training efficiency of neural TRFs as well as their scalability on large training corpus.

In this paper, the following contributions are made to improve both the training efficiency and the performance of neural TRF LMs. First, TRFs are defined in the form of exponential tilting of a reference distribution. As a result, only the difference between the data distribution and the reference distribution needs to be fitted by TRFs, which is much simpler than fitting the empirical distribution directly. Second, noise-contrastive estimation (NCE) [8] is introduced to train TRF LMs by optimizing a discriminator between the real sentences drawn from the training set and the noise sentences drawn from a noise distribution. The noise distribution in NCE is independent of the model distribution and hence drawing noise sentences can be parallelized with model estimation to accelerate the training process significantly. Meanwhile, the normalization constants in TRFs can be treated as the normal parameters and be jointly optimized with the model parameters during NCE training. Third, we enhance the neural TRF in [6] by marrying deep CNN and bidirectional LSTM into the potential function to extract the deep hierarchical features and bidirectionally sequential features.

Two experiments are designed in the paper, including a pilot experiment called short-word morphology to validate the NCE in TRF situation and an ASR experiment on CHiME-4 Challenge data to evaluate the scalability of neural TRF LMs. First, the pilot experiment reveals the effectiveness of NCE in TRF training where more than one normalization constant need to be estimated simultaneously. Then in ASR, a neural TRF LM with a LSTM LM as the reference distribution is trained using NCE in 40x larger training corpus with only a third of training time, compared with the results reported in [6]. The results show that the neural TRF models and the directed graphical modeling models are complementary. The lowest WERs are achieved by combining the neural TRF LMs with n-gram LMs and LSTM LMs, with relative reduction of 4.7% over state-of-the-art LSTM LMs, i.e. combining the n-gram LMs and LSTM

---

This work is supported by NSFC grant 61473168.

LMs.

The rest of the paper is organized as follows. We first discuss related work in Section 2. Then in Section 3, we describe the new formulation of neural TRFs with the neural network potential. The NCE training method is introduced in Section 4. After presenting the experimental results in Section 5, the conclusions are made in Section 6.

## 2. RELATED WORK

The noise-contrastive estimation is first proposed in [8] for the estimation of unnormalized statistical models, whose normalization constants can not be obtained in closed form. In language modeling, NCE is used to train the conditional neural network (NN) LMs, such as the feedforward neural network LMs [9] and LSTM LMs [10], by treating the learning as a binary classification problem between the target words and the noise samples. As the normalization terms of the prediction probabilities are dependent on the context and can not be enumerated, an approximate approach is to freeze them to an empirical value, which is 1 in [10] and  $e^9$  in [11]. Moreover, [10] accelerates the NCE training by sharing the noise samples for each target word in the mini-batch to reduce the sample times. Then [12] further omits the sampling process by taking the other target words in current mini-batch as the noise samples. In this paper, NCE is evaluated in the estimation of TRF LMs, which are defined on the trans-dimensional state space. To our best knowledge, this is the first time that NCE is applied in the trans-dimensional setting.

The idea of defining a model in the form of exponential tilting of a reference distribution has been studied in natural image generative modeling [13, 14] where the reference distribution is set to the Gaussian white noise distribution. Similar formulations have been used in language modeling with a class n-gram LM as the reference, including the maximum entropy LMs [15] and the whole-sentence LMs [16, 17, 18]. In this paper, a LSTM LM is used as the reference distribution to define a neural TRF LM.

