# ERNIE-DOC: A Retrospective Long-Document Modeling Transformer

**Siyu Ding***, **Junyuan Shang***, **Shuohuan Wang**, **Yu Sun**, **Hao Tian**, **Hua Wu** and **Haifeng Wang**
Baidu Inc., China
{dingsiyu, shangjunyuan, wangshuohuan, sunyu02, tianhao, wu_hua, wanghaifeng}@baidu.com

## Abstract

Transformers are not suited for processing long documents, due to their quadratically increasing memory and time consumption. Simply truncating a long document or applying the sparse attention mechanism will incur the context fragmentation problem or lead to an inferior modeling capability against comparable model sizes. In this paper, we propose ERNIE-DOC, a document-level language pretraining model based on Recurrence Transformers (Dai et al., 2019). Two well-designed techniques, namely the retrospective feed mechanism and the enhanced recurrence mechanism, enable ERNIE-DOC [1], which has a much longer effective context length, to capture the contextual information of a complete document. We pretrain ERNIE-DOC to explicitly learn the relationships among segments with an additional document-aware segment-reordering objective. Various experiments were conducted on both English and Chinese document-level tasks. ERNIE-DOC improved the state-of-the-art language modeling result of perplexity to 16.8 on WikiText-103. Moreover, it outperformed competitive pretraining models by a large margin on most language understanding tasks, such as text classification and question answering.

## 1 Introduction

Transformers (Vaswani et al., 2017) have achieved remarkable improvements in a wide range of natural language tasks, including language modeling (Dai et al., 2019), text classification (Yang et al., 2019), and question answering (Devlin et al., 2018; Radford et al., 2019). This success is largely due to the self-attention mechanism, which enables the network to capture contextual information from the entire input sequence. Nevertheless, the memory usage and computation complexity caused by the self-attention mechanism grows quadratically with the sequence length, incurring excessive cost when processing a long document on existing hardware.

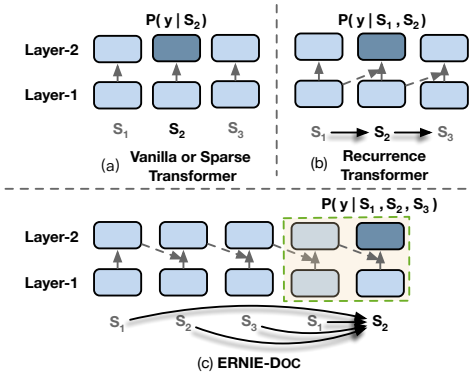


Figure 1: Available contextual information utilized by Transformer variants, where a long document $\mathcal{D}$ is partitioned into three segments $\mathbf{S}_i (i \in [1, 2, 3])$. When training on $\mathbf{S}_2$, (a) and (b) optimize the pretraining objective depending only on the contextual information from the current segment or segments in the forward pass, whereas ERNIE-DOC utilizes the contextual information of the entire document for each segment.

Currently, the most prominent pretrained models, such as BERT (Devlin et al., 2018), are used on fixed-length input segments of a maximum of 512 tokens owing to the aforementioned limitation. Thus, a long document input must be partitioned into smaller segments of manageable sizes. However, this leads to the loss of important cross-segment information, that is, the *context fragmentation* problem (Dai et al., 2019), as shown in Fig. 1(a). To mitigate the problem of insufficient interactions among the partitioned segments of long documents, *Recurrence Transformers* (Dai et al., 2019; Rae et al., 2019) permit the use of contextual information from previous segments in computing the hidden states for a new segment by maintaining a memory component from the previous activation;

*indicates equal contribution.

[1] Source code and pre-trained checkpoints can be found at https://github.com/PaddlePaddle/ERNIE/tree/repro/ernie-doc.



*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2914–2927
August 1–6, 2021. 

this enables the modeling of long documents. In addition, *Sparse Attention Transformers* (Child et al., 2019; Tay et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020) focus on reducing the complexity of self-attention operations to explicitly improve the modeling length, but only up to a restricted context length (4,096) due to resource limitations.

We argue that existing strategies are not sufficiently effective or reliable, because ***the contextual information of a complete document*** *is still not available for each segment during the training phase*. As depicted in Fig. 1, when training on segment $S_2$, the model is ideally optimized by maximizing $P(y \mid (S_1, S_2, S_3))$ conditioned on the contextual information of the entire document $\mathcal{D} = \{S_1, S_2, S_3\}$, in contrast to the following sub-optimal solutions: $P(y \mid S_2)$ for Vanilla/Sparse Transformers[2] and $P(y \mid (S_1, S_2))$ for Recurrence Transformers.

To address this limitation, we propose ERNIE-DOC (A Retrospective Long-Document Modeling Transformer) based on the Recurrence Transformer paradigm. Inspired by the human reading behavior of skimming a document first and then looking back upon it attentively, we design a **retrospective feed mechanism** in which segments from a document are fed twice as input. As a result, each segment in the retrospective phase could explicitly fuse the semantic information of the entire document learned in the skimming phase, which prevents context fragmentation.

