ADVERSARIAL MASK TRANSFORMER FOR SEQUENTIAL LEARNING

Hou Lio   *Shang-En Li*   *Jen-Tzung Chien*

Dept of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan

ABSTRACT

Mask language model has been successfully developed to build a transformer for robust language understanding. The transformer-based language model has achieved excellent results in various downstream applications. However, typical mask language model is trained by predicting the randomly masked words and is used to transfer the knowledge from rich-resource pre-training task to low-resource downstream tasks. This study incorporates a rich contextual embedding from pre-trained model and strengthens the attention layers for sequence-to-sequence learning. In particular, an *adversarial mask* mechanism is presented to deal with the shortcoming of random mask and accordingly enhance the robustness in word prediction for language understanding. The adversarial mask language model is trained in accordance with a minimax optimization over the word prediction loss. The worst-case mask is estimated to build an optimal and robust language model. The experiments on two machine translation tasks show the merits of the adversarial mask transformer.

Index Terms— Adversarial learning, transformer, mask language model, sequential learning

1. INTRODUCTION

Deep neural network (DNN) based methods have shown great success in various sequential learning tasks, such as speech recognition, text classification, sentiment analysis and machine translation. However, some studies [1] have illustrated the vulnerability of DNN models due to small perturbations so that it is possible to deliberately fool a target model to produce the incorrect results. Deep learning is basically vulnerable to those adversarial examples which are intentionally crafted by replacing, scrambling and erasing characters [2] or words [3, 4] under certain semantic and syntactic constraints. These adversarial examples are invisible to humans but can easily fool DNN models. In order to avoid such an incorrectness and improve the robustness of a model, the adversarial training is feasible to enhance the model by finding consistent predictions even in the presence of disturbances. Adversarial training aims to improve the robustness and enhance the generalization of the model by building a model where the original examples and the adversarial examples are both considered. Recently, the researches on adversarial training in computer vision and natural language processing have been increased significantly [5, 6, 7]. However, the adversarial examples in language domain are more difficult to generate. One reason is that adding the subtle changes to find imperceptible adversarial sentences is much more difficult than that in computer vision. It is more likely to incorporate invisible disturbances in a given image. In addition, text data exist in a discrete space, and the word-level disturbances may significantly change the meaning of a text stream. The adversarial disturbances to words should be made as little as possible. Such a restriction turns out to impose a sparsity on the changes in language generation. Recently, the generation tasks are getting more and more important. These tasks aim to generate the mostly-grammatical natural sentences from diverse input data. However, training a robust text generation model usually requires a large-scale set of training data.

To overcome this issue, one possible way is to enlarge the training data through data augmentation using back translation [8] or other schemes. But, these methods take more than twice the training time. Another possible way is to introduce the uncertainty such as dropout and text changes to overcome the overfitting during training, especially in case of limited data. However, the contextual information in text changes using this adversarial training is not considered. The robustness to different contextual data is not assured. Because contextual information is essential to improve the robustness in sequential learning, this paper presents the adversarial training and incorporates the contextual robustness in a sequential model based on the transformer [9, 10]. A new architecture called the adversarial mask transformer (AMT) is proposed. This method is motivated by enhancing the mask language model. This AMT applies the adversarial word masks when computing the loss from adversarial samples, which forces the model to run machine generation with no specific word information. Adversarial mask aims to find out a corrupted context to make the sequence generation or prediction more difficult. The advantage of using this worst-case context is two-fold. First, with the dynamic mask strategy which generates a new mask pattern in each training step, we feed in a different sequence to the decoder so that the decoder can see diverse training instances. Second, in order to encode such a noise, the model has to learn contextual information from other words without doing mask. Such an approach would enhance the model robustness due to a large-span contextual information.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

2. BACKGROUND SURVEY

2.1. Mask language model

Pre-training in language model has been demonstrated as a powerful avenue to boost the performance of sequential learning tasks, such as question answering, language generation and text summarization. Through unsupervised learning from the massive unlabeled data, the pre-trained language models are able to learn the contextual representations of input streams, which are extremely helpful to accomplish downstream tasks. To learn the representations of the tokens in a sequence, the previous works [9, 11, 12] used the attention-based mechanism to retrieve meaningful contextual information. Specifically, the training instances are created by replacing a subset of the tokens in input sequence with a mask token, and the objective is to predict the masked tokens. The bidirectional encoder representations from transformers (BERT) [13], as the most widely used pre-trained language models, are trained by using two unsupervised schemes, namely, masked language model and next sentence prediction. During training, given an input sequence $X = \{\mathbf{x}_m\}_{m=1}^{T_i}$ with length T_i, the masked language model aims to calculate

