

Transkimmer: Transformer Learns to Layer-wise Skim

Yue Guan^{1,2,*}, Zhengyi Li^{1,2,*}, Jingwen Leng^{1,2,*}, Zhouhan Lin^{1,§} and Minyi Guo^{1,2,†}

¹Shanghai Jiao Tong University, ²Shanghai Qizhi Institute

*{bonboru, hobbit, leng-jw}@sjtu.edu.cn,

[§]lin.zhouhan@gmail.com, [†]guo-my@cs.sjtu.edu.cn

Abstract

Transformer architecture has become the de-facto model for many machine learning tasks from natural language processing and computer vision. As such, improving its computational efficiency becomes paramount. One of the major computational inefficiency of Transformer-based models is that they spend the identical amount of computation throughout all layers. Prior works have proposed to augment the Transformer model with the capability of skimming tokens to improve its computational efficiency. However, they suffer from not having effectual and end-to-end optimization of the discrete skimming predictor. To address the above limitations, we propose the Transkimmer architecture, which learns to identify hidden state tokens that are not required by each layer. The skimmed tokens are then forwarded directly to the final output, thus reducing the computation of the successive layers. The key idea in Transkimmer is to add a parameterized predictor before each layer that learns to make the skimming decision. We also propose to adopt reparameterization trick and add skim loss for the end-to-end training of Transkimmer. Transkimmer achieves $10.97\times$ average speedup on GLUE benchmark compared with vanilla BERT_{base} baseline with less than 1% accuracy degradation.

1 Introduction

The Transformer model (Vaswani et al., 2017) has pushed the accuracy of various NLP applications to a new stage by introducing the multi-head attention (MHA) mechanism (Lin et al., 2017). Further, the BERT (Devlin et al., 2019) model advances its performances by introducing self-supervised pre-training, and has reached the state-of-the-art accuracy on many NLP tasks.

Compared to the recurrent fashion models, e.g. RNN (Rumelhart et al., 1986), LSTM (Hochreiter and Schmidhuber, 1997), the Transformer model leverages the above attention mechanism to process

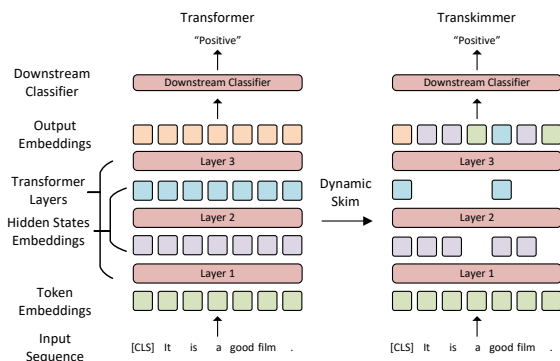


Figure 1: Overview of Transkimmer dynamic token skimming method. Tokens are pruned during the processing of Transformer layers. Note that actually we don't need all the tokens given to the downstream classifier in this sequence classification example. We show the full length output embedding sequence to demonstrate the forwarding design of Transkimmer.

all the input sequence. By doing so, extremely large scale and long span models are enabled, resulting in a huge performance leap in sequence processing tasks. However, the computation complexity of the attention mechanism is $O(N^2)$ with the input length of N , which leads to the high computation demand of the Transformer model.

Some prior works (Goyal et al., 2020; Kim and Cho, 2021; Kim et al., 2021; Ye et al., 2021) explore the opportunity on the dynamic reduction of input sequence length to improve the Transformer's computational efficiency. Its intuition is similar to the human-being's reading comprehension capability that does not read all words equally. Instead, some words are focused with more interest while others are skimmed. For Transformer models, this means adopting dynamic computation budget for different input tokens according to their contents. To excavate the efficiency from this insight, we propose to append a skim predictor module to the Transformer layer to conduct fine-grained dynamic token pruning as shown in Fig. 1. When processed by the Transformer layers, the sequence of token

hidden state embeddings are pruned at each layer with reference to its current state. Less relevant tokens are skimmed without further computation and forwarded to the final output directly. Only the significant tokens are continued for successive layers for further processing. This improves the Transformer model inference latency by reducing the input tensors on the sequence length dimension.

However, the optimization problem of such skim decision prediction is non-trivial. To conduct pruning of dynamic tensors, non-differentiable discrete skim decisions are applied. Prior works have proposed to use soft-masking approximation or reinforcement learning to resolve, which leads to approximation mismatch or nonuniform optimization. Transkimmer propose to adopt reparameterization technique (Jang et al., 2017) to estimate the gradient for skim prediction. As such, we can achieve the end-to-end joint optimization objective and training paradigm. By jointly training the downstream task and skim objective, the Transformer learns to selectively skim input contents. In our evaluation, we show Transkimmer outperforms all prior input reduction works on inference speedup gain and model accuracy. Specifically, BERT_{base} is accelerated for $10.97\times$ on GLUE benchmark and $2.81\times$ without counting the padding tokens. Moreover, we also demonstrate the method proposed by Transkimmer is generally applicable to pre-trained language models and compression methods with RoBERTa, DistillBERT and ALBERT models.

This paper contributes to the following 3 aspects.

- We propose the Transkimmer model which accelerates the Transformer inference with dynamic token skimming.
- We further propose an end-to-end joint optimization method that trains the skim strategy together with the downstream objective.
- We evaluate the proposed method on various datasets and backbone models to demonstrate its generality.