However, simply incorporating the retrospective feed mechanism into Recurrence Transformers is infeasible because the maximum effective context length is limited by the number of layers (Dai et al., 2019), as shown in Fig. 1 (b). Thus, we present an **enhanced recurrence mechanism**, a drop-in replacement for a Recurrence Transformer, by changing the shifting-one-layer-downwards recurrence to the same-layer recurrence. In this manner, the maximum effective context length can be expanded, and past higher-level representations can be exploited to enrich future lower-level representations.

Moreover, we introduce a **segment-reordering objective** to pretrain a document-level model. Specifically, it is a document-aware task of predicting the correct order of the permuted set of segments of a document, to model the relationship among segments directly. This allows ERNIE-DOC to build full document representations for prediction. This is analogous to the sentence-reordering task in ERNIE 2.0 (Sun et al., 2020b) but at a segment level of granularity, spanning (commonly) multiple training steps.

We first evaluate ERNIE-DOC on autoregressive word-level language modeling using the enhanced recurrence mechanism, which, in theory, allows the model to process a document with infinite words. ERNIE-DOC achieves state-of-the-art (SOTA) results on the WiKiText-103 benchmark dataset, demonstrating its effectiveness in long-document modeling. Then, to evaluate the potential of ERNIE-DOC on document-level natural language understanding (NLU) tasks, we pretrained the English ERNIE-DOC on the text corpora utilized in BigBird (Zaheer et al., 2020) from the RoBERTa-released checkpoint, and the Chinese ERNIE-DOC on the text corpora utilized in ERNIE 2.0 (Sun et al., 2020b) from scratch. After pretraining, we fine-tuned ERNIE-DOC on a wide range of English and Chinese downstream tasks, including text classification, question answering and keypharse extraction. Empirically, ERNIE-DOC consistently outperformed RoBERTa on various benchmarks and showed significant improvements over other high-performance long-text pretraining models for most tasks.

## 2 Related Work

**Sparse Attention Transformers** have been extensively explored (Child et al., 2019; Tay et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020). The key idea is to sparsify the self-attention operation, which scales quadratically with the sequence length. For instance, the Sparse Transformer (Child et al., 2019) uses a dilated sliding window that reduces the complexity to $\mathcal{O}(L\sqrt{L})$, where $L$ is the sequence length. Reformer (Kitaev et al., 2020) further reduces the complexity to $\mathcal{O}(L \log L)$ using locality-sensitive hashing attention to compute the nearest neighbors. BP-Transformers (Ye et al., 2019) employs a binary partition for the input sequence. Recently, Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) have been proposed, and both achieved state-of-the-art performance on a variety of long-document tasks. They reduce the complexity of self-attention to $\mathcal{O}(L)$ by combining random attention, window attention, and global attention. However, it has been proven in Zaheer et al. (2020) that sparse attention mech-

[2]For Sparse Transformers, the length of segment $S_2$ could be up to 4,096 in Beltagy et al. (2020); Zaheer et al. (2020).

anisms cannot universally replace dense attention mechanisms; moreover, solving the simple problem of finding the furthest vector requires $\Omega(n)$-layers of a sparse attention mechanism but only $\mathcal{O}(1)$-layers of a dense attention mechanism. In addition, the aforementioned methods require customized CUDA kernels or TVM programming to implement sparse attention, which are not maintainable and are difficult to use. In this study, we adopt a different approach to adapting Recurrence Transformers for a pretraining-then-finetuning setting, to model a long document.

**Recurrence Transformers** (Dai et al., 2019; Rae et al., 2019) have been successfully applied in generative language modeling. They employ the Transformer decoder as a parametric model for each conditional distribution in $p(\boldsymbol{x}) = \prod_{t=1}^{L} p(x_t|\boldsymbol{x}_{<t})$, where $\boldsymbol{x}$ denotes a text sequence. To capture long dependencies, they process the text in segments from left to right based on the segment recurrence mechanism (Dai et al., 2019). This mechanism maintains a memory bank of past activations at each layer to preserve a history of context. Compressive Transformer (Rae et al., 2019) adds a compressive memory bank to sufficiently store old activations instead of discarding them, which facilitates long-range sequence learning. However, these methods operate from left to right, which limits their capacity for discriminative language understanding tasks that require bidirectional information. XLNet (Yang et al., 2019) proposed a permutation language modeling objective to construct bidirectional information and achieve superior performance in multiple NLP tasks; however, its application to long-document modeling tasks remains largely unexplored. ERNIE-DOC builds on the ideas of the Recurrence Transformers to 1) tackle the limitation of Recurrence Transformers for utilizing bidirectional contextual information and 2) improve the behavior of the segment recurrence mechanism to capture longer dependencies.

**Hierarchical Transformers** (Zhang et al., 2019; Lin et al., 2020) have enabled significant progress on numerous document-level tasks, such as document summarization (Zhang et al., 2019) and document ranking (Lin et al., 2020). Similar to Vanilla Transformers, Hierarchical Transformers also split long documents into shorter segments with manageable lengths and then feed them independently to produce corresponding segment-level semantic representations. Unlike in Vanilla Transformers, however, separate Transformer layers are used in Hierarchical Transformers to process the concatenation of these representations. Hierarchical Transformers ignore the contextual information from the remaining segments when processing each segment of a long document, thus suffering from the *context fragmentation* problem.

