Investigating Learning Dynamics of BERT Fine-Tuning

Yaru Hao[†]*, **Li Dong**[‡], **Furu Wei**[‡], **Ke Xu**[†]
[†]Beihang University
[‡]Microsoft Research
{haoyaru@,kexu@nlsde.}buaa.edu.cn
{lidong1,fuwei}@microsoft.com

Abstract

The recently introduced pre-trained language model BERT advances the state-of-the-art on many NLP tasks through the fine-tuning approach, but few studies investigate how the fine-tuning process improves the model performance on downstream tasks. In this paper, we inspect the learning dynamics of BERT fine-tuning with two indicators. We use JS divergence to detect the change of the attention mode and use SVCCA distance to examine the change to the feature extraction mode during BERT fine-tuning. We conclude that BERT fine-tuning mainly changes the attention mode of the last layers and modifies the feature extraction mode of the intermediate and last layers. Moreover, we analyze the consistency of BERT fine-tuning between different random seeds and different datasets. In summary, we provide a distinctive understanding of the learning dynamics of BERT fine-tuning, which sheds some light on improving the fine-tuning results.

1 Introduction

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2019) is a large pre-trained language model. It obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks. Unlike other previous pre-trained language models (Peters et al., 2018a; Radford et al., 2018), BERT employs the multi-layer bidirectional Transformer encoder as the model architecture and proposes two novel pre-training tasks: the masked language modeling and the next sentence prediction.

There are two approaches to adapt the pre-trained language representations to the downstream tasks. One is the feature-based approach, where the parameters of the original pre-trained model are frozen when applied on the downstream tasks (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018a). Another one is the fine-tuning approach, where the pre-trained model and the task-specific model are trained together (Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018). Take the classification task as an example, the new parameter added for BERT fine-tuning is a task-specific fully-connected layer, then all parameters of BERT and the classification layer are trained together to minimize the loss function.

Peters et al. (2019) demonstrate that the fine-tuning approach of BERT generally outperforms the feature-based approach. We know that BERT encodes task-specific representations during fine-tuning, but it is unclear about the learning dynamics of BERT fine-tuning, i.e., how fine-tuning helps BERT to improve performance on downstream tasks.

We investigate the learning dynamics of BERT fine-tuning with two indicators. First, we use Jensen-Shannon divergence to measure the change of the attention mode during BERT fine-tuning. Second, we use Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al. (2017)) distance to measure the change of the feature extraction mode.

We conclude that during the fine-tuning procedure, BERT mainly changes the attention mode of the last layers, and modifies the feature extraction mode of intermediate and last layers. At the same time, BERT has the ability to avoid catastrophic forgetting of knowledge in low layers. Moreover, we also analyze the consistency of the fine-tuning procedure. Across different random seeds and different datasets, we observe that the changes of low layers (0-9th layer) are generally consistent, which indicates that BERT has learned some common transferable language knowledge in low layers during the pre-training process, while the task-specific

*Contribution during internship at Microsoft Research.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 87–92

information is mostly encoded in intermediate and last layers.

2 Experimental Setup

We employ the BERT-large model[1] on a diverse set of NLP tasks: natural language inference (NLI), sentiment analysis (SA) and paraphrase detection (PD).

For NLI, we use both the Multi-Genre Natural Language Inference dataset (MNLI; Williams et al. 2018) and the Recognizing Textual Entailment dataset (RTE; aggregated from Dagan et al. 2006, Haim et al. 2006, Giampiccolo et al. 2007, Bentivogli et al. 2009). For SA, we use the binary version of the Stanford Sentiment Treebank dataset (SST-2; Socher et al. 2013). For PD, we use the Microsoft Research Paraphrase Corpus dataset (MRPC; Dolan and Brockett 2005).

Dataset	LR	BS	NE
MNLI	3e-5	64	3
RTE	1e-5	32	5
SST-2	3e-5	64	4
MRPC	1e-5	16	5

Table 1: Hyperparameter configuration for BERT fine-tuning. LR: learning rate, BS: batch size, NE: number of epochs.