2 Related Works

Recurrent Models with Skimming. The idea to skip or skim irrelevant sections or tokens of input sequence has been studied in NLP models, especially recurrent neural networks (RNN) (Rumelhart et al., 1986) and long short-term memory network (LSTM) (Hochreiter and Schmidhuber,

1997). When processed recurrently, skimming the computation of a token is simply jumping the current step and keep the hidden states unchanged. LSTM-Jump (Yu et al., 2017), Skim-RNN (Seo et al., 2018), Structural-Jump-LSTM (Hansen et al., 2019) and Skip-RNN (Campos et al., 2018) adopt this skimming design for acceleration in recurrent models.

Transformer with Input Reduction. Unlike the sequential processing of the recurrent models, the Transformer model calculates all the input sequence tokens in parallel. As such, skimming can be regarded as the reduction of hidden states tensor on sequence length dimension. Universal Transformer (Dehghani et al., 2019) proposes a dynamic halting mechanism that determines the refinement steps for each token. DeFormer (Cao et al., 2020) proposes a dual-tower structure to process the question and context part separately at shallow layers specific for QA task. The context branch is pre-processed off-line and pruned at shallow layers. Also dedicated for QA tasks, Block-Skim (Guan et al., 2021) proposes to predict and skim the irrelevant context blocks by analyzing the attention weight patterns. Progressive Growth (Gu et al., 2021) randomly drops a portion of input tokens during training to achieve better pre-training efficiency.

Another track of research is to perform such input token selection dynamically during inference, which is the closest to our idea. POWER-BERT (Goyal et al., 2020) extracts input sequence at token level while processing. During the fine-tuning process for downstream tasks, Goyal et al. proposes a soft-extraction layer to train the model jointly. Length-Adaptive Transformer (Kim and Cho, 2021) improves it by forwarding the inflected tokens to final downstream classifier as recovery. Learned Token Pruning (Kim et al., 2021) improves POWER-BERT by making its pre-defined sparsity ratio a parameterized threshold. TR-BERT (Ye et al., 2021) adopts reinforcement learning to independently optimize a policy network that drops tokens. Comparison to these works are discussed in detail in Sec. 3. Moreover, SpAttn (Wang et al., 2021) facilitate POWER-BERT design with a domain-specific hardware design for better acceleration and propose to make skimming decisions with attention values from all layers.

Early Exit Early exit (Panda et al., 2016; Teerapittayanon et al., 2016) is another method to execute the neural network with input-dependent computational complexity. The idea is to halt the execution during model processing at some early exits. Under the circumstance of processing sequential inputs, early exit can be viewed as a coarse-grained case of input skimming. With the hard constraint that all input tokens are skimmed at the same time, early exit methods lead to worse accuracy and performance results compared to input skimming methods. However, the early exit method is also generally applicable to other domains like convolutional neural networks (CNN). DeeBERT (Xin et al., 2020), PABEE (Zhou et al., 2020), FastBERT (Liu et al., 2020) are some recent works adopting early exit in Transformer models. Magic Pyramid (He et al., 2021) proposes to combine the early exit and the input skimming ideas together. Tokens are skimmed with fine-grained granularity following POWER-BERT design and the whole input sequence is halted at some early exits.

Efficient Transformer. There are also many efforts for designing efficient Transformers (Zhou et al., 2020; Wu et al., 2020; Tay et al., 2020). For example, researchers have applied well studied compression methods to Transformers, such as pruning (Guo et al.), quantization (Wang and Zhang, 2020; Guo et al., 2022), distillation (Sanh et al., 2019), and weight sharing. Other efforts focus on dedicated efficient attention mechanism considering its quadratic complexity of sequence length (Kitaev et al., 2020; Beltagy et al., 2020; Zahoor et al., 2020) or efficient feed-forward neural network (FFN) design regarding its dominant complexity in Transformer model (Dong et al., 2021). Transkimmer is orthogonal to these techniques on the input dimension reduction.

3 Input Skimming Search Space

In this section, we discuss the challenges of dynamic input skimming idea in details. Moreover, we compare techniques and design decisions from prior works described in Tbl. 1.

3.1 Optimization Method

The first challenge of input skimming is the optimization with discrete skimming decisions. In specific, the decision for pruning the hidden state tensors (i.e., reducing their sequence length) is

Models	Optimization	Input	Discard	Strategy
POWER-BERT (Goyal et al., 2020)	Soft-Masking	Attention	Discard	Searched
LAT (Kim and Cho, 2021)	Soft-Masking	Attention	Forward	Searched
LTP (Kim et al., 2021)	Soft-Masking	Attention	Discard	Learned
TR-BERT (Ye et al., 2021)	RL	Embedding	Forward	Searched
Transkimmer	Reparameterize	Embedding	Forward	Learned

Table 1: Summary of prior token reduction works and their design choices including POWER-BERT, Length-Adaptive Transformer (LAT), Learned Token Pruning (LTP) and TR-BERT. The design details are discussed in Sec. 3.

a binary prediction. As such, the skim prediction model is non-differentiable and unable to be directly optimized by gradient back propagation. Prior works handle the discrete binary skimming decision by using a set of complicated training techniques, which we categorize in Tbl. 1.