## 3 Proposed Method

In this section, we first describe the background (Sec. 3.1) that ERNIE-DOC builds on. Then, we present the implementation of ERNIE-DOC, including the retrospective feed mechanism in Sec. 3.2, the enhanced recurrence mechanism in Sec. 3.3, and the segment-reordering objective in Sec. 3.4.

### 3.1 Background

Formally, a long document $\mathcal{D}$ is sliced into $T$ sequential segments, denoted as $\{S_1, S_2, ..., S_T\}$, where $S_\tau = \{x_{\tau,1}, x_{\tau,2}, ..., x_{\tau,L}\}$ is the $\tau$-th segment with $L$ tokens; $x$ denotes a single token. Vanilla, Sparse, and Recurrence Transformers employ different strategies to produce the hidden state $\mathbf{h}_\tau^n \in \mathbb{R}^{L\times d}$ for segment $S_\tau$ at the $n$-th layer:

$$\begin{aligned}
\widetilde{\mathbf{h}}_{\tau+1}^{n-1} &= \begin{cases} \mathbf{h}_{\tau+1}^{n-1}, & \text{Vanilla or Sparse Transformers} \\ [\mathbf{SG}(\mathbf{h}_\tau^{n-1}) \circ \mathbf{h}_{\tau+1}^{n-1}], & \text{Recurrence Transformers,} \end{cases} \\
\mathbf{q}_{\tau+1}^n, \mathbf{k}_{\tau+1}^n, \mathbf{v}_{\tau+1}^n &= \mathbf{h}_{\tau+1}^{n-1}\mathbf{W}_q^\top, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_k^\top, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_v^\top. \\
\mathbf{h}_{\tau+1}^n &= \text{Transformer-Block}\,(\mathbf{q}_{\tau+1}^n, \mathbf{k}_{\tau+1}^n, \mathbf{v}_{\tau+1}^n).
\end{aligned} \tag{1}$$

where $\mathbf{q} \in \mathbb{R}^{L\times d}$, $\mathbf{k}$, and $\mathbf{v} \in \mathbb{R}^{(L+m)\times d}$ are the query, key and value vectors, respectively with hidden dimension $d$ and memory length $m$ (Note that $m = 0$ for Vanilla or Sparse Transformers); $\widetilde{\mathbf{h}}_{\tau+1}^{n-1} \in \mathbb{R}^{(L+m)\times d}$ is the extended context; $\mathbf{W}_* \in \mathbb{R}^{d_*\times d}$ represents learnable linear projection parameters; the function $\mathbf{SG}(\cdot)$ denotes the stop-gradient operation; and the notation $[\circ]$ denotes the concatenation of two hidden states along the length dimension. In contrast to Vanilla or Sparse Transformers, where $\mathbf{h}_{\tau+1}^n$ is produced using only itself, Recurrence Transformers introduce a segment-level recurrence mechanism to promote interaction across segments. The hidden state computed for the previous segment $\mathbf{h}_\tau^{n-1}$ is cached as an auxiliary context to help process the current segment $\mathbf{h}_\tau^n$. However, from the concatenation part in Eq. 1, i.e., $[\mathbf{SG}(\mathbf{h}_\tau^{n-1}) \circ \mathbf{h}_{\tau+1}^{n-1}]$, there is apparently a constraint that the current hidden state can only fuse information from the previous segments. In