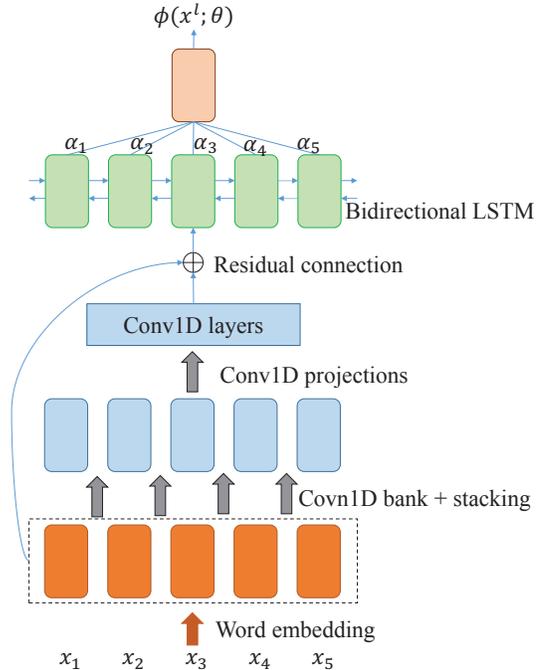
## 3. NEURAL TRANS-DIMENSIONAL RANDOM FIELD LMS

Different from the neural TRFs in [6], here the joint probability of a sequence  $x^l$  and its length  $l$  ( $l = 1, \dots, m$ ) is assumed to be in the form of exponential tilting of a reference distribution  $q(x^l)$ :

$$p(l, x^l; \theta) = \frac{\pi_l}{Z_l(\theta)} q(x^l) e^{\phi(x^l; \theta)} \quad (1)$$

where  $x^l = (x_1, \dots, x_l)$  is a word sequence of length  $l$ ,  $\pi_l$  is the prior length probability,  $\theta$  indicates the set of parameters,  $\phi$  is the potential function, and  $Z_l(\theta)$  is the normalization constant of length  $l$ , i.e.  $Z_l(\theta) = \sum_{x^l} q(x^l) e^{\phi(x^l; \theta)}$ . Different with the TRFs defined in [5, 6], a reference distribution  $q(x^l)$  is introduced as the baseline distribution, and the role of TRFs is to fit the difference between the data distribution and the reference distribution. If  $q(x^l)$  is a good approximation of the data distribution, such as the LSTM LMs used in our experiments, fitting the difference between the data distribution and the reference distribution  $q(x^l)$  shall be much simpler than fitting the data distribution directly.

To compute the potential function, we define a neural network by combining the deep CNN structure and the bidirectional LSTM structure to extract both the deep hierarchical features and bidirectionally sequential features, as shown in Figure 1. The architecture is



**Fig. 1.** Neural network architecture used to define the potential function  $\phi(x^l; \theta)$

motivated by the encoder module in [19] with some simplifications and is described as follows.

In the bottom, a deep CNN is used to extract deep hierarchical features, whose architecture is similar to the neural network used in [6] except that the weighted summation in the “CNN-stack” module is removed. There are three steps. First, each word  $x_i$  ( $i = 1, \dots, l$ ) in a sentence is mapped to an embedding vector. Then, these embedding vectors in a sentence are fed into a “CNN-bank” module, which contains a set of 1-D convolutional filters with widths ranging from 1 to  $K$ . These filters explicitly model local contextual information (akin to modeling unigrams, bigrams, up to  $K$ -grams) [20]. Third, the output feature maps from multiple filters with varying widths are spliced together, and fed into a few fixed-width 1-D convolutions to further extract hierarchical features, which resembles the “CNN-stack” module in [6]. The output of the last 1-D convolution is added with the word embedding vectors via a residual connection.

A bidirectional LSTM (BLSTM) is stacked on top of the deep CNN to extract long-range sequential features from the forward and backward contexts. Note that all the convolutions mentioned above are *half convolutions*<sup>1</sup>, i.e. convolutions are performed after padding zeros to the beginning and the end of the input sequences to preserve the time resolution.

Finally, the attention mechanism is introduced to summate the feature vectors of BLSTM at all positions, and the potential function  $\phi(x^l; \theta)$  is computed as follows.

$$\phi(x^l; \theta) = \lambda^T \sum_{i=1}^l \alpha_i h[:, i] + c \quad (2)$$

$$\alpha_i = \beta^T h[:, i], \quad i = 1, \dots, l \quad (3)$$

where  $h \in R^{2d \times l}$  is the hidden vectors in the BLSTM,  $d$  is the

<sup>1</sup><http://deeplearning.net/software/theano/library/tensor/nnet/conv.html>

number of hidden units of the BLSTM,  $h[:, i]$  is the  $i$ -th column of  $h$  and  $\lambda, \beta \in R^{2d}$ ,  $c \in R$  are the parameters. In summary,  $\theta$  denotes the collection of all the parameters defined in the neural networks.