Soft-Masking. Some works (Goyal et al., 2020; Kim and Cho, 2021; Kim et al., 2021) propose to use the soft-masking training trick which uses a continuous value for predicting the skimming prediction. During the training process, the predicted value is multiplied to the hidden states embedding vectors so that no actual pruning happens. In the inference phase, this continuous skimming prediction value is binarized by a threshold-based step function. The threshold value is pre-defined or determined through a hyper-parameter search process. Obviously, there exists a training-inference paradigm mismatch where the actual skimming **only** happens at the inference time. Such a mismatch leads to a significant accuracy degradation.

Reinforcement Learning. TR-BERT (Ye et al., 2021) proposes to use the reinforcement learning (RL) to solve the discrete skimming decision problem. It uses a separated policy network as the skimming predictor, and the backbone Transformer model is considered as the value network. At first, the backbone Transformer is fine-tuned separately. It then updates the skimming policy network by using the RL algorithm. This multi-step training paradigm is tedious. And training the backbone Transformer and skimming policy network separately is sub-optimal compared to the joint optimization paradigm. Moreover, the large search space of such RL objective is difficult to converge especially on small downstream datasets.

Reparameterization. In this work, we propose to use the reparameterization technique to address the discrete skimming decision challenge. Its core idea is to sample the backward propagation gradient during training, whose details we describe in Sec. 4. The advantage of our method is that it enables the joint optimization of skim predictor and backbone Transformer model and therefore achieves the optimal solution. For example, we will later demonstrate in Fig. 4 that the different tasks or datasets prefer different layer-wise skimming strategies, which are learned by our method. We will further explain the results in Sec. 5.4.

3.2 Design Choices

In our work, we also jointly consider other design choices regarding the skimming optimization, which includes the choice of input to the skimming module and how to deal with the skimmed input. We first explain the choices made by prior works, and then explain the choice of our method.

Strategy. For the skimming optimization methods described above, there can be different strategies regarding the implementation details. Generally, the skimming strategy can be categorized into search-based or learning-based approach, as described in Tbl. 1. However, when applied to various downstream NLP tasks and datasets, the dynamic skimming scheme prefers different layer-wise strategies as we mentioned above. This layer-wise skimming characteristics makes the search-based approach not scalable and generally applicable. In contrast, our method enables the joint training of skimming strategy and downstream task, which leads to better skimming decisions with reference to both efficiency and accuracy. LTP is the only by prior works adopting learning-based method, which, however, uses the soft-masking approach and suffers from the training-inference mismatch.

Input for Skimming. POWER-BERT, LAT and LTP treat the attention weight value as importance score and utilize it as the criterion for making the skimming decision. Compared to this value-based method (Guan et al., 2020), TR-BERT uses hidden state embeddings as input feature. In our work, we use the hidden state embeddings because they enclose contextual information of the corresponding input token. Our work shows that the joint training of skimming module and backbone Transformer

model leads to that the embeddings also learn to carry features for skimming prediction.

Skimming Tokens. For the tokens pruned dynamically by the skimming decision during processing, it is natural to remove them from all the successive layers. However, LAT and TR-BERT propose to forward such tokens to the final output of the Transformer encoder, which keeps the dimension of the Transformer output unchanged. Our work adopts the forward-based design because it is more friendly for the Transformer decoder module on downstream tasks.

4 Transkimmer Methodology

4.1 Transformer with Skim Predictor

To predict which tokens to be pruned, we append an extra prediction module before each layer as shown in Fig. 2. This prediction module outputs a skimming mask M , which is used to gather the hidden state embedding H at the sequence length dimension. The pruned embedding is then feed to the Transformer layer as its input.

$$\begin{aligned} H^{i+1} &= \text{Transkimmer}^i(H^i) \\ &= \text{Transformer}^i(\text{Gather}(H^i, M^i)) \end{aligned} \quad (1)$$

In the skim mask, we use output 1 to denote remaining tokens and 0 to denote pruned tokens. The gathering operation is to select the input tensor with a provided mask. By optimizing this stand-alone skim module, syntactically redundant and semantically irrelevant tokens are skimmed and pruned. The proposed skim predictor module is a multi-layer perceptron (MLP) network composed of 2 linear layers with a layer norm operation (Ba et al., 2016) and GeLU activation (Hendrycks and Gimpel, 2016). The activation function is an arbitrary function with discrete output as skim decision.

$$\begin{aligned} M^i &= \text{SkimPredictor}(H^i) \\ &= \text{Activation}(\text{MLP}(H^i)) \end{aligned} \quad (2)$$

where $\text{MLP} = \text{Linear}(\text{GeLU}(\text{LN}(\text{Linear})))$

This skim predictor introduces extra model parameters and computation overhead. However, both of them are very small compared to the vanilla Transformer model, which are about 7.9% and 6.5% respectively. We demonstrate later that the computation overhead of skim module is much smaller than the benefits brought by the reduction of input tensor through skimming.