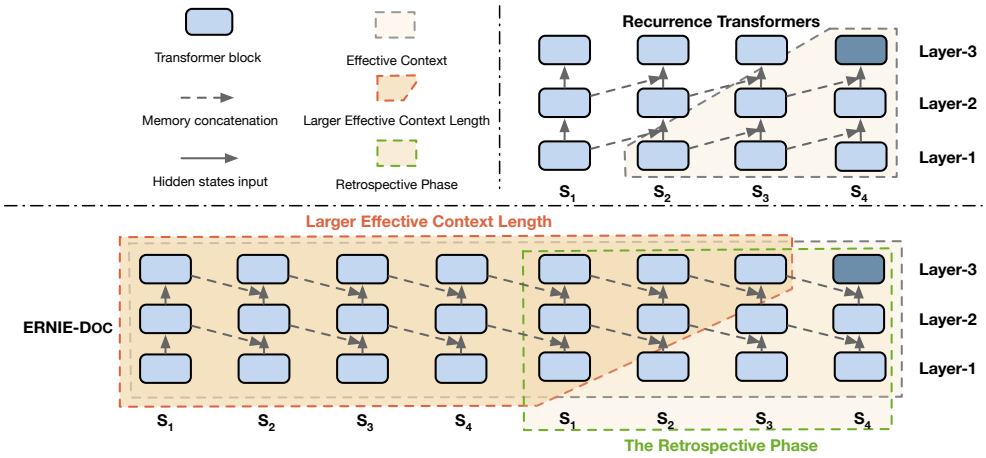


Figure 2: Illustrations of ERNIE-DOC and Recurrence Transformers, where models with three layers take as input a long document $\mathcal{D}$ which is sliced into four segments $\mathbf{S}_i, i \in [1, 2, 3, 4]$. **Recurrence Transformers (upper-right)**: When training on $\mathbf{S}_4$, it can only fuse the contextual information of the previous two consecutive segments $\mathbf{S}_2, \mathbf{S}_3$, since the largest effective context length grows linearly w.r.t the number of layers. **ERNIE-DOC (lower)**:The effective context length is much larger aided by the enhanced recurrence mechanism (Sec. 3.3). Thus, $\mathbf{S}_4$ can fuse the information of $\mathbf{S}_1$ discarded by Recurrence Transformers. Moreover, segments in the retrospective phase contains the contextual information of an entire document, powered by the retrospective feed mechanism (Sec. 3.2).

other words, the contextual information of an entire document is not available for each segment.

### 3.2 Retrospective Feed Mechanism

ERNIE-DOC employs a retrospective feed mechanism to address the unavailability of *the contextual information of a complete document for each segment*. The segments from a long document are twice fed as input. Mimicking the human reading behavior, we refer to the first and second input-taking phases as the skimming and retrospective phases, respectively. In the skimming phase, we employ a recurrence mechanism to cache the hidden states for each segment. In the retrospective phase, we reuse the cached hidden states from the skimming phase to enable bi-directional information flow. Naively, we can rewrite Eq. 1 to obtain the contextual information of an entire document in the skimming phase to be utilized in the retrospective phase as follows,

$$\begin{aligned} \widehat{\mathbf{H}} &= [\widehat{\mathbf{H}}^1_{1:T} \circ \widehat{\mathbf{H}}^2_{1:T} \cdots \circ \widehat{\mathbf{H}}^N_{1:T}], \text{ (skim. phase)} \\ \widetilde{\mathbf{h}}^{n-1}_{\tau+1} &= [\mathbf{SG}(\widehat{\mathbf{H}} \circ \mathbf{h}^{n-1}_{\tau}) \circ \mathbf{h}^{n-1}_{\tau+1}], \text{ (retro. phase)} \end{aligned} \tag{2}$$

where $\widehat{\mathbf{H}} \in \mathbb{R}^{(L*T*N)\times d}$ denotes the cached hidden states in the skimming phase with $T$ segments, $L$ length of each segment and total $N$ layers, and $\widehat{\mathbf{H}}^i_{1:T} = [\widehat{\mathbf{h}}^i_1 \circ \widehat{\mathbf{h}}^i_2 \cdots \circ \widehat{\mathbf{h}}^i_T]$ is the concatenation of $i$-th layer's hidden states of the skimming phase. Thus, the extended context $\widetilde{\mathbf{h}}^{n-1}_{\tau+1}$ is guaranteed to capture the bidirectional contextual information of the entire document. However, it will incur massive memory and computation cost for directly employing $\widehat{\mathbf{H}}$ in self-attention mechanism. Henceforth, the main issue is how $\widehat{\mathbf{H}}$ should be implemented in a memory- and computation-efficient manner.

By rethinking segment-level recurrence (Dai et al., 2019), we observe that the largest possible context dependency length increases linearly w.r.t the number of layers ($N$). For instance, at $i$-th layer, $\widehat{\mathbf{h}}^i_\tau$ have the longest dependency to $\widehat{\mathbf{h}}^1_{\tau-(i-1)}$. Thus, to minimize memory and computation consumption, hidden states from the $N$-th layer (top-layer) are included at a stride of $N$, which is sufficient to build the contextual information of an entire document. Formally, $\widehat{\mathbf{H}}$ can be reduced to $\widehat{\mathbf{H}}_r = [\widehat{\mathbf{h}}^N_N \circ \widehat{\mathbf{h}}^N_{2*N} \cdots \circ \widehat{\mathbf{h}}^N_{\lfloor T/N \rfloor * N}]$ (Note that when $T$ is not evenly divisible by $N$, the last hidden state $\widehat{\mathbf{h}}^N_T$ need to be included). However, for a long document input, the extra computational and memory cost of $\widehat{\mathbf{H}}_r \in \mathbb{R}^{\lceil T/N \rceil \times d}$ where $T \gg N$ is still excessive on existing hardware.

### 3.3 Enhanced Recurrence Mechanism

To effectively utilize the retrospective feed mechanism in practice, an ideal strategy is to ensure that the cached hidden state $\mathbf{h}^{n-1}_\tau$ already contains the contextual information of an entire document without explicitly taking $\widehat{\mathbf{H}}$ or $\widehat{\mathbf{H}}_r$ as input. Essentially, we should tackle the problem of limited effective context length in the segment-level recurrence mechanisms. Herein, we introduce the enhanced recurrence mechanism, a drop-in replacement for the segment-level recurrence mechanism,

by changing the shifting-one-layer-downwards recurrence to the same-layer recurrence as follows:

$$\widetilde{\mathbf{h}}_{\tau+1}^{n-1} = [\,\mathbf{SG}(\mathbf{h}_{\tau}^{n}) \circ \mathbf{h}_{\tau+1}^{n-1}] \tag{3}$$

where the cached hidden state $\mathbf{h}_{\tau}^{n-1}$ in Eq. 1 and Eq. 2 is replaced with $\mathbf{h}_{\tau}^{n}$ in Eq. 3.