$$p\left(\mathbf{x}_m | \mathbf{x}_1, \ldots, \mathbf{x}_{m-1}, [\text{mask}], \mathbf{x}_{m+1}, \ldots, \mathbf{x}_{T_i}\right) \qquad (1)$$

where [mask] is a masked token over the m-th word. Actually, a masked language model can recover multiple masks together. For simplicity, this is the case with only one masked token. Unlike the traditional language model that is in left-to-right order $p(\mathbf{x}_m|\mathbf{x}_1, \cdots, \mathbf{x}_{m-1})$, the masked language model is able to use both the left and the right contexts. By adding a few layers in the top of the model, a mask language model can be easily adapted into task-specific model, which is then fine-tuned by using the labeled data to achieve optimal performance. BERT has been widely used in various sequential learning and has achieved state-of-the-art results.

2.2. Adversarial learning

With the strength of generative adversarial network [14, 15, 16], adversarial learning has shown stunning results in the areas of computer vision especially in generation tasks. Comparing to the images, even though text data are discrete, the adversarial learning still receives high attentions in sequential learning tasks. Many research efforts have been taken to estimate the adversarial examples. These examples are basically obtained with small perturbations to training samples which are indistinguishable to humans but enough to produce misclassifications by a trained neural network. Recent works showed that augmenting examples to the training set can make a neural network model robust to perturbations. The work in [17] adapted the adversarial training to text classification and improved the performance on a few supervised and semi-supervised tasks. Such an adversarial learning is eligible to incorporate the adversarial examples to improve generalization. In general, there are two categories of approaches. The first is to keep the perturbations small such as a norm-bound on the gradient or replacing the words by their synonyms. This category of works aims to make the model robust to both source and target perturbations which are simulated by swapping the word embedding of a word with that of its synonym. Small perturbations are made by considering the word swapping which cause the smallest increase in loss gradient. In [17], the tasks on Chinese-English and English-German machine translations showed the desirable performance. On the other hand, a minimax formulation can be introduced where the adversarial examples are generated to maximize a loss function and the mode is trained to minimize the loss function. These considerations have motivated us to design an adversarial algorithm to generate a mask to perturb the actual text instead of adapting the embedding.

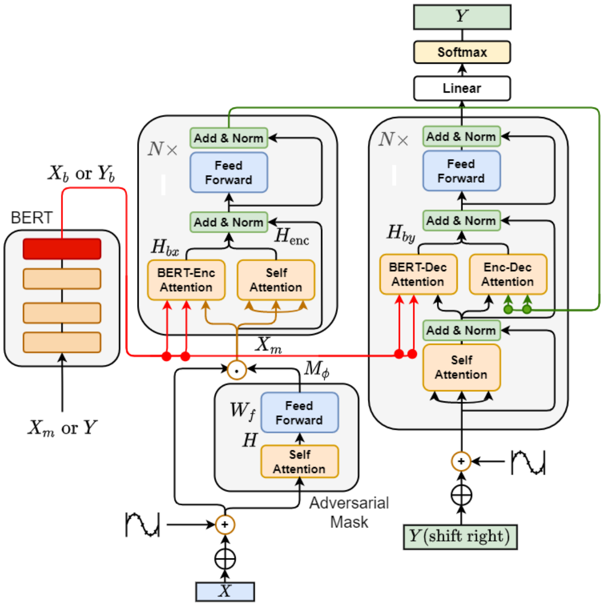


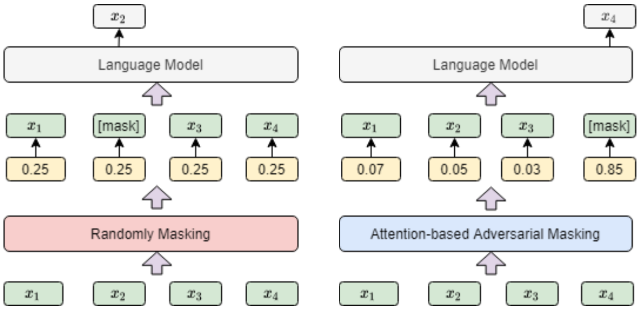
Fig. 1: Adversarial mask transformer for sequential learning from X to Y. Pre-trained BERT is merged to implement BERT enhanced encoder and decoder attentions.

3. ADVERSARIAL MASK TRANSFORMER

Adversarial mask transformer contains the BERT enhanced attention layers in both encoder and decoder as well as the adversarial mask module in the encoder as depicted in Figure 1 which is addressed in details as follows.