The hyperparameter choice for fine-tuning is task-specific. We choose relatively optimal parameters for every dataset as suggested in Devlin et al. (2019). The detailed hyperparameter configuration is shown in Table 1. Moreover, we use Adam optimizer with the slanted triangular learning rate schedule (Howard and Ruder, 2018) and keep the dropout probability at 0.1.

3 Fine-tuning changes the attention mode of the last layers

The model architecture of BERT is essentially based on the multi-layer bidirectional Transformer, the core function of which is the self-attention mechanism (Vaswani et al., 2017). We use Jensen-Shannon divergence between two attention scores to detect changes of the attention mode in different layers during fine-tuning.

Jensen-Shannon divergence JS divergence is a method of measuring the distance between two probability distributions, it is defined as:

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||R) + \frac{1}{2}D_{KL}(Q||R)$$

where P and Q are two different probability distributions, $R = \frac{P+Q}{2}$ is the average probability distribution of them and D_{KL} represents the Kullback-Leibler divergence.

For every layer of BERT, there are 16 attention heads, each head produces an attention score of the input sequence. Each attention score is a probability distribution about how much attention a target word pays to other words. We compute JS divergence of attention scores between the original BERT model M_0 and the fine-tuned model M_t on the development set, by calculating the average of the sum of JS divergence at each word and each attention head for every layer, the specific calculation formula is as follows:

$$D_{JS}(M_t||M_0) = \frac{1}{N}\frac{1}{H}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{W}\sum_{i=1}^{W} D_{JS}(A_t^h(word_i)||A_0^h(word_i))$$

where N denotes the number of development examples, H denotes the number of attention heads, W denotes the number of tokens in a sequence and $A_t^h(word_i)$ denotes the attention score of the attention head h at $word_i$ in model M_t.

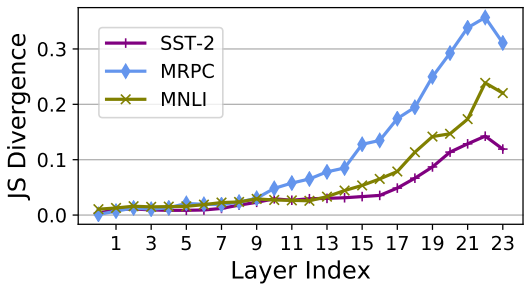

Figure 1: JS divergence of attention scores of every layer between the original BERT model and the fine-tuned model.

We present JS divergence results in Figure 1, from which we observe the attention mode in low layers and intermediate layers do not change seriously, while the attention mode of last layers changes drastically. It indicates that the fine-tuning procedure has the ability to keep the attention mode of low layers consistent with the original BERT model, and changes the attention mode of the last layers to adapt BERT on specific tasks.

[1] github.com/google-research/bert

4 Fine-tuning modifies the feature extraction mode of the intermediate and the last layers

While the attention score implies the inherent dependencies between different words, the output representation of every layer is the practical feature that the model extracts. We use SVCCA distance (Raghu et al., 2017) to quantify the change of these output representations during fine-tuning, which indicates the change of the feature extraction mode of BERT.

Singular Vector Canonical Correlation Analysis. SVCCA distance is used as a metric to measure the differences of hidden representations between the original BERT model M_0 and the fine-tuned model M_t at a target layer. It is calculated by:

$$D_{SVCCA}(M_t||M_0) = 1 - \frac{1}{c}\sum_{i=1}^{c}\rho^{(i)}$$

where c denotes the hidden size of BERT, ρ is the Canonical Correlation Analysis (CCA) resulting in a value between 0 and 1, which indicates how well correlated the two representations derived by two models are. For a detailed explanation of SVCCA, please see Raghu et al. (2017).

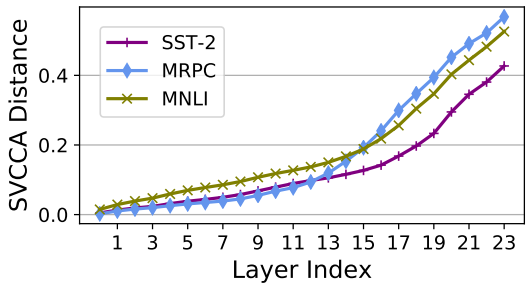

Figure 2: SVCCA distance of individual layers between the original BERT model and the fine-tuned model.