As shown in Fig. 2, when the retrospective feed mechanism is combined with the enhanced recurrence mechanism, every segment in the retrospective phase (shown in the box with a green dotted border) has bidirectional contextual information of the entire text input. We successfully modeled a larger effective context length (shown in the box with a orange dotted border) than traditional Recurrence Transformers can without extra memory and computation costs. Another benefit of the enhanced recurrence scheme is that past higher-level representations can be exploited to enrich future lower-level representations.

### 3.4 Segment-Reordering Objective

In addition to the **masked language model (MLM) objective** (Devlin et al., 2018), we introduce an additional document-aware task called **segment-reordering objective** for pretraining. Benefitting from the much larger effective context length provided by the enhanced recurrence mechanism, the goal of the segment-reordering objective is to predict the correct order for the permuted set of segments of a long document, to explicitly learn the relationships among segments. During the pretraining process of this task, a long text input $\mathcal{D}$ is first randomly partitioned into 1 to $m$ chunks; then, all the combinations are shuffled in a random order. As shown in Fig. 3, $\mathcal{D}$ is partitioned into three chunks and then permuted, that is, $\mathcal{D} = \{C_1, C_2, C_3\} \implies \hat{\mathcal{D}} = \{C_2, C_3, C_1\}$, where $C_i$ denotes the $i$-th chunk. Subsequently, the permuted long context $\hat{\mathcal{D}}$ is split into T sequential segments as a common practice, denoted as $\hat{\mathcal{D}} = \{S_1, S_2, ..., S_T\}$. We let the pretrained model reorganize these permuted segments, modeled as a $K$-class classification problem, where $K = \sum_{i=1}^{m} i!$.

The pretraining objective is summarized as follows for the $\tau$-th input segment:

$$\max_{\theta} \log p_\theta(S_\tau|\hat{S}_\tau) + \mathbb{1}_{\tau=T} \log p_\theta(\mathcal{D}|\hat{\mathcal{D}})$$

where $\hat{S}_\tau$ is the corrupted version of $S_\tau$, which is obtained by randomly setting a portion of tokens to `[MASK]`; $\hat{\mathcal{D}}$ is the permutated version of $\mathcal{D}$; $\theta$ is the model parameter; and $\mathbb{1}_{\tau=T}$ indicates that the segment-reordering objective is optimized only at the $T$-th step.

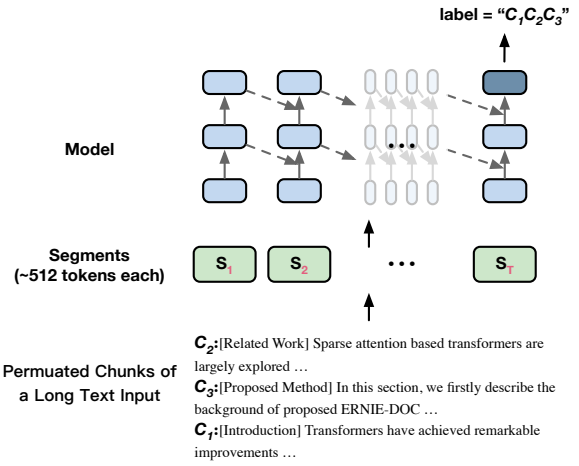


Figure 3: Illustrations of segment-reordering objective.

## 4 Experiments

### 4.1 Autoregressive Language Modeling

Autoregressive language modeling aims to estimate the probability distribution of an existing token/character based on previous tokens/characters in an input sequence. For comparison with previous work, we conducted experiments on word-level LM, that is, WikiText-103 (Merity et al., 2016), which is a document-level language modeling dataset.

#### 4.1.1 Experimental Setup

For autoregressive language modeling, we use a memory-enhanced Transformer-XL (Dai et al., 2019), that is, we employ our enhanced recurrence mechanism to replace the primitive one used in the Transformer-XL. Additionally, as proposed by Segatron (Bai et al., 2020), we introduce the segment-aware mechanism into Transformer-XL. Based on Transformer-XL, we trained a base-size model (L=16, H=410, A=10) and a large-size model (L=18, H=1,024, A=16)[3]. The models were trained for 200K/400K steps using a batch size of 64/128 for the base/large configurations. During the training phase, the sequence length and memory length were limited to 150 and 384 for the base and the large model, respectively. The remaining hyper-parameters were identical to those of Transformer-XL.

