陽明交大
NYCU

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

#3301

Introduction

- 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.
- A new architecture called the adversarial mask transformer (AMT) is proposed. We present the adversarial training and incorporate the contextual robustness in a sequential model based on the transformer.

Mask Language Model

- 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

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

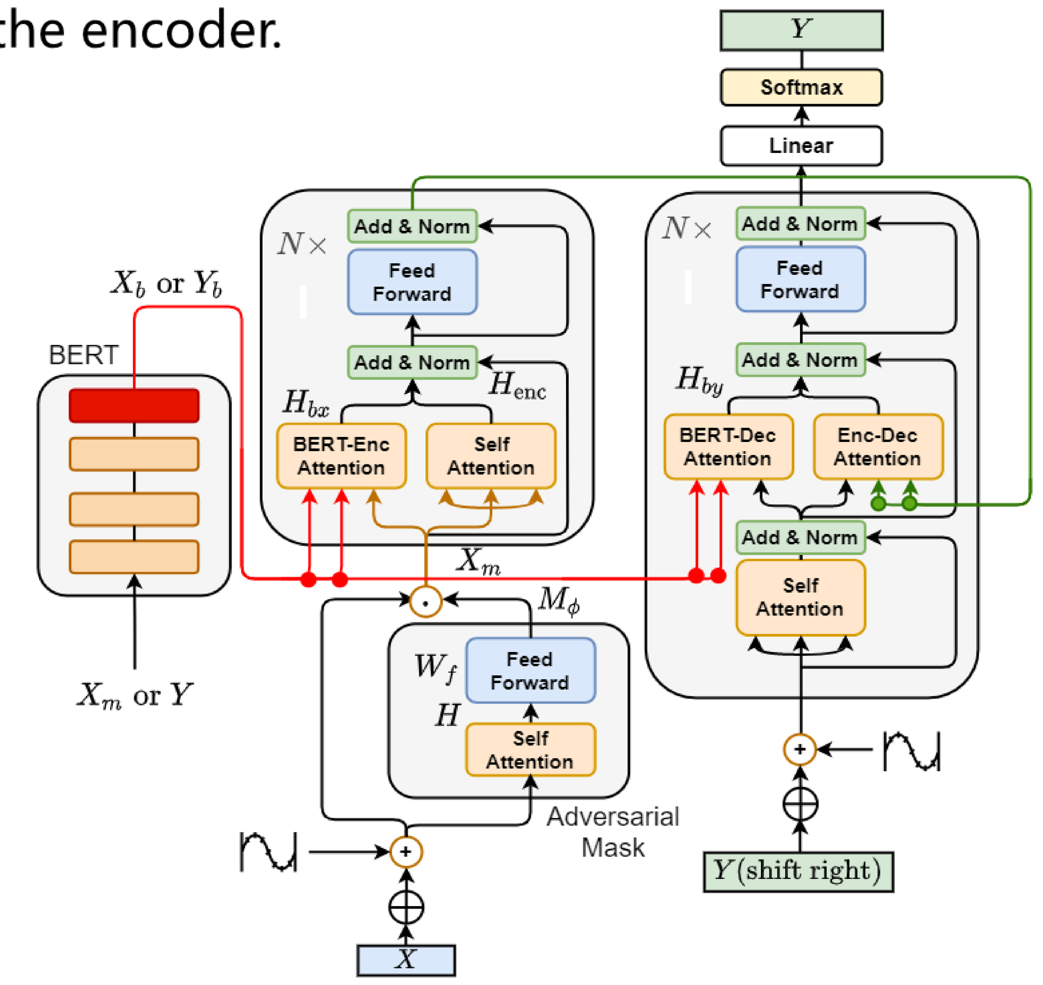
- Unlike the traditional language model that is in left-to-right order $p(\mathbf{x}_m | \mathbf{x}_1, \cdots, \mathbf{x}_{m-1})$, the masked language model is able to use both the left and the right contexts.
- A mask language model can be easily adapted into task-specific model, which is then fine-tuned by using the labeled data to achieve optimal performance.

Adversarial Learning

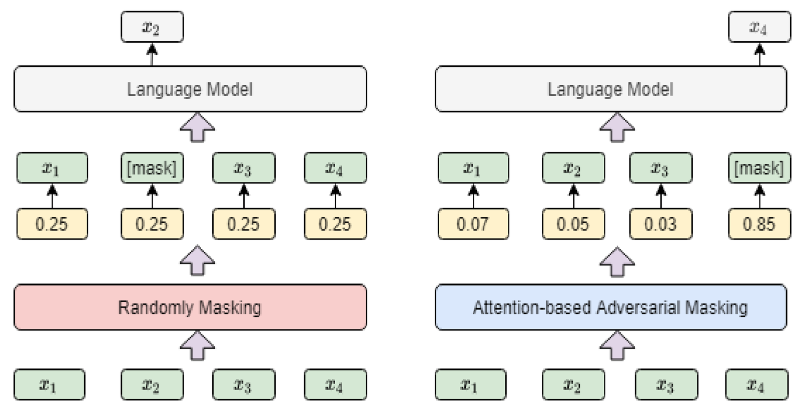
- Adversarial learning is eligible to incorporate the adversarial examples to improve generalization.
- A minimax formulation can be introduced where the adversarial examples are generated to maximize a loss function and the mode is trained to minimize the loss function.
- These considerations have motivated us to design an adversarial algorithm to generate a mask to perturb the actual text instead of adapting the embedding.

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.



- Different from the mask language model using random mask, we present a new transformer with the attention based adversarial mask.



- 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.

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

- The encoder head is integrated from the heads using X_m and X_b as

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

- The conditional likelihood for prediction of an output sample $\mathbf{y}_n$ of Y is calculated via the decoder or classifier

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

- The adversarial learning objective of using AMT for sequence-to-sequence learning is

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

Experiments

- This study conducted the evaluation on machine translation over different languages with various sizes of training data.
- IWSLT and WMT datasets were used to evaluate different machine translation models.
- The following two table report the evaluation results using IWSLT and WMT datasets, respectively.

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**

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**

- The following table reports the translation results using different mask language models (MLMs).

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

Conclusions

- We 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.