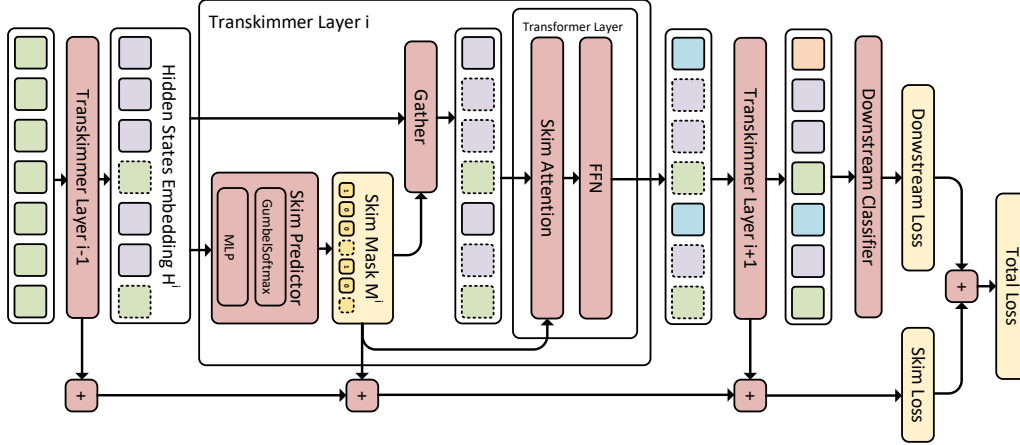


Figure 2: Architecture and end-to-end optimization objective of Transkimmer. The dashed token embeddings are directly forwarded to the final output of Transformer layers without further processing.

For the tokens pruned by the skim module at each layer, we forward these pruned hidden state embeddings to the last Transformer layer. As such, the final output of the whole Transformer model is composed of token embeddings skimmed at all layers and the ones processed by all layers without being skimmed.

$$H^L = \sum_{i=0}^{L-1} H^i \cdot M^i \quad (3)$$

And this output is used for classification layers on various downstream tasks. This makes the skimming operation also compatible for token classification tasks such as extractive question answering (QA) and named entity recognition (NER). This also restores the once abandoned information for downstream tasks.

4.2 End-to-End Optimization

In the above discussion, we have described that Transkimmer can be easily augmented to a backbone model without modification to its current structure. Furthermore, Transkimmer is also capable to utilize the pre-trained model parameters and finetune the Transkimmer activated Transformer-based models on downstream tasks. With an extra skim loss appended to the optimization object, this fine-tuning process is also performed end-to-end without changing its origin paradigm.

Skim Attention. In the training procedure, Transkimmer does not prune the hidden state tensors as it does in the inference time. Because the gathering and pruning operation of a portion of tokens prevents the back-propagation of their gradients.

The absence of error signal from negative samples interferes the convergence of the Transkimmer model. Therefore, we propose skim-attention to mask the reduced tokens in training instead of actually pruning them. The attention weights to the skimmed tokens are set to 0 and thus unreachable by the other tokens.

$$\text{SkimAttn}(H^i) = \text{Attn}(H^i) \cdot M^i \quad (4)$$

By doing so, the remaining tokens will have the identical computational value as actually pruning. And the gradient signal is passed to the skim predictor module from the skim attention multiplication.

Gumbel Softmax. Following the discussion in Sec. 3.1, the output decision mask of skim predictor is discrete and non-differentiable. To conquer this inability of back propagation, we use the reparameterization method (Jang et al., 2017) to sample the discrete skim prediction from the output probability distribution π^i of the MLP. The gradient of the non-differentiable activation function is estimated from the Gumbel-Softmax distribution during back propagation.

$$\begin{aligned} M_j^i &= \text{Activation}(\pi_j^i), \text{ for } j = 0, 1 \\ &= \text{GumbelSoftmax}(\pi_j^i) \\ &= \frac{\exp((\log(\pi_j^i) + g_j^i)/\tau)}{\sum_{k=0}^1 \exp((\log(\pi_k^i) + g_k^i)/\tau)} \end{aligned} \quad (5)$$

g_j^i are independent and identically sampled from $\text{Gumbel}(0, 1)$ distribution. τ is the temperature hyper-parameter controlling the one-hot prediction distribution. We take $\tau = 0.1$ for all experiments.

Dataset	CoLA	RTE	QQP	MRPC	SST-2	MNLI	WNLI	QNLI	STS-B	SQuAD	IMDB	YELP	20News
Task	Acceptability	NLI	Similarity	Paraphrase	Sentiment	NLI	NLI	QA	Similarity	QA	Sentiment	Sentiment	Sentiment
Average Sample Length	11	64	30	53	25	39	37	51	31	152	264	179	551
Input Sequence Length	64	256	128	128	64	128	128	128	64	384	512	512	512
Harmony Coefficient	0.3	0.8	0.2	0.5	0.3	0.2	0.5	0.1	0.3	0.8	0.5	0.5	0.5

Table 2: Summary of evaluation datasets. The input sequence length matches the setting of prior works POWER-BERT and LTP. It is determined by covering 99 percentile of input samples without truncation.

To achieve better token sparsification ratio, we further add a skim loss term to the overall optimization objective as follows

$$Loss_{skim} = \frac{1}{L} \sum_{L-1}^1 \frac{sum(M^i)}{len(M^i)}. \quad (6)$$

The skim loss is essentially the ratio of tokens remained in each layer thus representing the computation complexity speedup. By decreasing this objective, more tokens are forced to be pruned during processing. To collaborate with the original downstream task loss, we use a harmony coefficient λ to balance the two loss terms. As such, the total loss used for training is formulated as

$$Loss_{total} = Loss_{downstream} + \lambda Loss_{skim}. \quad (7)$$

With the use of the previous settings, the Transkimmer model is trained end-to-end without any change to its original training paradigm.