[3]We denote the number of Transformer layers as L, the hidden size as H, and the number of self-attention heads as A.

| Models | #Param. | PPL |
|---|---|---|
| *Results of* ***base models*** | | |
| LSTM (Grave et al., 2016) | - | 48.7 |
| LSTM+Neural cache (Grave et al., 2016) | - | 40.8 |
| GCNN-14 (Dauphin et al., 2017) | - | 37.2 |
| QRNN (Merity et al., 2018) | 151M | 33.0 |
| Transformer-XL Base (Dai et al., 2019) | 151M | 24.0 |
| SegaTransformer-XL Base (Bai et al., 2020) | 151M | 22.5 |
| ERNIE-DOC Base | 151M | **21.0** |
| *Results of* ***large models*** | | |
| Adaptive Input (Baevski and Auli, 2018) | 247M | 18.7 |
| Transformer-XL Large (Dai et al., 2019) | 247M | 18.3 |
| Compressive Transformer (Rae et al., 2019) | 247M | 17.1 |
| SegaTransformer-XL Large (Bai et al., 2020) | 247M | 17.1 |
| ERNIE-DOC Large | 247M | **16.8** |

Table 1: Comparison between Transformer-XL and competitive baseline results on WikiText-103.

#### 4.1.2 Results

Tab. 1 summarizes the evaluation results for WikiText-103. ERNIE-DOC achieves an impressive improvement compared with Transformer-XL: the perplexity (PPL) decreases by 3.0 for the base model and by 1.5 for the large model. Finally, we improve the state-of-the-art result of PPL to **21.0** (the base model) and **16.8** (the large model).

### 4.2 Pretraining and Finetuning

#### 4.2.1 Pretraining Text Corpora

| Dataset | # tokens | Avg len | Size |
|---|---|---|---|
| WIKIPEDIA | 2.7B | 480 | 8G |
| BOOKSCORPUS | 1.2B | 2,010 | 3.5G |
| CC-NEWS | 14B | 560 | 42G |
| STORIES | 7.5B | 1,891 | 22G |

Table 2: English datasets used for pretraining.

**English Data.** To allow ERNIE-DOC to capture long dependencies in pretraining, we compiled a corpus from four standard datasets: WIKIPEDIA, BOOKSCORPUS (Zhu et al., 2015), CC-NEWS[4], and STORIES (Trinh and Le, 2018) (details listed in Tab. 2). We tokenized the corpus using the RoBERTa wordpieces tokenizer (Liu et al., 2019) and duplicated the pretraining data 10 times.

**Chinese Data.** The Chinese text corpora used in ERNIE 2.0 (Sun et al., 2020b) were adopted for pretraining ERNIE-DOC.

#### 4.2.2 Experimental Setup

**Pretraining.** We trained three sizes of models for English tasks: small (L=6, H=256, A=4), base (L=12, H=768, A=12), and large (L=24, H=1,024, A=16). For Chinese tasks, we used only one size, i.e., base (L=12, H=768, A=12). We limited the length of the sentences in each mini-batch to 512 tokens and the length of the memory to 128. The models were trained for 500K/400K/100K steps using a batch size of 2,560/2,560/3,920 sentences for the small/base/large configurations. ERNIE-DOC was optimized with the Adam (Kingma and Ba, 2014) optimizer. The learning rate was warmed up over the first 4,000 steps to a peak value of 1e-4, and then it linearly decayed. The remaining pretraining hyperparameters were the same as those of RoBERTa (Liu et al., 2019) (see Tab. 12). Additionally, we employed relative positional embedding (Shaw et al., 2018) in our model pretraining because it is necessary for reusing hidden state without causing temporal confusion (Dai et al., 2019).

[4]We used `news-please` to crawl English news articles published between September 2016 and February 2019 and adopted Message Digest Algorithm5 (MD5) for deduplication.

| Models | IMDB | | HYP |
|---|---|---|---|
| | Acc. | F1 | F1 |
| RoBERTa (Liu et al., 2019) | 95.3 | 95.0 | 87.8 |
| Longformer (Beltagy et al., 2020) | 95.7 | - | 94.8 |
| BigBird (Zaheer et al., 2020) | - | 95.2 | 92.2 |
| ERNIE-DOC | **96.1** | **96.1** | **96.3** |
| XLNet-Large (Yang et al., 2019) | 96.8 | - | - |
| ERNIE-DOC-Large | **97.1** | **97.1** | **96.6** |

Table 3: Results on the IMDB and HYP dataset for long-text classification.

**Finetune.** In contrast to previous models, such as BERT, RoBERTa, and XLNet, the proposed model employs the retrospective feed mechanism and the enhanced recurrence mechanism during the finetuning phase to fully utilize the advantages of these two strategies.