3.1. BERT enhanced attention layer

The attention mechanism [9, 18] works on a set of query vectors $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$, key vectors $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_m\}$, and value vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$, which are all embedded in the same dimension d. The query, key and value vectors are packed into matrices Q, K and V, respectively, where $Q \in \mathbb{R}^{n \times d}$ and $K, V \in \mathbb{R}^{m \times d}$. The attention weights are calculated by a softmax function and the attended context vector or a head vector is given by $\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$. There are two kinds attention weights in this transformer. One is the self-attention while the other is the cross-attention (or encoder-decoder attention). For the case of self-attention, the matrices Q, K and V are all projected from the same input matrix X using three different parameter matrices W^q, W^k and W^v, respectively. Self-attention is calculated to find the context vector $H = \mathrm{Attn}(W^q X, W^k X, W^v X)$ to fulfill multi-head attention. To enhance the contextual understanding in the source sequence X, this study calculates the large-span embedding from the last layer of a pre-trained BERT by $X_b = \mathrm{BERT}(X)$, and used it as the input to encoder to find the BERT head $H_{bx} = \mathrm{Attn}(W^q X, W^k X_b, W^v X_b)$. The same operation is run to find H_{by} by using BERT embedding Y_b of target sequence Y and applying it in decoder.



(a) Random mask

(b) Adversarial mask

Fig. 2: Random versus adversarial masks. Random mask randomly and uniformly selects the token for masking. Adversarial mask finds the masked token via adversarial learning.

3.2. Adversarial mask language model

Different from the mask language model using random mask, this paper presents a new transformer with the attention-based adversarial mask in Figure 2. Typically, random mask is likely to choose the commonly used words for masking because these words appear frequently. Instead of repeatedly choosing these common words such as prepositions and auxiliary verbs, this model adaptively masks those words which have considerable influence on the meaning of sentences. A learnable mask is estimated by selecting relatively important words for masking the input sequence X. Adversarial mask is run by $M_\phi = g_\phi(HW_f)$ using $X_m = M_\phi X$ where g_ϕ is the mapping function to find binary mask M_ϕ and W_f is the parameter of feedforward network with the outputs which are used to calculate the unnormalized log probability for different masked tokens. The parameters of language model θ and mask model ϕ are estimated via an adversarial learning

$$J(\theta, \phi) = \min_{\phi} \max_{\theta} \mathbb{E}_{X \sim p(X)}[p_\theta(X | X_m(M_\phi))]. \quad (2)$$

The worst-case mask is estimated to train the robust language model via a minimax optimization in term of prediction likelihood $p_\theta(X|X_m)$ or classification loss for word prediction.

3.3. Integration with transformer

Overall, the BERT enhanced attention layer and the adversarial mask module are combined in transformer to construct the adversarial mask transformer (AMT) for supervised sequence-to-sequence (seq2seq) learning. AMT is seen as a new transformer powered by the adversarial enhanced module which is utilized to map from input sequence $X = \{\mathbf{x}_m\}_{m=1}^{T_i}$ to output sequence $Y = \{\mathbf{y}_n\}_{n=1}^{T_o}$ with different lengths. Different from vanilla transformer, the encoder of AMT aims to learn the robust representation with an adversarial mask and extract the contextual information from the BERT enhanced attention layers during training. This new transformer carries out the enhanced multi-head attention in the encoder with parameter $\theta_e = \{W_m^q, W_m^k, W_m^v, W_b^q, W_b^k, W_b^v\}$ where $Q_m = W_m^q X_m$, $K_m = W_m^k X_m$, $V_m = W_m^v X_m$, $Q_b = W_b^q X_b$, $K_b = W_b^k X_b$, $V_b = W_b^v X_b$. The encoder head is integrated from the heads using X_m and X_b as

$$H_{\text{enc}} = \frac{1}{2}(\mathrm{Attn}(Q_m, K_m, V_m) + \mathrm{Attn}(Q_b, K_b, V_b)). \quad (3)$$

Then, a decoder with parameter θ_d is incorporated to predict $\mathbf{y}_n$ of the target sequence $\mathbf{y}_{1:T_o}$ or $\{\mathbf{y}_n\}_{n=1}^{T_o}$ sequentially by given the previous tokens $\mathbf{y}_{0:n-1}$. Segmentation is run for sequence-to-sequence learning in the self-attention layers of decoder or the cross-attention layers between encoder and decoder. The conditional likelihood for prediction of an output sample $\mathbf{y}_n$ of Y is calculated via the decoder or classifier

$$p(\mathbf{y}_n | \mathbf{y}_{0:n-1}, X) = \mathrm{Decoder}(\mathbf{y}_{0:n-1}, H_{\text{enc}}; \theta_d). \quad (4)$$