#### 4. NOISE-CONTRASTIVE ESTIMATION

Noise-contrastive estimation (NCE) is proposed by [8] for the estimation of unnormalized statistical models, and has been successfully used in the estimation of neural network LMs [9, 10, 11, 12]. The basic idea of NCE is “learning by comparison”, i.e. to perform nonlinear logistic regression to discriminate between the data samples drawn from the training set and the noise samples drawn from a known noise distribution. The normalization constants can be treated as the normal parameters and updated together with the model parameters.

To apply NCE to estimate the neural TRFs defined in (1), first we rewrite the formulation in (1) by introducing a new parameter  $\zeta$ :

$$p(l, x^l; \theta, \zeta) = \pi_l q(x^l) e^{\phi(x^l; \theta) - \zeta_l} \quad (4)$$

where  $\zeta = (\zeta_1, \dots, \zeta_l)$  denotes the hypothesized values of the true logarithmic normalization constants  $\zeta_l^* = \log Z_l$ , which can be estimated in NCE. Assuming for each sequence in the training set,  $\nu$  noise sequences of varying lengths are generated from a noise distribution  $p_n(l, x^l)$ . Then the probabilities of a sequence  $(l, x^l)$  belonging to the data distribution  $P(C = 0|l, x^l; \theta, \zeta)$  and the noise distribution  $P(C = 1|l, x^l; \theta, \zeta)$  are given by

$$P(C = 0|l, x^l; \theta, \zeta) = \frac{p(l, x^l; \theta, \zeta)}{p(l, x^l; \theta, \zeta) + \nu p_n(l, x^l)} \quad (5)$$

$$P(C = 1|l, x^l; \theta, \zeta) = 1 - P(C = 0|l, x^l; \theta, \zeta) \quad (6)$$

Given the training set  $D$ , NCE maximizes the following conditional log-likelihood:

$$J(\theta, \zeta) = \frac{1}{|D|} \sum_{(l, x^l) \in D} \log P(C = 0|l, x^l; \theta, \zeta) + \nu \frac{1}{|B|} \sum_{(l, x^l) \in B} \log P(C = 1|l, x^l; \theta, \zeta) \quad (7)$$

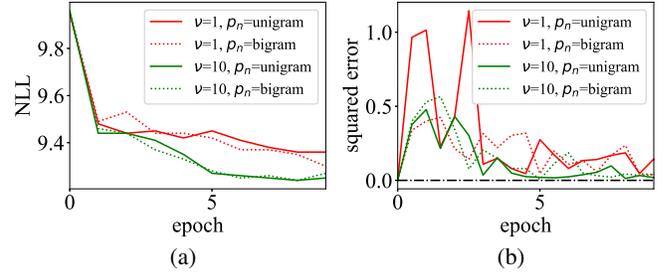
where  $B$  is the noise sample set, which is generated from the noise distribution  $p_n(l, x^l)$ ,  $|D|$  and  $|B|$  are the number of samples in  $D$  and  $B$  respectively, and satisfy  $\nu = |B|/|D|$ . To maximize the objective  $J(\theta, \zeta)$ , the gradient with respect to  $\theta$  and  $\zeta$  can be computed as follows:

$$\frac{\partial J(\theta, \zeta)}{\partial \theta} = \sum_{(l, x^l) \in D} w_t(l, x^l; \theta, \zeta) \frac{\partial \phi(x^l; \theta)}{\partial \theta} + \sum_{(l, x^l) \in B} w_n(l, x^l; \theta, \zeta) \frac{\partial \phi(x^l; \theta)}{\partial \theta} \quad (8)$$

$$\frac{\partial J(\theta, \zeta)}{\partial \zeta_j} = - \sum_{(l, x^l) \in D} w_t(l, x^l; \theta, \zeta) 1(l = j) - \sum_{(l, x^l) \in B} w_n(l, x^l; \theta, \zeta) 1(l = j) \quad (9)$$

where

$$\begin{cases} w_t(l, x^l; \theta, \zeta) = \frac{1 - P(C = 0|l, x^l; \theta, \zeta)}{|D|} \\ w_n(l, x^l; \theta, \zeta) = -\frac{P(C = 0|l, x^l; \theta, \zeta)}{|D|} \end{cases} \quad (10)$$



**Fig. 2.** The results of short-word morphology. (a) is the negative log-likelihood (NLL) on the valid set using the true normalization constants. (b) is the squared error between the estimated normalization constants  $\zeta$  and the true normalization constants  $\zeta^*$ . Different configurations for number of noise samples  $\nu = 1, 10$  and noise distribution  $p_n(x^l) = \text{unigram}, \text{bigram}$  are investigated.

and  $1(l = j)$  is 1 if  $l = j$  and 0 otherwise. The gradient of the potential function  $\phi(x^l; \theta)$  with respect to the parameters  $\theta$  can be computed through the back-propagation algorithm. Then any gradient method can be used to optimize the parameters and normalization constants, such as stochastic gradient descent (SGD) or Adam [21].