#### 4.2.3 Results on English Tasks

**Results on Long-Text Classification Tasks**. We consider two datasets: IMDB reviews (Maas et al., 2011) and Hyperpartisan News Detection (HYP) (Kiesel et al., 2019). The former is a widely used sentiment analysis dataset containing 50,000 movie reviews, labeled as positive or negative. The latter contains news that takes extreme left-wing or right-wing standpoints. The documents in HYP are extremely long (50% of the samples contain more than 537 tokens) and are thus suitable for testing long-text classification ability. Tab. 3 summarizes the results of the ERNIE-DOC-Base and ERNIE-DOC-Large models for long-text classification tasks, and ERNIE-DOC achieves a SOTA result. On IMDB, we observed a modest perfor-

| Models | TQA | HQA | | |
|---|---|---|---|---|
| | F1 | Span | Supp | Joint |
| RoBERTa | 74.3 | 73.5 | 83.4 | 63.5 |
| Longformer | 75.2 | 74.3 | 84.4 | 64.4 |
| BigBird | 79.5 | 75.5 | **87.1** | 67.8 |
| ERNIE-DOC | **80.1** | **79.4** | 86.3 | **70.5** |
| Longformer-Large | 77.8 | 81.0 | 85.8 | 71.4 |
| BigBird-Large | - | 81.3 | **89.4** | - |
| ERNIE-DOC-Large | **82.5** | **82.2** | 87.6 | **73.7** |

Table 4: Results on TQA and HQA dev dataset for document-level QA. HQA metrics are F1.

| OpenKP dataset | F1@1 | F1@3 | F1@5 |
|---|---|---|---|
| BLING-KPE (Xiong et al., 2019) | 26.7 | 29.2 | 20.9 |
| JointKPE (Sun et al., 2020a) | 39.1 | 39.8 | 33.8 |
| ETC (Ainslie et al., 2020) | - | 40.2 | - |
| ERNIE-DOC | **40.2** | **40.5** | **34.4** |

Table 5: Results on OpenKP dev dataset. The baseline results are obtained from corresponding papers under no-visual-features setting.

mance gain compared with RoBERTa. This is because nearly 90% of the samples in the dataset consist of fewer than 569 tokens. Unlike on IMDB, ERNIE-DOC surpasses the baseline models on HYP by a substantial margin, demonstrating its capability of utilizing information from a long document input. Note that we include XLNet-Large, the previous SOTA pretraining model on the IMDB dataset, as the baseline for a large model setting; ERNIE-DOC achieves a result comparable to that of XLNet-Large.

**Results on Document-level Question-Answering Tasks**. We utilized two document-level QA datasets (Wikipedia setting of TriviaQA (TQA) (Joshi et al., 2017) and distractor setting of HotpotQA (HQA) (Yang et al., 2018)) to evaluate the reasoning ability of the models over long documents. TQA and HQA are extractive QA tasks, and we follow the simple QA model of BERT (Devlin et al., 2018) to predict an answer with the maximum sum of start and end logits across multiple segments of a sample. In addition, we use a modified cross-entropy loss (Clark and Gardner, 2017) for the TQA dataset and use a two-stage model (Groeneveld et al., 2020) with the backbone of ERNIE-DOC for the HQA dataset. Tab. 4. shows that ERNIE-DOC outperforms RoBERTa and Longformer by a considerable margin on these two datasets, and is comparable to current SOTA long-document model, i.e., BigBird on HQA in large-size model setting.

**Results on the Keyphrase Extraction Task**. We include OpenKP (Xiong et al., 2019) dataset to evaluate ERNIE-DOC's ability to extract keyphrases from a long document. Each document contains up to three short keyphrases and we follow the model setting of JointKPE (Sun et al., 2020a) and ETC (Ainslie et al., 2020) by applying CNNs on BERT's output to compose n-gram embeddings for classification. We report the results of base-size models in Tab. 5 under no-visual-features setting for easy and fair comparison with baselines. ERNIE-DOC performs stably better on all metrics on the OpenKP dataset.

#### 4.2.4 Results on Chinese Tasks

We conducted extensive experiments on seven Chinese natural language understanding (NLU) tasks, including machine reading comprehension (CMRC2018 (Cui et al., 2018), DRCD (Shao et al., 2018), DuReader (He et al., 2017), $C^3$ (Sun et al., 2019a)), semantic similarity (CAIL2019-SCM (CAIL) (Xiao et al., 2019)), and long-text classification (IFLYTEK (IFK) (Xu et al., 2020), THUCNews (THU)[5] (Sun et al., 2016)). The documents in all the aforementioned datasets are sufficiently long to be used to evaluate the effectiveness of ERNIE-DOC on long-context tasks (see detailed datasets statistics in Tab. 9). We reported the mean results with five runs for the seven Chinese tasks in Tab. 6, and summarized the hyperparameters in Tab. 16. ERNIE-DOC outperforms previous models across these Chinese NLU tasks by a significant margin in the base-size model group.

#### 4.2.5 Ablation Studies

| No. | Models | TQA | HYP |
|---|---|---|---|
| **I** | ERNIE-DOC | 64.56 | 86.10 |
| **II** | I w/o segment-reordering | 63.59 | 84.60 |
| **III** | II w/o retrospective feed | 63.38 | 83.27 |
| **IV** | III w/o enhanced recurrence | 61.09 | 81.67 |
| **V** | IV w/o recurrence | 58.35 | 77.72 |

Table 7: Performance of ERNIE-DOC-Small after ablating each proposed component (F1 result is reported).