This AMT is finally trained by a hybrid objective including the adversarial mask and the classifier for sequence generation. The adversarial learning objective of using AMT for sequence-to-sequence learning is extended from Eq. (2) as

$$\begin{aligned} J(\theta_e, \theta_d, \phi) = & \min_{\phi} \max_{\theta_e, \theta_d} \mathbb{E}_{X \sim p(X)}[p_{\theta_e}(X | X_m(M_\phi))] \\ & + \mathbb{E}_{X,Y \sim p(X,Y)}[\log p_{\theta_e, \theta_d}(Y|X)]. \end{aligned} \quad (5)$$

4. EXPERIMENTS

This study conducted the evaluation on machine translation over different languages with various sizes of training data.

4.1. Experimental settings

IWSLT [19] and WMT datasets [20] were used to evaluate different machine translation models. IWSLT dataset contained about 200k pairs of sentences for training, 7K for validation, and 7K for test. The corpus consisted of transcriptions and their translations from TED talks. WMT dataset was a much bigger dataset compared to IWSLT. WMT had 4M pairs of sentences for training, 6K for validation, and 7K for test. We applied BPE [21] with 32K merge operations to segment words into subword units using these two datasets. For both IWSLT and WMT translation tasks, all attention models were built with 6 blocks. The word embedding dimension, hidden state dimension and number of heads were 512, 1024 and 8, respectively, the same as those in baseline models. The dropout rate of each attention and feed forward layers was set to 0.1. All models were optimized by using Adam optimizer with initial learning rate 1e-3. We chose the pre-trained $\text{BERT}_{\text{base}}$ for both IWSLT and WMT tasks, which confirmed that the dimensions of BERT and neural machine translation model were matched. The pre-trained weights of $\text{BERT}_{\text{base}}$ were provided by FAIRSEQ [22] framework, and the evaluations were also done by using FAIRSEQ framework.

4.2. Experimental results

In the experiment, the proposed AMT is compared with several strong baselines including the convolution seq2seq [23], transformer [9], weighted transformer [24], evolved transformer [25], BERT-fused model [26]. Weighted transformer replaces the multi-head attention by multiple self-attention branches so that the model learns to combine during the training process. Evolved transformer proposes the progressive dynamic hurdles method which is allowed to find the promising candidate models dynamically. BERT-fused model uses BERT to extract the representation of input sequence, and then introduces this embedding to each layer of encoder and decoder of a translation model through the attention mechanism. Tables 1 and 2 report the evaluation results using IWSLT and WMT datasets, respectively. In IWSLT English-to-German (En-De) and German-to-English (De-En), it is shown that the transformer trained with adversarial mask receives the highest BLEU. In WMT English-to-German (En-De), the adversarial mask transformer obtains the highest BLEU even in the translation for long sentences in WMT.

Table 3 reports the translation results using different mask language models (MLMs). The first baseline is BERT+LM, which uses the MLM in BERT to pre-train the encoder, and the standard language model (LM) to pre-train the decoder. The second baseline is the Mask-Predict [27] method. The resulting transformer architecture randomly masks the tokens in the sentence and decode them according to the confidence value in the iterative process. The third baseline is MASS [28]. MASS takes the sentence with the random and continuous mask as the input. MASS is basically consistent with the BERT. BERT randomly selected 15% of the words and used three different strategies including replace/mask/unchanged. MASS handles the balance between encoding and decoding for 50% of the total sentence length instead of random selection. The result shows that AMT achieves the highest BLEU even training with a small mask.

Model	En→De	De→En
ConvS2S [23]	26.1	31.9
Transformer [9]	28.6	34.4
Weighted Transformer [24]	28.9	35.1
Evolved Transformer [25]	30.4	36.0
BERT-fused model [26]	30.5	36.1
Adversarial Mask Transformer	**30.9**	**36.6**

Table 1: Comparison of BLEU scores using different methods on IWSLT En-De and De-En translation tasks.

Model	BLEU
ConvS2S [23]	25.2
Transformer [9]	26.2
Weighted Transformer [24]	27.2
Evolved Transformer [25]	28.4
BERT-fused model [26]	28.3
Adversarial Mask Transformer	**28.9**

Table 2: Comparison of BLEU scores using different methods on WMT En-De translation tasks.

Model	BLEU
BERT+LM [13]	24.9
Transformer with Mask-Predict [27]	27.7
MASS [28]	28.3
Adversarial Mask Transformer	**28.9**

Table 3: Comparison between different MLM-fused models on WMT En-De translation tasks.

