

A COMPARABLE STUDY OF MODELING UNITS FOR END-TO-END MANDARIN SPEECH RECOGNITION

Wei Zou, Dongwei Jiang, Shuaijiang Zhao, Xiangang Li

AI Labs, Didi Chuxing, Beijing, China

{zouwei, jiangdongwei, zhaoshuaijiang, lixiangang}@didichuxing.com

Abstract

End-To-End speech recognition have become increasingly popular in mandarin speech recognition and achieved delightful performance. Mandarin is a tonal language which is different from English and requires special treatment for the acoustic modeling units. There have been several different kinds of modeling units for mandarin such as phoneme, syllable and Chinese character. In this work, we explore two major end-to-end models: connectionist temporal classification (CTC) model and attention based encoder-decoder model for mandarin speech recognition. We compare the performance of three different scaled modeling units: context dependent phoneme(CDP), syllable with tone and Chinese character. We find that all types of modeling units can achieve approximate character error rate (CER) in CTC model and the performance of Chinese character attention model is better than syllable attention model. Furthermore, we find that Chinese character is a reasonable unit for mandarin speech recognition. On DidiCallcenter task, Chinese character attention model achieves a CER of 5.68% and CTC model gets a CER of 7.29%, on the other DidiReading task, CER are 4.89% and 5.79%, respectively. Moreover, attention model achieves a better performance than CTC model on both datasets.

Index Terms: automatic speech recognition, connectionist temporal classification, attention model, modeling units, mandarin speech recognition

1. Introduction

Traditional speech recognition includes separate modeling components, including acoustic, phonetic and language models. These components of the system are trained separately, thus each components errors would extend during the process. Besides, building the components requires expert knowledge, for example, building a language model requires linguistic knowledge. The acoustic model is used to recognize context-dependent (CD) states or phonemes [1, 2], by bootstrapping from an existing model which is used for alignment. The pronunciation model maps the phonemes sequences into word sequences, then the language model scores the word sequences. A weighted finite state transducer (WFST) [3] integrates these models and do the decoding for the final result.

Recently, end-to-end speech recognition systems have become increasingly popular and achieve promising performance in mandarin [4]. End-to-end speech recognition methods predict graphemes directly from the acoustic data without linguistic knowledge, thus reducing the effort of building ASR systems greatly and making it easier for new language. The end-to-end ASR simplifies the system into a single network architecture, and it is likely to be more robust than a multi-module architecture. There are two major types of end-to-end architectures for ASR: The connectionist temporal classification (CTC) cri-

terion [5, 6, 7, 8], which has been used to train end-to-end systems that can directly predict grapheme sequences. The other is attention-based encoder-decoder model [9, 10, 11, 12] which applies an attention mechanism to perform alignment between acoustic frames and recognized symbols.

Attention-based encoder-decoder models have become increasingly popular [13, 7, 14, 15]. These models consist of an encoder network, which maps the input acoustic sequence into a higher-level representation, and an attention-based decoder that predicts the next output symbol conditioned on the full sequence of previous predictions.

A recent comparison of sequence-to-sequence models for speech recognition [9] has shown that Listen, Attend and Spell (LAS) [16], a typical attention-based approach, offered improvements over other sequence-to-sequence models, and attention-based encoder-decoder model performs considerably well in mandarin speech recognition [17].

For Mandarin speech recognition, modeling units of acoustic model affect the performance significantly [18]. As we all know, CDP is most commonly used as the acoustic modeling units for speech recognition in mandarin [4]. In fact, there have been several different kinds of modeling units for Mandarin [19] such as phoneme, syllable and Chinese character. Compared with CDP, it will be easier to use syllable or character which does not need other prior model for alignment. Under current end-to-end speech recognition framework, we can get target output syllable sequence and character sequence directly from training transcripts and lexicon. Especially, in the case of using Chinese character models, we can get the desired results directly without lexicon and language model.

In order to find a more suitable end-to-end system and modeling unit in Mandarin speech recognition, we explore two major end-to-end models: CTC model and attention based encoder-decoder model. Meanwhile, We compare the performance of three different scaled modeling units: context dependent phoneme (CDP), syllable with tone and Chinese character.

The rest of this paper is organized as follows. Section 2 introduces the details of end-to-end speech recognition. Various model units for end-to-end speech recognition in mandarin are studied in Section 3. Section 4 describes the detail of the experiments. Section 5 draws some conclusions and outlines our future work.