Unbalanced Initialization. Another obstacle is that skimming tokens during the training process makes it much unstable and decreases its accuracy performance. With the pre-trained language modeling parameters, the skim predictor module is random initialized and predicts random decisions. This induces significant processing mismatch in the backbone Transformer model, where all tokens are accessible. Consequently, the randomly initialized skim predictor makes the training unstable and diverged. We propose an unbalance initialization technique to solve this issue. The idea is to force positive prediction at first and learn to skim gradually. Generally, parameters are initialized by zero mean distribution as

$$\omega \sim N(0, \sigma). \quad (8)$$

We propose to initialize the bias vector of the last linear layer in the skim predictor MLP with unbalanced bias as

$$\beta_i \sim N((-1)^{i+1} \mu_0, \sigma), \quad (9)$$

where i stands for the bias vector for prediction 1 or 0. Consequently, the skim predictor tends to reserve tokens rather than skimming them when innocent. The mean value μ_0 of the unbalanced distribution set to 5 for all the experiments.

5 Evaluation

5.1 Setup

Datasets. We evaluate the proposed Transkimmer method on various datasets. We use the GLUE(Wang et al., 2019) benchmark including 9 classification/regression datasets, extractive question answering dataset SQuAD-v2.0, and sequence classification datasets 20News (Lang, 1995), YELP (Zhang et al., 2015) and IMDB (Maas et al., 2011). These datasets are all publicly accessible and the summary is shown in Tbl. 2. The diversity of tasks and text contexts demonstrates the general applicability of the proposed method.

Models. We follow the setting of the BERT model to use the structure of the Transformer encoder and a linear classification layer for all the datasets. We evaluate the base setting with 12 heads and 12 layers in prior work (Devlin et al., 2019). We implement Transkimmer upon BERT and RoBERTa pre-trained language model on downstream tasks.

Baselines. We compare our work to prior token reduction works including POWER-BERT (Goyal et al., 2020), Length-Adaptive Transformer (LA-Transformer) (Kim and Cho, 2021), Learned Token Pruning (LTP) (Kim et al., 2021), DeFormer (Cao et al., 2020) and TR-BERT (Kim et al., 2021). We also compare our method with model compression methods of knowledge distillation and weight sharing. Knowledge distillation uses a teacher model to transfer the knowledge to a smaller student model. Here we adopt DistilBERT (Sanh et al., 2019) setting to distill a 6-layer model from the BERT_{base} model. By sharing weight parameters among layers, the amount of weight parameters reduces. Note that weight sharing does not impact the computa-

Method	Padding	COLA		RTE		QQP		MRPC		SST-2		MNLI		WNLI		QNLI		STS-B		
		Matthews	FLOPs	Acc.	FLOPs	Acc.	FLOPs	F1	FLOPs	Acc.	FLOPs	Acc.	FLOPs	Acc.	FLOPs	Acc.	FLOPs	Pearson	FLOPs	
BERT _{base}	Baseline	-	57.8	1.00×	65.7	1.00×	91.3	1.00×	88.9	1.0×	93.0	1.00×	84.9	1.00×	56.3	1.00×	91.4	1.00×	88.6	1.00×
	DeeBERT	-	-	-	66.7	1.50×	-	-	85.2	1.79×	91.5	1.89×	80.0	1.59×	-	-	87.9	1.79×	-	-
	POWER-BERT	Sequence	52.3	4.50×	67.4	3.40×	90.2	4.50×	88.1	2.70×	92.1	2.40×	83.8	2.60×	-	-	90.1	2.00×	85.1	2.00×
	LAT	Sequence	-	-	-	-	-	-	-	-	92.8	2.90×	84.4	2.80×	-	-	-	-	-	-
	Transkimmer	No	58.9	1.75×	68.9	2.85×	90.8	2.79×	88.5	3.13×	92.3	1.58×	83.2	2.02×	56.3	5.56×	90.5	2.33×	87.4	3.45×
	Transkimmer	Sequence	58.9	18.9×	68.9	4.67×	90.8	11.72×	88.5	7.45×	92.3	10.89×	83.2	6.65×	56.3	18.10×	90.5	6.01×	87.4	18.20×
	DistilBERT	-	55.7	1.98×	58.8	1.98×	90.3	1.98×	88.3	1.98×	90.6	1.98×	87.5	1.98×	53.5	1.98×	89.3	1.98×	87.0	1.98×
	+Transkimmer	No	55.1	3.52×	59.2	4.12×	90.1	4.95×	87.8	9.92×	89.5	5.01×	86.7	4.40×	56.3	10.41×	87.5	4.04×	86.5	3.47×
	ALBERT	-	58.3	0.99×	70.7	0.99×	90.2	0.99×	90.4	0.99×	90.9	0.99×	81.8	0.99×	56.3	0.99×	89.2	0.99×	90.4	0.99×
	+Transkimmer	No	53.4	1.52×	71.5	1.57×	90.2	3.09×	90.6	1.94×	90.1	3.25×	81.5	1.67×	57.7	6.19×	90.1	2.30×	89.8	1.46×
RoBERTa _{base}	Baseline	-	61.8	1.00×	78.0	1.00×	90.4	1.00×	92.1	1.00×	94.3	1.00×	87.5	1.00×	56.6	1.00×	92.9	1.00×	90.9	1.00×
	LTP	Batch	-	-	78.0	1.81×	89.7	2.10×	91.6	2.10×	93.5	2.09×	86.5	1.88×	-	-	92.0	1.87×	90.0	1.95×
	Transkimmer	No	61.3	1.52×	76.2	1.79×	91.0	4.92×	91.9	2.67×	93.5	2.08×	86.7	2.19×	56.3	8.41×	91.7	2.85×	90.5	2.70×

Table 3: Performance and FLOPs (speedup) on GLUE benchmark with BERT_{base} and RoBERTa_{base} as backbone model. Transkimmer is adopted on DistilBERT and ALBERT to shows its applicability to general model compression methods.