5. CONCLUSIONS

This paper presented an approach to mask the important information in sentences. The masked sentence was used as the input to a new transformer, where the encoder was used to predict the masked words. We developed the adversarial learning to allow the model to learn different masks adaptively instead of random methods. By compensating the missing information and improving the model robustness, the ability to language understanding was improved accordingly. Experimental results showed that the proposed adversarial mask was more robust than random mask and the other models.

6. REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *Proc. of International Conference on Learning Representations*, 2014.

[2] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, “Black-box generation of adversarial text sequences to evade deep learning classifiers,” in *Proc. of IEEE Security and Privacy Workshops*, 2018, pp. 50–56.

[3] Z. Zhao, D. Dua, and S. Singh, “Generating natural adversarial examples,” in *Proc. of International Conference on Learning Representations*, 2018.

[4] C.-T. Chu, M. Rohmatillah, C. h. Lee, and J.-T. Chien, “Augmentation strategy optimization for language understanding,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[5] Y. Tu, M.-W. Mak, and J.-T. Chien, “Variational domain adversarial learning with mutual information maximization for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020.

[6] C.-E. Hsu, M. Rohmatillah, and J.-T. Chien, “Multitask generative adversarial imitation learning for multi-domain dialogue system,” in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 954–961.

[7] Y. Tu, M.-W. Mak, and J.-T. Chien, “Variational Domain Adversarial Learning for Speaker Verification,” in *Proc. of Annual Conference of International Speech Communication Association*, 2019, pp. 4315–4319.

[8] S. Edunov, M. Ott, M. Auli, and D. Grangier, “Understanding back-translation at scale,” in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2018, pp. 489–500.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[10] C.-H. Leong, Y.-H. Huang, and J.-T. Chien, “Online Compressive Transformer for End-to-End Speech Recognition,” in *Proc. of Annual Conference of International Speech Communication Association*, 2021, pp. 2082–2086.

[11] J.-T. Chien and W.-H. Chang, “Dualformer: a unified bidirectional sequence-to-sequence learning,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7718–7722.

[12] J.-T. Chien and W.-H. Chang, “Collaborative regularization for bidirectional domain mapping,” in *Proc. of International Joint Conference on Neural Networks*, 2021, pp. 1–8.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of Conference of North American Chapter of Association for Computational Linguistics*, 2019, pp. 4171–4186.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[15] J.-T. Chien and C.-L. Kuo, “Variational Bayesian GAN,” in *Proc. of European Signal Processing Conference*, 2019, pp. 1–5.

[16] J.-T. Chien and C.-W. Huang, “Stochastic adversarial learning for domain adaptation,” in *Proc. of International Joint Conference on Neural Networks*, 2020, pp. 1–7.

[17] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, “FreeLB: Enhanced adversarial training for natural language understanding,” in *Proc. of International Conference on Learning Representations*, 2020.

[18] J.-T. Chien and Y.-H. Chen, “Continuous-time attention for sequential learning,” in *Proc. of AAAI Conference on Artificial Intelligence*, 2021, pp. 7116–7124.

[19] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stuker, K. Sudoh, K. Yoshino, and C. Federmann, “Overview of the IWSLT 2017 evaluation campaign,” in *Proc. of International Workshop on Spoken Language Translation*, 2017, pp. 1–14.

[20] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. S., and M. Turchi, “Findings of the 2017 conference on machine translation,” in *Proc. of the Conference on Machine Translation.*, 2017, pp. 169–214.

[21] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2016, pp. 1715–1725.

[22] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, “FAIRSEQ: A fast, extensible toolkit for sequence modeling,” in *Proc. of Conference of North American Chapter of Association for Computational Linguistics*, 2019, pp. 48–53.

[23] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” in *Proc. of International Conference of Machine Learning*, 2017, pp. 1243–1252.

[24] K. Ahmed, N. S. Keskar, and R. Socher, “Weighted transformer network for machine translation,” *arXiv preprint arXiv:1711.02132*, 2017.

[25] D. R. So, Q. V. Le, and C. Liang, “The evolved transformer,” in *Proc. of International Conference of Machine Learning*, 2019, pp. 5877–5886.

[26] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu, “Incorporating BERT into neural machine translation,” in *Proc. of International Conference on Learning Representations*, 2020.

[27] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, “Mask-Predict: Parallel decoding of conditional masked language models,” in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 6111–6120.

[28] K. Song, X. Tan, T. Qin, J. Lu, and Tie-Yan Liu, “MASS: masked sequence to sequence pre-training for language generation,” in *Proc. of International Conference of Machine Learning*, 2019, pp. 5926–5936.