2. End-to-End Speech Recognition

Recently, end-to-end speech recognition systems have become increasingly popular and achieve encouraging performance in mandarin.

2.1. Connectionist Temporal Classification(CTC)

The CTC criterion was proposed by Graves et al. [5] as a way of training end-to-end models without requiring a frame-level alignment of the target labels for a training utterance. To achieve this, an extra blank label denoted $\langle b \rangle$ is introduced to map frames and labels to the same length, which can be interpreted as no target label. CTC computes the conditional probability by marginalizing all possible alignments and assuming conditional independence between output predictions at different time steps given aligned inputs.

Given a label sequence y corresponding to the utterance x , where y is typically much shorter than the x in speech recognition. Let $\beta(y, x)$ be the set of all sequences consisting of the labels in $\mathcal{Y} \cup \langle b \rangle$, which are of length $|x| = T$, and which are identical to y after first collapsing consecutive repeated targets and then removing any blank symbols (e.g., $A\langle b \rangle AA\langle b \rangle B \rightarrow AAB$). CTC model defines the probability of the label sequence conditioned on the acoustics as Equation 1.

$$P_{CTC}(y|x) = \sum_{\hat{y}=\beta(y,x)} P(\hat{y}|x) = \sum_{\hat{y}=\beta(y,x)} \prod_{t=1}^T P(\hat{y}_t|x) \quad (1)$$

With the conditional independent assumption, $P_{CTC}(\hat{y}|x)$ can be decomposed into a product of posterior $P(\hat{y}_t|x)$ in each frame t . The conditional probability of the labels at each frame, $P_{CTC}(\hat{y}_t|x)$, can be estimated using BLSTM, which we refer to as the encoder. The model can be trained to maximize Equation 1 by using gradient descent, where the required gradients can be computed using the forward-backward algorithm [5].

CTC models have a conditional independence assumption on its outputs, wherein it will become difficult to model the interdependencies between words. During the beam search process, language model and word count are introduced. The beam search process of CTC [20] is to find

$$\arg \max_y (\log(P_{CTC}(y|x)) + \alpha \log(P_{LM}(y)) + \beta \text{wordcount}(y)) \quad (2)$$

where a language model and word count are included, and α and β are the weights of them respectively.

2.2. Attention based models

Chan et al. [16] proposed Listen, Attend and Spell (LAS), a kind of neural network that learns to transcribe speech utterances to characters. As an attention-based encoder-decoder network, LAS is often used to deal with variable length input and output sequences. Using the attention mechanism, the attention model can align the input and output sequence.

As section 2.1 mentioned, the CTC assumes monotonic alignment, and it explicitly marginalizes over alignments. And because of the conditional independence assumption, the CTC model can not explicitly learn co-articulation patterns, which exist in speech commonly. Attention based models remove the conditional independence assumption in the label sequence that CTC requires, then the $p(y|x)$ defines as Equation 3

$$P_{Attention}(y|x) = P(y|h) = \prod_{t=1}^T P(y_t|c_t, y_{<t}) \quad (3)$$

where c_t is the context at decoding time step t .

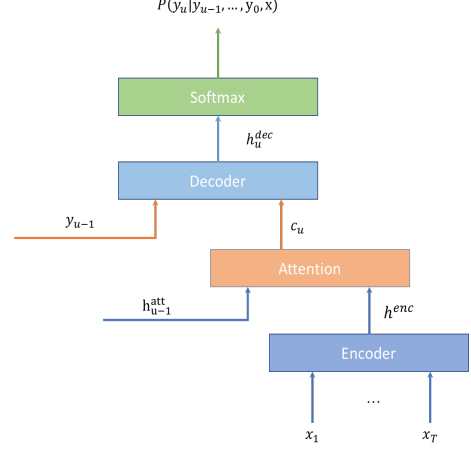


Figure 1: Schematic diagram of the attention-based encoder-decoder network.

An attention-based model contains an encoder network and an attention based decoder network. The encoder network maps the input acoustics into a higher-level representation. The attention based decoder network predicts the next output symbol conditioned on the full sequence of previous predictions and acoustics, which can be defined as $P(y_u | y_{u-1}, \dots, y_1, x)$. The attention mechanism selects or weights the input frames to generate the next output label.