Model	Padding	SQuADv2.0		20News		Yelp		IMDB	
		F1	FLOPs	Acc.	FLOPs	Acc.	FLOPs	Acc.	FLOPs
BERT _{base}	-	77.1	1.00×	86.7	1.00×	69.9	1.00×	94.0	1.00×
TR-BERT	No	75.7	2.08×	87.4	4.22×	70.0	2.19×	93.6	2.26×
POWER-BERT	Sequence	-	-	86.5	2.91×	67.4	2.75×	92.1	3.05×
LAT	Batch	-	-	-	-	-	-	92.5	2.70×
DeFormer	Sequence	71.4	2.19×	-	-	-	-	-	-
Transkimmer	No	75.7	2.10×	86.1	5.27×	70.1	2.51×	93.7	2.70×

Table 4: Performance and FLOPs evaluation on several downstream tasks and datasets with BERT_{base} as backbone model. The speedup results are emphasized considering the padding setting.

tion FLOPs (floating-point operations). We evaluate Transkimmer on ALBERT (Lan et al., 2020) that shares weight parameters among all layers. To express that token reduction method is compatible with these model compression methods, we further implement Transkimmer method with this works to demonstrate their cooperation effect. Besides, DeeBERT (Xin et al., 2020) is a Transformer early exit baseline which can be regarded as coarse-grained input skimming.

Padding. While processing batched input samples, Transformer models perform a padding operation on the input sequences to align the input length. Sequences are appended with a special padding token [PAD] to a predefined sequence length for the convenience of successive computing. This is a trivial setting for general evaluation but could lead to possible pseudo speedup for token reductions works. Because the padded tokens can be pruned without prediction. For the prior works, there are three evaluation settings with reference to padding, padding to a fixed sequence length, padding to mini-batch maximum length and no padding (denoted as **Sequence**, **Batch** and **No** in Fig. 3 & 4). We indicate the padding methods of prior works and evaluate Transkimmer with differ-

ent padding settings for a fair comparison. The speedup of padding to mini-batch maximum length setting is related to batch size and processing order of input samples. So it is difficult to make a direct comparison under this setting. However, it can be estimated with padding to fixed sequence length as upper bound and no padding as lower bound. The sequence length on different datasets is determined following prior works’ settings (Goyal et al., 2020; Kim et al., 2021). We measure the inference FLOPs as a general measurement of the model computational complexity on all platforms. We use the TorchProfile(?) tool to calculate the FLOPs for each model.

Training Setting. We implement the proposed method based on open-sourced library from Wolf et al. (2020)¹. For each baseline model, we use the released pre-trained checkpoints². We follow the training setting used by Devlin et al. (2019) and Liu et al. (2019) to perform the fine-tuning on the above datasets. We perform all the experiments reported with random seed 42. We use four V100 GPUs for training experiments.

The harmony coefficient λ is determined by hyper-parameter grid search on development set with 20% data random picked from training set set. The search space is from 0.1 to 1 with a step of 0.1.

5.2 Overall Results

We show the overall results on several datasets and demonstrate our observations. Tbl. 3 demonstrates the accuracy and speedup evaluated on GLUE benchmark. And Tbl. 4 further demonstrates the results on other datasets with longer input.

¹The source code is available at <https://github.com/ChandlerGuan/Transkimmer>.

²We use pre-trained checkpoints from Wolf et al. (2020).

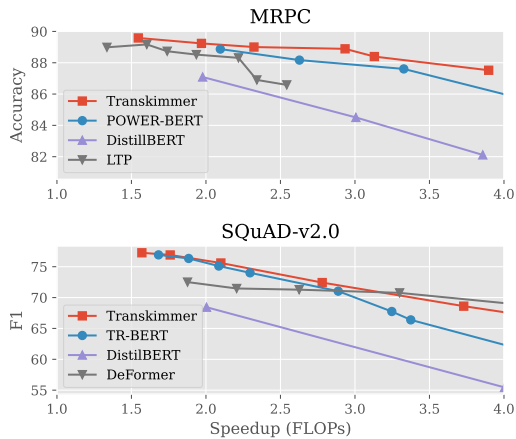


Figure 3: Trade-off results between accuracy and speedup of MRPC and SQuAD-v2.0 datasets by tuning the harmony coefficient. Note that different padding settings are used for each baseline while Transkimmer doesn't count any padding.

Comparison to vanilla model baseline. Generally, Transkimmer achieves considerably speedup to the vanilla models with a minor accuracy degradation, which is less than 1% for nearly all cases. The average speedup is $2.81\times$ on GLUE benchmark and over $2\times$ on the other datasets. This demonstrates the inference efficiency improvement of the Transkimmer input reduction method. We also evaluate Transkimmer with RoBERTa model as backbone and reach $3.24\times$ average speedup on GLUE benchmark. This result further expresses the general applicability of Transkimmer with different Transformer-based pre-trained language models. Among all the datasets we evaluated, Transkimmer tends to have better acceleration ratio on the easier ones. For example, sequence classification tasks like QQP and STS-B are better accelerated than QA or NLI datasets. We suggest that the Transformer backbone is able to process the information at shallower layers and skim the redundant part earlier. This is also demonstrated in the following post-hoc analysis Sec. 5.4.