In our experiments, the noise distribution is defined as

$$p_n(l, x^l) = \pi_l p_n(x^l), \quad (11)$$

where  $\pi_l, l = 1, \dots, m$  is the prior length probability and  $p_n(x^l)$  is a  $n$ -gram LM. This makes the noise distribution  $p_n(l, x^l)$  close to the data distribution, and we can achieve a good performance with a medium sample number  $\nu$ , such as 20 in our following experiments, as suggested in [8].

## 5. EXPERIMENTS

In this section, two experiments are conducted to evaluate NCE training for neural TRFs. First, a pilot experiment - short-word morphology is designed to exactly evaluate the performance of NCE. Then neural TRFs and NCE training are applied to language modeling in ASR using CHiME-4 Challenge data [22].

### 5.1. Short-word morphology

In this section, we design a simulation experiment to validate the NCE training for neural TRF models, where more than one normalization constants need to be estimated simultaneously. The training set and valid set include 5,143 and 69 different English words with at most 3 characters respectively, which are extracted from English Gigaword dataset [23]. Every word is decomposed to a character sequence and the objective of this experiment is to assign a probability to the character sequences. The vocabulary contains a total of 28 symbols, i.e. the 26 English letters and two auxiliary symbols - the beginning symbol and the end symbol, which are added to the beginning and the end of every character sequence respectively. As the length of each sequences is no more than 3, the normalization constants  $Z_l, l = 1, 2, 3$  can be calculated exactly.

A neural TRF defined in section 3 is applied to model these character sequences, with  $q(x^l)$  being uniform. For the potential function, each character is embedded to a 16 dimensional vector and directly fed into the BLSTM without passing the CNN layers. The BLSTM contains one hidden layer for each direction and 16 hidden units for each layer. The prior length distribution  $\pi_l$  in model

$\pi_l$	empirical length probabilities	
$q(x^l)$	a LSTM with 512 embedding size, 2 hidden layers and 512 hidden units per layer [1]	
$\phi(x^l; \theta)$	embedding size	200
	Conv1D bank	cnn- $k$ -128-ReLU, with $k$ ranging from 1 to 10
	Conv1D layers	cnn-3-128-ReLU $\rightarrow$ cnn-3-128-ReLU $\rightarrow$ cnn-3-128-ReLU
	BLSTM	1 layer and 128 hidden units

**Table 1.** The configuration of neural TRFs. “cnn- $k$ - $n$ -ReLU” denotes a 1-D CNN with filter width  $k$ , output dimension  $n$  and rectified linear unit (ReLU) activation. “ $A \rightarrow B$ ” denotes that the output of layer  $A$  is fed into layer  $B$ .

distribution (4) and the noise distribution (11) are both set to the empirical length distribution. The parameters  $\theta$  are initialized randomly within an interval from -0.1 to 0.1, and the initial value of normalization constants are  $\zeta_l = l \times \log V$  where  $V = 28$  denotes the vocabulary size. We use Adam to update the parameter  $\theta$  and the normalization constants  $\zeta$  with the mini-batch size being 10. The learning rates for  $\theta$  and  $\zeta$  are fixed to 0.001 and 0.01 respectively. We investigate different configurations for sample number  $\nu$  and the noise distribution  $p_n(x^l)$ . The results are summarized in Figure 2. The main conclusions are as follows.