As shown in Figure 1, the attention-based encoder-decoder network can be defined as:

$$h = \text{Encoder}(x) \quad (4)$$

$$P(y_u | x, y_{1:t-1}) = \text{AttentionDecoder}(h, y_{1:t-1}) \quad (5)$$

where $\text{Encoder}(\cdot)$ can be long short-term memory(LSTM) or bidirectional LSTM (BLSTM) and $\text{AttentionDecoder}(\cdot)$ can be LSTM or gated recurrent unit(GRU).

The beam search process of attention is to find

$$\arg \max_y (\log(P_{Att}(y|x))/|y|^\gamma + \beta \text{cov}(\alpha) + \lambda \log(P_{LM}(y))) \quad (6)$$

where γ is the length normalization hyperparameter. The coverage term "cov" encourages the model to attend over all encoder time steps, and stops rewarding repeated attendance over the same time steps. The coverage term addresses both short as well as infinitely long decoding.


3. Acoustic Modeling Units

In mandarin speech recognition, modeling units of acoustic model affect the performance significantly. There have been kinds of different acoustic representations for Mandarin in recent years [18, 19, 21]. For example, There have been syllable initial/final approach, syllable initial/final with tone approach, syllable approach, syllable with tone approach, Chinese Character approach and preme/toneme approach [22]. In this study, we select context dependent syllable initial/final with tone, syllable with tone and Chinese Character as study object. Figure 2 shows an example of various modeling units.

3.1. Context Dependent Phoneme (CDP)

For CTC based end-to-end speech recognition in mandarin, CDP is commonly used as the acoustic modeling unit. We usu-

CDP: sil-d+a4 d-a4+j a4-j+ia1
j-ia1+h ia1-h+ao3 h-ao3+sil

Utterance: 大家好  Syllable with tone: da4 jia1 hao3

Character: 大 家 好

Figure 2: An example of converting one Chinese utterance into various modeling units.

ally use syllable initial/final with tone as phoneme, such as syllable initial *d* and syllable final with tone *a4*, and the context dependent phoneme is like *sil-d+a4*.

3.2. Syllable

A syllable with tone consists of a syllable initial and a syllable final with tone, such as *da4*. Chinese is naturally a syllabic language and each basic language unit (Chinese character) can be phonetically represented by a syllable [23]. Furthermore, each Chinese syllable also has syllable Initial-Final structure. According to the official released scheme for Chinese phonetic alphabet, each syllable is regarded as the combination of these aspects are very helpful for the design of acoustic models.

3.3. Character

Like English words, Chinese characters are the basic symbols of the recording language. In most cases, in mandarin speech recognition, our goal is to transcribe the speech sequence into the Chinese Character sequence. Therefore, in the end-to-end speech recognition framework, Chinese character is a perfect modeling unit which can be decoded without language model and lexicon.

There is no exact number of Chinese characters, the number is about one hundred thousand, and there are only a few thousand characters in the daily use of Chinese characters. In our work, We chose 4977 common Chinese characters and the coverage is 99.92% on our datasets.

4. Experiments

Several experiments have been done to compare the performance of three kinds of acoustic modeling units by the two types of end-to-end methods. We find that both On the DidiReading dataset and DidiCallcenter dataset the Character based attention models achieve the best performance.

4.1. Data

We do all the experiments both on DidiCallcenter dataset and DidiReading dataset, which are different not only on the data size but also the dialogue scene. The DidiCallcenter dataset contains more than 2.2M utterances (about 2,800 hours), which is a spontaneous style dataset. The DidiReading dataset contains more than 16.2M utterances (about 12,000 hours), which is a reading style dataset. There are also two test sets, which are randomly extracted from the two datasets respectively. The DidiCallcenter test set includes 2000 utterances and the DidiReading test set includes 5000 utterances. 40 mel-scale filterbanks coefficients computed every 10ms are used as input fea-

Table 1: Detailed composition of various labels

Models	Modeling Units	Composition of the label
CTC	CDP	DidiCallcenter: $\sim 12,100$ CDP + 1 BLANK(<i>b</i>) DidiReading: $\sim 12,200$ CDP + 1 BLANK(<i>b</i>)
	Syllable	1313 tonal syllable with tone + 1 BLANK(<i>b</i>)
	Character	4977 Chinese character + 1 BLANK(<i>b</i>)
Attention	Syllable	1313 tonal syllable with tone + 1 unknown token(<i>unk</i>) + 1 sentence start token(<i>sos</i>) + 1 sentence end token(<i>eos</i>)
	Character	4977 Chinese characters + 1 unknown token(<i>unk</i>) + 1 sentence start token(<i>sos</i>) + 1 sentence end token(<i>eos</i>)

tures for both datasets. Global mean and variance normalization is conducted for each dataset.