From Figure 2, we observe that changes in SVCCA distance in higher layers are more distinct than lower layers. This phenomenon is reasonable because the output representation of higher layers undergoes more transformations, so the change of SVCCA distance in higher layers is more dramatic.

As the output representation of the last layer is directly used for classification, we aim to compare the effect of each layer on the final output representation respectively. We replace the parameters of every layer in the fine-tuned model with their original values in the BERT model before fine-tuning and compute the SVCCA distance of the last layer output representation. The results are shown in Figure 3, we observe that whether the low layers (0-10) are replaced with their original values or not, it has little effect on the final output representation. Moreover, the change in the intermediate and last layers will increase the SVCCA distance, which reflects that fine-tuning mainly changes the feature extraction mode of intermediate and last layers.

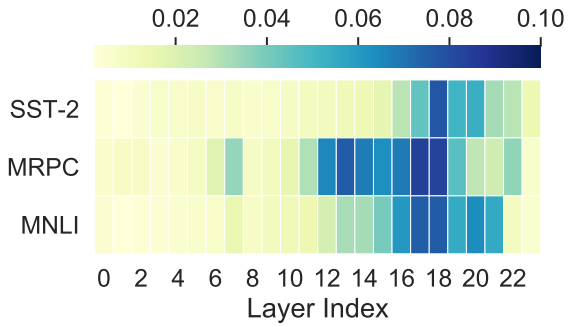

Figure 3: SVCCA distance of the last layer between the original fine-tuned model and the fine-tuned model with parameters of a target layer replaced with their pre-trained values.

5 Consistency of Fine-tuning

In this section, we investigate the consistency of different fine-tuning procedures, including the consistency between different random seeds and the consistency between different datasets.

5.1 Consistency between different random seeds

We fine-tune two models on every dataset with the same hyperparameters but different random seeds. We compute the pairwise JS divergence and SVCCA distance of each layer between the two models with different random seeds.

As shown in Figure 4, for large dataset MNLI and SST-2, the attention mode of low and intermediate layers is basically consistent between two different random seeds, whereas the attention mode of last layers is relatively divergent. For MRPC, the attention mode appears to be divergent at the 9th layer.

Figure 5 illustrates SVCCA distance between different random seeds, we observe that the SVCCA distance gradually increases in all layers. For MNLI and SST-2, the increase of last layers is

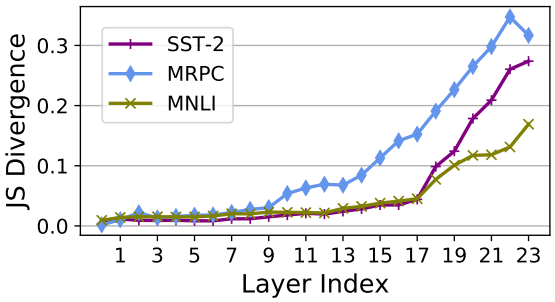

Figure 4: JS divergence between two models with different random seeds.

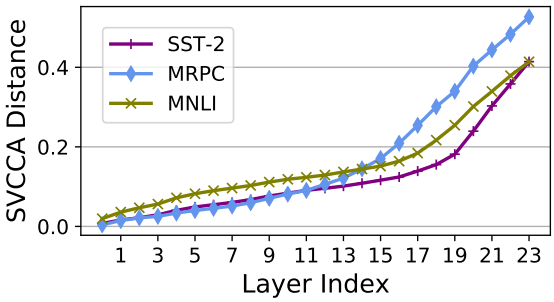

Figure 5: SVCCA distance between two models with different random seeds.

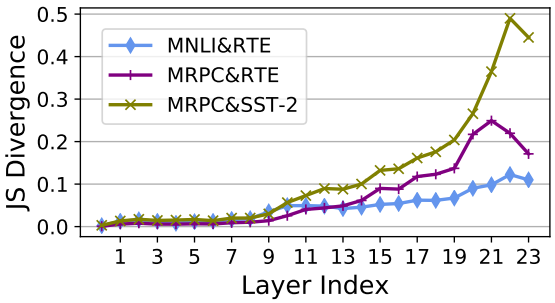

Figure 6: JS divergence between different datasets.