First, NCE performs well in TRF training, with both the NLL and normalization constants converging after several training epochs. Second, increasing the sample number  $\nu$  from 1 to 10 can lead to a lower NLL (Figure 2(a)) and make the normalization constants  $\zeta$  converge fast to the true  $\zeta^*$  (Figure 2(b)). Third, for a small sample number  $\nu = 1$ , changing the noise distribution  $p_n(x^l)$  from unigram (red solid line) to bigram (red dot line) will improve the convergence of NCE. If we increase the sample number to  $\nu = 10$ , the choice of the noise distribution makes less differences. The above observations of NEC in TRF training is consistent with [8]

## 5.2. Neural TRF LMs in speech recognition

In this section, we evaluate the performance and scalability of neural TRF LMs trained by NCE over CHiME-4 Challenge data. The training corpus for language modeling contains about 37 millions tokens, which is about 40 times of the Penn Treebank (PTB) training set used in [6]. The vocabulary is limited to 5 K words, including a special token  $\langle \text{UNK} \rangle$  denoting the out-of-vocabulary words.

For evaluation in terms of speech recognition WERs, various LMs are applied to rescore the 100-best lists from recognizing CHiME-4 development and test data. For each utterance, the 100-best list of candidate sentences are generated by the multi-channel ASR system developed by our team under CHiME-4 Challenge rule, which is detailed in [24]. All the hyper-parameters of training LMs, such as learning rates and training epochs, are tuned on the development utterances to achieve the lowest WER. In CHiME-4 Challenge, different systems are compared in terms of the WERs on the real test set, which consists of speech recordings collected in real environments.

The configuration of the neural TRF LMs used in this experiments are listed in Table 1. First a LSTM LM [1] with 2 hidden layers and 512 hidden units per layer is trained on the training set using SGD method, and serves as the reference distribution  $q(x^l)$  of neural TRF LMs in (1). Then NCE with  $\nu = 20$  is used to train the

model	Dev		Test	
	real	simu	real	simu
KN5	5.03	4.79	7.38	5.78
LSTM2x512	3.63	3.24	5.70	4.53
neural TRF	3.53	3.20	5.68	4.36
KN5+LSTM2x512	3.56	3.29	<b>5.71</b>	4.18
KN5+neural TRF	3.53	3.22	5.54	4.20
KN5+LSTM2x512+neural TRF	3.42	3.10	<b>5.44</b>	4.13

**Table 2.** Speech recognition WERs on CHiME-4 Challenge data. “Dev” denotes the development set and “Test” denotes the test set. “real” denotes the speech recorded in real environments and “simu” denotes the simulated speech. “+” denotes the log-linear interpolation with equal weights.

neural TRF LM by fixing the parameters of the reference distribution  $q(x^l)$ , and updating the model parameters  $\theta$  and normalization constants  $\zeta$  simultaneously based SGD method. The learning rates for both  $\theta$  and  $\zeta$  are same which are initialized to 0.01 and reduce by half per epoch. The parameters  $\theta$  are initialized randomly within an interval from -0.1 to 0.1, and the normalization constants are initialized to  $\zeta_l = l$  for  $l = 1, \dots, m$ . We perform the training process for 2 epochs before the neural TRF achieves the lowest WER in the development set. The total training time is 1 day, which is one third of the training time reported in [6].

The WERs of various LMs are shown in Table 2, including a 5-gram LM with modified Kneser-Ney smoothing [25] (denoted by “KN5”), and the reference LSTM LMs  $q(x^l)$  (denoted by “LSTM2x512” to emphasize the hidden layers and hidden units). First, the LSTM LM achieves significant WER reduction compared with “KN5” with relative WER reduction 22.8% on real test set. Combining the LSTM LM and the n-gram LM leads to the state-of-the-art results, denoted by “KN5+LSTM2x512”. Building on top of the LSTM LM, our neural TRF LM can further reduce the WER with only 2-epoch NCE training. Remarkably, the lowest WER 5.44% on real test set is achieved by combining neural TRF LMs with n-gram LMs and LSTM LMs (denoted by “KN5+LSTM2x512+neural TRF”), with relative reduction of 4.7% compared to the strong state-of-the-art system “KN5+LSTM2x512”. This reveals that neural TRF LMs and various LMs in the directed graphical modeling approach are complementary, and a combination of them leads to WER reduction.

## 6. CONCLUSION

This paper presents our continuous effort to develop the TRF approach to language modeling. We make the following contributions to improve the convergence and scalability of neural TRF LMs. First, a reference distribution is introduced to serve as a baseline distribution. Then NCE is used to estimate the parameters and normalization constants jointly. Finally, new neural network structures are investigated in neural TRF LMs, by combing the CNN and bidirectional LSTM into the potential function. Utilizing all these techniques leads to the superior performance of TRF LMs in the CHiME-4 Challenge dataset, which is a supporting evidence of the power of TRF LMs on a medium size training set.