Table 1 shows the detailed information of the labels for various modeling units.

4.2. CTC models

In this work, CTC models are trained to predict CDP, syllable and character as output targets, respectively.

4.2.1. Training

The network architecture of CTC is described in Figure 3, which contains one convolutional-2D layer, two residual blocks [24], four LSTM [25] layers and one full-connection layer. Each residual block includes two convolutional-2D layers. Each LSTM layer contains 1024 nodes and followed by layer normalization. The parameters of CDP-CTC, Syllable-CTC and Character-CTC model are about 86M, 30M, 46M, respectively. During training stage, Adam [26] optimization method is used and L2 weight decay is $1e-5$, the learning rate is decayed from $1e-3$ to $1e-6$ during training.

4.2.2. Decoding

These models are decoded using external 4-gram Chinese word language models. For DidiCallcenter task, the size of language model is 40GB which contains 1.9G gram tokens, on the other DidiReading task, we use a 55GB language model which contains 2.7G gram tokens.

4.3. Attention models

In this work, attention models are trained to predict syllable and Character as output targets, respectively.

4.3.1. Training

For syllable attention experiments, our models are LAS models with 2 convolutional layers, followed by 4 bi-directional LSTM layers with 256 LSTM units per-direction, interleaved with 3 time-pooling layers which resulted in an 8-fold reduction of the input sequence length. The Decoder was a 1 layer LSTM with 256 LSTM units and output has 1316 labels. For Character ex-

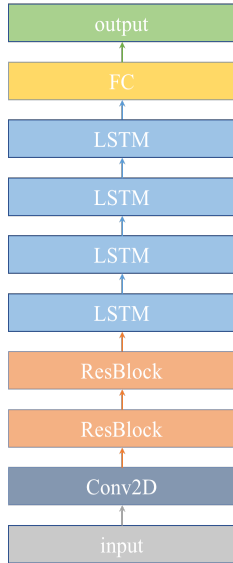


Figure 3: The network architectures of CTC model.

periments, our LAS models has the same architecture as the Syllable model, except that the output has 4980 labels. The syllable attention model has about 8.79M parameters and character attention model has about 12.54M parameters. During training stage, schedule sampling and unigram label smoothing is applied as described in [7, 10, 11]. Adam optimization method with gradient clipping is used for optimization. We initialized all the weights randomly from an isotropic Gaussian distribution with variance 0.1 and learning rate is decayed from $5e-4$ to $5e-6$ during training. All models are trained with the cross-entropy criterion and are trained using TensorFlow [27].

4.3.2. Decoding

A left-to-right beam search over modeling unit sequences was used during decoding. Beam search was stopped when the sentence end token $\langle eos \rangle$ was emitted. We also integrated external language models during decode stage and all the language models were trained with the training transcripts.

4.4. Results

We first conduct experiments on different modeling unit for CTC model. From Table 2, we can find that all modeling units can achieve similar CER but syllable based CTC model achieves the best performance both on the DidiCallcenter dataset and DidiReading dataset. Meanwhile, because syllable-based model has much less parameters than CDP-based model, the time model needed to converge is much less and decoding is much faster. Therefore, we believe that syllable is a more suitable modeling unit for CTC model.

Then, we compare the results of attention-based models in Table 3. By comparing the performance of the syllable attention model and character attention model, it's clear to see that the performance of character-model is better than syllable-model. We believe it's because the language model we're using isn't strong enough and the implicit RNN language model decoder learned helps to boost the performance of character-model. At the same time, we find external language model to be helpful for both of our tasks. But DidiReading task benefits more from

Table 2: CER(%) of CTC-based method on various modeling units (#Param is the number of model parameters)

Models	#Param	Callcenter	Reading
CDP-CTC	86.05M	7.42	5.81
Syllable-CTC	30.04M	7.31	5.62
Character-CTC	46.10M	7.45	5.79

external language model, we think it's because the DidiReading task is a domain-specific task which contains a lot of special terms, and the external language model can help solve this problem effectively.