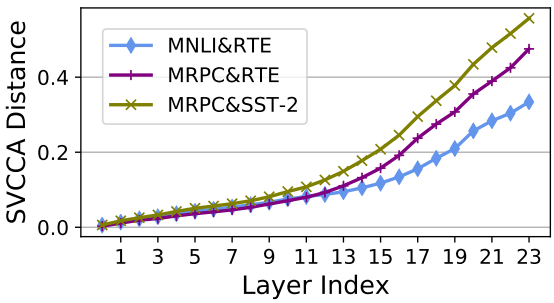

Figure 7: SVCCA distance between different datasets.

more obvious, and for MRPC, the increase appears to be obvious from the 13th layer.

5.2 Consistency between different datasets

Besides the consistency between different random seeds, we also aim to investigate the consistency between different datasets. We fine-tune two models on two different datasets then evaluate on a combined dataset containing 200 examples respectively from both two datasets.

For different datasets of the same domain, we use two models fine-tuned on RTE and MNLI dataset. For different domains, we examine the consistency between MRPC and RTE, which both have pairwise input sequences, and the consistency between MRPC and SST-2, which have different patterns of input sequences. The JS divergence results and SVCCA distance results between different datasets are shown in Figure 6 and Figure 7.

Figure 6 and Figure 7 demonstrate that no matter two datasets are from the same domain or the different domain, the attention mode and the feature extraction mode of low layers (0-7 layer) are consistent, which indicates BERT studies some common language knowledge during the pre-training procedure and low layers are stable to change their original modes. JS divergence of the attention scores and SVCCA distance of the output representations in intermediate and last layers between two models are more distinct when the difference between two training datasets increases. The consistency between datasets from similar tasks like RTE and MNLI is still relatively strong in last layers compared to the consistency between datasets from the different domain. And when the input sequence pattern and the domain of two datasets are different, the consistency of intermediate and last layers is weak as expected.

6 Related Work

Pre-trained language models (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Dong et al., 2019; Yang et al., 2019; Clark et al., 2020; Bao et al., 2020) stimulate the research interest on the interpretation of these black-box models. Peters et al. (2018b) show that the biLM-based models learn representations that vary with network depth, the lower layers specialize in local syntactic relationships and the higher layers model longer range relationships. Kovaleva et al. (2019) propose a methodology and offer the analysis of BERTs capacity to capture different kinds of linguistic information by encoding it in its self-attention weights. Hao et al. (2019) visualize the loss landscapes and

optimization trajectories of the BERT fine-tuning procedure and find that low layers of the BERT model are more invariant and transferable across tasks. Merchant et al. (2020) find that fine-tuning primarily affects the top layers of BERT, but with noteworthy variation across tasks. Hao et al. (2020) propose a self-attention attribution method to interpret information flow within Transformer.

7 Discussions

We use JS divergence to detect the change of the attention mode in different layers during BERT fine-tuning and use SVCCA distance to detect the change of the feature extraction mode. We observe that BERT fine-tuning mainly changes the attention mode of last layers and modifies the feature extraction mode of intermediate and last layers.

We also demonstrate that the changes of low layers are consistent between different random seeds and different datasets, which indicates that BERT learns common transferable language knowledge in low layers. In future research, we would like to explore learning dynamics for cross-lingual pre-trained models (Conneau and Lample, 2019; Conneau et al., 2020; Chi et al., 2020).

Acknowledgements

The work was partially supported by National Natural Science Foundation of China (NSFC) [Grant No. 61421003].

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-masked language models for unified language model pre-training. *arXiv preprint arXiv:2002.12804*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth pascal recognizing textual entailment challenge. In *In Proc Text Analysis Conference (TAC09*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *CoRR*, abs/2007.07834.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. *CoRR*, abs/1511.01432.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4143–4152, Hong Kong, China. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Self-attention attribution: Interpreting information interactions inside transformer. *CoRR*, abs/2004.11207.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to bert embeddings during fine-tuning?

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.