Comparison to input reduction prior works. As shown in Tbl. 3, Transkimmer outperforms all the input reduction methods by a margin on GLUE benchmark. To make a fair comparison, we evaluate Transkimmer with two padding settings, padding to fixed sequence length or no padding. For most cases, Transkimmer has better accuracy performance and higher speedup ratio at the same time. When taking the special padding token into account, Transkimmer is able to accelerate

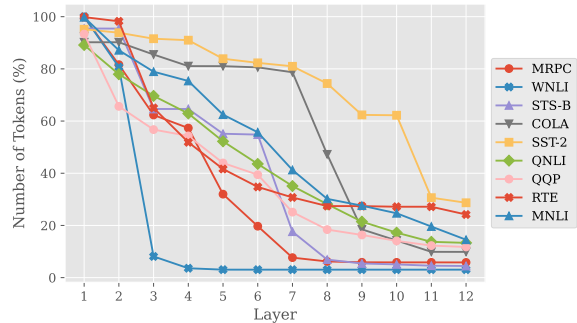


Figure 4: Layer-wise skim strategies analysis of datasets from GLUE benchmark. The normalized area under curve is viewed as an approximate speedup ratio with reference to sequence length.

BERT_{base} model for $10.97\times$ on GLUE benchmark. Transkimmer also outperforms the other methods on tasks shown in Tbl. 4. TR-BERT has the closest performance compared with Transkimmer but with a much complicated RL paradigm and larger search space.

Comparison to model compression methods. The comparison to two model compression methods is shown in Tbl. 3. Transkimmer outperforms the knowledge distillation and weight sharing baseline by a margin. Besides, the dynamic skimming idea itself is orthogonal to this existing model compression methods. To elaborate, we further adopt the proposed Transkimmer method on DistilBERT and ALBERT models. With the proposed end-to-end training objective, Transkimmer is easily augmented to these methods. There is also no need to change the original training process. The result shows that the Transkimmer method further accelerates the inference efficiency of compressed models with nearly no extra accuracy degradation.

5.3 Accuracy and Performance Trade-Off

Fig. 3 demonstrates the accuracy and performance trade-off analysis by tuning the harmony coefficient. We show the results on MRPC and SQuAD-v2.0 datasets to give comparisons with different baselines. It is shown that Transkimmer achieves a better accuracy to speedup Pareto curve compared to prior works. Transkimmer is able to provide better acceleration gain with less accuracy degradation. Especially, Transkimmer has a $1.5\times$ speedup without accuracy loss. The result validates our design decisions analyzed in the input reduction search space choices.

Dataset	Example
SST-2	[CLS] Even horror fans will most likely not find what they're seeking with trouble every day; the movie lacks both thrills and humor. [SEP]
SQuAD	Question: [CLS] In what country is Normandy located? [SEP] Context: The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries. [SEP] Answer: France

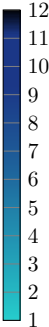


Table 5: Post-hoc case study of SST-2 sentimental analysis and SQuAD QA tasks from Transkimmer model with BERT_{base} setting. The color indicated by the colorbar represents the Transformer layer index where the token is pruned. Specifically, the black tokens are fully processed without being skimmed.

5.4 Post-hoc Analysis

Skim Strategy. Fig. 4 is the result of the number of tokens remained for the processing of each Transformer layer. The normalized area under each curve is a rough approximation of the speedup ratio with reference to the tokens number. By end-to-end optimization, Transkimmer learns significant distinguished strategies on different tasks. On WNLI dataset, over 90% of tokens are pruned within the first 3 layers and guarantees a high acceleration gain. The steep cliff at layer 7 on COLA demonstrates a large portion of skipping at this particular position. We suggest that this is because the processing of contextual information is sufficient for the skipping decision at this specific layer.

Post-Hoc Case Study. Moreover, several post-hoc case studies are demonstrated with Tbl. 5. In the SST-2 sentimental analysis example, the definite articles and apostrophes are discarded at the beginning. And all words are encoded in contextual hidden states embeddings and gradually discarded except for a few significant key words. Only the special token [CLS] is fully processed in this example for final sentimental classification. However, on the token classification task example from SQuAD dataset, all tokens are given to the downstream classifier to predict the answer position. The answer tokens are processed by all Transformer layers. Similarly, the question part is also kept with tokens containing enough information. Another detail worth mentioning is that we use subword tokenization for the SQuAD dataset. As such, subword tokens of the same word might be discarded at different layers. For instance, the word *Francia* is tokenized into *fran-* and *-cia* two subword tokens, which are pruned at layer 4 and 6 respectively.

6 Conclusion

Input skimming or dynamic input reduction is an emerging Transformer model acceleration method studied by many works recently. This idea utilizes the semantic structure of language and the syntactic information of the input context for inference acceleration. Compared to static model weight compression methods, input skimming explores the redundancy in the input and hidden state tensors. As such, it is orthogonal and compatible with those model compression algorithms with its dynamic feature.

In this work, we propose an accurate and efficient Transformer inference acceleration method by teaching it how to skim input contents. The proposed Transkimmer method is trained with an easy and end-to-end paradigm. Furthermore, Transkimmer is also generally applicable to various Transformer-based model structures. It is even compatible with the static model compression methods like knowledge distillation and weight sharing. We believe that the above features guarantee the Transkimmer method a wide range of applicable production scenarios.