It is worthwhile to further investigate techniques to improve the training efficiency of TRF LMs. Moreover, integrating richer non-linear and structured features is an important future direction. The neural TRF models can also be applied to other sequential and trans-dimensional data modeling tasks in general.

## 7. REFERENCES

- [1] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, “Recurrent neural network regularization,” *arXiv preprint arXiv:1409.2329*, 2014.
- [2] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu, “Exploring the limits of language modeling,” *arXiv preprint arXiv:1602.02410*, 2016.
- [3] Gakuto Kurata, Abhinav Sethy, Bhuvana Ramabhadran, and George Saon, “Empirical exploration of novel architectures and objectives for language models,” in *Proc. INTERSPEECH*, 2017.
- [4] Bin Wang, Zhijian Ou, and Zhiqiang Tan, “Trans-dimensional random fields for language modeling,” in *Proc. Annu. Meeting of the Association for Computational Linguistics (ACL)*, 2015.
- [5] Bin Wang, Zhijian Ou, and Zhiqiang Tan, “Learning trans-dimensional random fields with applications to language modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017.
- [6] Bin Wang and Zhijian Ou, “Language modeling with neural trans-dimensional random fields,” in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2017.
- [7] Haotian Xu and Zhijian Ou, “Joint stochastic approximation learning of helmholtz machines,” *International Conference on Learning Representations (ICLR) Workshop Track*, 2016.
- [8] Michael Gutmann and Aapo Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.
- [9] Ashish Vaswani, Yingdong Zhao, Victoria Fossum, and David Chiang, “Decoding with large-scale neural language models improves translation,” in *EMNLP*, 2013.
- [10] Barret Zoph, Ashish Vaswani, Jonathan May, and Kevin Knight, “Simple, fast noise-contrastive estimation for large rnn vocabularies,” in *HLT-NAACL*, 2016.
- [11] Xie Chen, Xunying Liu, Mark JF Gales, and Philip C Woodland, “Recurrent neural network language model training with noise contrastive estimation for speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.
- [12] Youssef Oualil and Dietrich Klakow, “A batch noise contrastive estimation approach for training large vocabulary language models,” *arXiv preprint arXiv:1708.05997*, 2017.
- [13] Jifeng Dai, Yang Lu, and Ying-Nian Wu, “Generative modeling of convolutional neural networks,” *arXiv preprint arXiv:1412.6296*, 2014.
- [14] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu, “A theory of generative convnet,” in *International Conference on Machine Learning*, 2016.
- [15] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra, “A maximum entropy approach to natural language processing,” *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
- [16] Ronald Rosenfeld, “A whole sentence maximum entropy language model,” in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, 1997.
- [17] Fredy Amaya and José Miguel Benedí, “Improvement of a whole sentence maximum entropy language model using grammatical features,” in *Proc. Ann. Meeting of the Association for Computational Linguistics (ACL)*, 2001.
- [18] Teemu Ruokolainen, Tanel Alumäe, and Marcus Dobrinkat, “Using dependency grammar features in whole sentence maximum entropy language model for speech recognition,” in *Baltic HLT*, 2010.
- [19] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., “Tacotron: A fully end-to-end text-to-speech synthesis model,” *arXiv preprint arXiv:1703.10135*, 2017.
- [20] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush, “Character-aware neural language models,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [21] Diederik Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv:1412.6980 [cs.LG]*, 2014.
- [22] “Chime-4 speech separation and recognition challenge,” [http://spandh.dcs.shef.ac.uk/chime\\_challenge/chime2016/index.html](http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/index.html).
- [23] “Ldc english gigaword,” <https://catalog.ldc.upenn.edu/LDC2003T05>.
- [24] Hongyu Xiang, Bin Wang, and Zhijian Ou, “The thu-spmi chime-4 system: Lightweight design with advanced multi-channel processing, feature enhancement, and language modeling,” in *CHI-ME-4 Workshop*, 2016.
- [25] Stanley F. Chen and Joshua Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech & Language*, vol. 13, pp. 359–394, 1999.