Table 3: CER(%) of Attention-based method on various modeling units (#Param is the number of model parameters)

Models	Units	#Param	Callcenter	Reading
Attention + LM	Syllable	8.79M	-	-
	Syllable	8.79M	6.34	5.78
Attention + LM	Character	12.54M	5.86	6.22
	Character	12.54M	5.68	4.89

A comparison of the CTC models and attention models reveals some interesting conclusions. First, we note that the performance of attention model is significantly better than CTC model. On the DidiCallcenter task, the CTC model achieves a best CER of 7.31%, by using the character attention model, we improved the CER to 5.68%. In the same way, we improved the CER from 5.62% to 4.89% on the DidiReading task. It is also interesting to compare the structure of the two types of models, our CTC models used un-directional LSTM and attention models used bi-directional LSTM, but the number of parameters in CTC models is three times as much as attention models. In future work, we hope to compare a bigger attention model and bi-directional LSTM CTC model which has the same number of parameter as attention model.

5. Conclusions

In this work, we studied the performance of various acoustic modeling units on different end-to-end models in Mandarin speech recognition.

In experimental evaluations, we find that on CTC model, syllable achieve the best CER on both DidiCallcenter dataset and DidiReading dataset. Moreover, we find that all modeling units can achieve approximate CER in CTC model. On attention model, however, we find that character model outperforms syllable model. Namely, Chinese character is a appropriate modeling unit for acoustic modeling. Finally, we compare two end-to-end models and find that attention model is much better than CTC model in Mandarin speech recognition, even if the size of the attention model is much smaller than CTC model.

6. Acknowledgements

The authors would like to thank Liqiang He, Caixia Gong, Xiaohui Li and Mingxin Liang for their help.

7. References

- [1] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, “Recent advances in deep learning for speech research at microsoft,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8604–8608.
- [2] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.
- [3] Y. Miao, M. Gowayyed, and F. Metze, “Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding,” in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 167–174.
- [4] 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 *International Conference on Machine Learning*, 2016, pp. 173–182.
- [5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.
- [6] S. Kim, T. Hori, and S. Watanabe, “Joint ctc-attention based end-to-end speech recognition using multi-task learning,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4835–4839.
- [7] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, “State-of-the-art speech recognition with sequence-to-sequence models,” *arXiv preprint arXiv:1712.01769*, 2017.
- [8] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *International Conference on Machine Learning*, 2014, pp. 1764–1772.
- [9] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, “A comparison of sequence-to-sequence models for speech recognition,” in *Proc. Interspeech*, 2017, pp. 939–943.
- [10] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.
- [11] J. Chorowski and N. Jaitly, “Towards better decoding and language model integration in sequence to sequence models,” *arXiv preprint arXiv:1612.02695*, 2016.
- [12] W. Chan and I. Lane, “On online attention-based speech recognition and joint mandarin character-pinyin training,” in *INTER-SPEECH*, 2016, pp. 3404–3408.
- [13] C. Shan, J. Zhang, Y. Wang, and L. Xie, “Attention-based end-to-end speech recognition in mandarin,” *arXiv preprint arXiv:1707.07167*, 2017.
- [14] 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.
- [15] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm,” *arXiv preprint arXiv:1706.02737*, 2017.
- [16] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *ICASSP*, 2016.
- [17] C. Shan, J. Zhang, Y. Wang, and L. Xie, “Attention-based end-to-end speech recognition on voice search.”
- [18] E. Chang, J. Zhou, S. Di, C. Huang, and K.-F. Lee, “Large vocabulary mandarin speech recognition with different approaches in modeling tones,” in *Sixth International Conference on Spoken Language Processing*, 2000.
- [19] L. Xiangang, Y. Yuning, P. Zaihu, and W. Xihong, “A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary chinese speech recognition,” *Neurocomputing* 170 (2015) 251–256, 2015.
- [20] S. Wiseman and A. M. Rush, “Sequence-to-sequence learning as beam-search optimization,” *arXiv preprint arXiv:1606.02960*, 2016.
- [21] J. Li, F. Zheng, and J. Zhang, “Context dependent initial/final acoustic modeling for continuous chinese speech recognition,” *JOURNAL-TSINGHUA UNIVERSITY*, vol. 44, no. 1, pp. 61–64, 2004.
- [22] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, “New methods in continuous mandarin speech recognition,” in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [23] H. Wu and X. Wu, “Context dependent syllable acoustic model for continuous chinese speech recognition,” in *Eighth Annual Conference of the International Speech Communication Association*, 2007.
- [24] Y. Zhang, W. Chan, and N. Jaitly, “Very deep convolutional networks for end-to-end speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4845–4849.
- [25] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.
- [26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [27] 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.