Acknowledgement

This work was supported by the National Key R&D Program of China under Grant 2021ZD0110104, the National Natural Science Foundation of China (NSFC) grant (U21B2017, 62106143, 62072297, and 61832006), and Shanghai Pujiang Program. We would like to thank the reviewers of ACL rolling review for their supportive comments and suggestions. Jingwen Leng and Minyi Guo are the corresponding authors of this paper.

References

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. [Layer normalization](#). *arXiv preprint arXiv:1607.06450*.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *arXiv preprint arXiv:2004.05150*.
- Víctor Campos, Brendan Jou, Xavier Giró-i-Nieto, Jordi Torres, and Shih-Fu Chang. 2018. [Skip RNN: learning to skip state updates in recurrent neural networks](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjana Balasubramanian. 2020. [DeFormer: Decomposing pre-trained transformers for faster question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4497, Online. Association for Computational Linguistics.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. [Universal transformers](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- 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.
- Chenhe Dong, Guangrun Wang, Hang Xu, Jiefeng Peng, Xiaozhe Ren, and Xiaodan Liang. 2021. [Efficientbert: Progressively searching multilayer perceptron via warm-up knowledge distillation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1424–1437.
- Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Rajee, Venkatesan T. Chakaravarthy, Yogish Sabharwal, and Ashish Verma. 2020. [Power-bert: Accelerating BERT inference via progressive word-vector elimination](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research*, pages 3690–3699. PMLR.
- Xiaotao Gu, Liyuan Liu, Hongkun Yu, Jing Li, Chen Chen, and Jiawei Han. 2021. [On the transformer growth for progressive BERT training](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5174–5180, Online. Association for Computational Linguistics.
- Yue Guan, Jingwen Leng, Chao Li, Quan Chen, and Minyi Guo. 2020. [How far does BERT look at: Distance-based clustering and analysis of BERT’s attention](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3853–3860, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Yue Guan, Zhengyi Li, Jingwen Leng, Zhouhan Lin, Minyi Guo, and Yuhao Zhu. 2021. [Block-skim: Efficient question answering for transformer](#). *arXiv preprint arXiv:2112.08560*.
- Cong Guo, Bo Hsueh, Jingwen Leng, Yuxian Qiu, Yue Guan, Zehuan Wang, Xiaoying Jia, Xipeng Li, Minyi Guo, and Yuhao Zhu. [Accelerating sparse dnn models without hardware-support via tile-wise sparsity](#). In *2020 SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 204–218. IEEE Computer Society.
- Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. 2022. [SQuant: On-the-fly data-free quantization via diagonal hessian approximation](#). In *International Conference on Learning Representations*.
- Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. [Neural speed reading with structural-jump-lstm](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Xuanli He, Iman Keivanloo, Yi Xu, Xiang He, Belinda Zeng, Santosh Rajagopalan, and Trishul Chilimbi. 2021. [Magic pyramid: Accelerating inference with early exiting and token pruning](#). *arXiv preprint arXiv:2111.00230*.
- Dan Hendrycks and Kevin Gimpel. 2016. [Gaussian error linear units \(gelus\)](#). *arXiv preprint arXiv:1606.08415*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with gumbel-softmax](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Gyuwan Kim and Kyunghyun Cho. 2021. [Length-adaptive transformer: Train once with length drop, use anytime with search](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6501–6511, Online. Association for Computational Linguistics.

- Sehoon Kim, Sheng Shen, David Thorsley, Amir Ghلامي, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. 2021. [Learned token pruning for transformers](#). *arXiv preprint arXiv:2107.00910*.
- Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. [Reformer: The efficient transformer](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [ALBERT: A lite BERT for self-supervised learning of language representations](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier.
- Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. [A structured self-attentive sentence embedding](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. 2020. [FastBERT: a self-distilling BERT with adaptive inference time](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6035–6044, Online. 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*.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Priyadarshini Panda, Abhronil Sengupta, and Kaushik Roy. 2016. Conditional deep learning for energy-efficient and enhanced pattern recognition. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 475–480. IEEE.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *arXiv preprint arXiv:1910.01108*.
- Min Joon Seo, Sewon Min, Ali Farhadi, and Hannaneh Hajishirzi. 2018. [Neural speed reading via skim-rnn](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *arXiv e-prints*, pages arXiv–2009.
- Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469. IEEE.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Chunpei Wang and Xiaowang Zhang. 2020. Q-bert: A bert-based framework for computing sparql similarity in natural language. In *Companion Proceedings of the Web Conference 2020*, pages 65–66.
- Hanrui Wang, Zhekai Zhang, and Song Han. 2021. Spatten: Efficient sparse attention architecture with cascade token and head pruning. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 97–110. IEEE.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020. [Lite transformer with long-short range attention](#). In *8th International Conference on*

Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

- Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. [DeeBERT: Dynamic early exiting for accelerating BERT inference](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online. Association for Computational Linguistics.
- Deming Ye, Yankai Lin, Yufei Huang, and Maosong Sun. 2021. [TR-BERT: Dynamic token reduction for accelerating BERT inference](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5798–5809, Online. Association for Computational Linguistics.
- Adams Wei Yu, Hongrae Lee, and Quoc Le. 2017. [Learning to skim text](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1880–1890, Vancouver, Canada. Association for Computational Linguistics.
- Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. [Big bird: Transformers for longer sequences](#). *arXiv preprint arXiv:2007.14062*.
- Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.
- Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. 2020. Bert loses patience: Fast and robust inference with early exit. *Advances in Neural Information Processing Systems*, 33.