**Effect of proposed components**. Tab. 7 shows the performance of ERNIE-DOC-Small on two English tasks after ablating each proposed component. All models were pretrained and fine-tuned with the same experimental setup, and we report the mean results of five runs. We observed a stable performance gain across these two tasks by incorporating each proposed component. By comparing

[5]We use a subset of THUCNews which can be found at https://github.com/gaussic/text-classification-cnn-rnn.

| Models | DRCD EM/F1 | | CMRC2018 EM/F1 | DuReader EM/F1 | CAIL Acc. | | THU Acc. | | IFK Acc. | $C^3$ Acc. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dev | Test | Dev | Dev | Dev | Test | Dev | Test | Dev | Dev | Test |
| BERT (Devlin et al., 2018) | 85.7/91.6 | 84.9/90.9 | 66.3/85.9 | 59.5/73.1 | 61.9 | 67.3 | 97.7 | 97.3 | 60.3 | 65.7 | 64.5 |
| BERT-wwm-ext* | 85.0/91.2 | 83.6/90.4 | 67.1/85.7 | -/- | - | - | 97.6 | 97.6 | 59.4 | 67.8 | 68.5 |
| RoBERTa-wwm-ext* | 86.6/92.5 | 85.2/92.0 | 67.4/87.2 | -/- | - | - | - | - | 60.3 | 67.1 | 66.5 |
| MacBERT (Cui et al., 2020a) | 88.3/93.5 | 87.9/93.2 | 69.5/87.7 | -/- | - | - | - | - | - | - | - |
| XLNet-zh (Cui et al., 2020b) | 83.2/92.0 | 82.8/91.8 | 63.0/85.9 | -/- | - | - | - | - | - | - | - |
| ERNIE 1.0 (Sun et al., 2019b) | 84.6/90.9 | 84.0/90.5 | 65.1/85.1 | 57.9/72/1 | - | - | 97.7 | 97.3 | 59.0 | 65.5 | 64.1 |
| ERNIE 2.0 (Sun et al., 2020b) | 88.5/93.8 | 88.0/93.4 | 69.1/88.6 | 61.3/74.9 | 64.9 | 67.9 | 98.0 | 97.5 | 61.7 | 72.3 | 73.2 |
| ERNIE-DOC | **90.5/95.2** | **90.5/95.1** | **76.1/91.6** | **65.8/77.9** | **65.6** | **68.8** | **98.3** | **97.7** | **62.4** | **76.5** | **76.5** |

Table 6: Results on seven Chinese NLU tasks for ERNIE-DOC-base model.The results of the models with "∗" are from Cui et al. (2019). The XLNet-zh is the abbreviation of Chinese-XLNet. Notably, the result of BERT on CAIL was obtained from Xiao et al. (2019), where BERT was post-pretrained with a legal dataset.

No.IV and No.V, we see that segment-level recurrence is necessary for modeling long documents and produces 2.74 and 3.95 % points improvement on the TQA and HYP dateset, respectively. Moreover, a substantial improvement is achieved using the enhance recurrence mechanism (2.29% point on TQA and 1.40% point on HYP, see No.III - IV). Retrospective feed mechanism further improves 0.21% point on TQA and 1.33% point on HYP (No.II - No.III). Considering different types of tasks, we observe that on HYP, an extremely long text classification dataset, a substantial improvement is achieved using the segment-reordering objective (1.5% point). This indicates that the `[CLS]` token, pretrained using the segment-reordering objective, is more adaptable to the document-level text classification task.

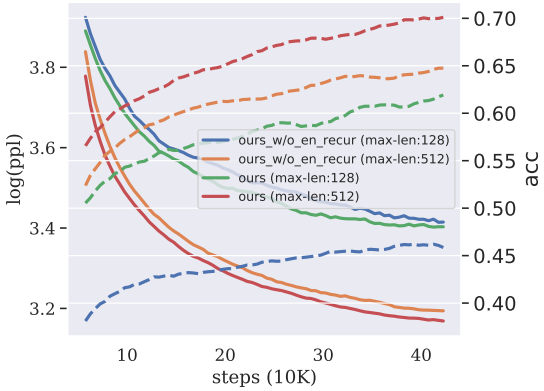


Figure 4: Acc. (dotted line) and PPL (solid line) metrics for variants of our small models with different maximum sequence length during pretraining.

**Effect of enhanced recurrence mechanism with regard to different maximum sequence lengths**. As depicted in Fig. 4, the enhanced recurrence mechanism plays an important role in pretraining an effective language model with lower PPL and higher accuracy under both the maximum sequence input lengths of 128 and 512. The effect of the enhanced recurrence mechanism is more significant under a smaller maximum sequence length, even makes the ERNIE-DOC-Small (max-len:128) comparable to ERNIE-DOC-Small_w/o_en_recur (max-len:512) w.r.t accuracy. This intriguing property of the enhanced recurrence mechanism enables more efficient model training and inference by reducing maximum sequence length while remaining comparable modeling capability.

## 5 Conclusion

In this paper, we proposed ERNIE-DOC, a document-level language pretraining model based on the Recurrence Transformers paradigm. Two well-designed mechanisms, namely the retrospective feed mechanism and the enhanced recurrent mechanism, enable ERNIE-DOC, which theoretically has the longest possible dependency, to model bidirectional contextual information of a complete document. Additionally, ERNIE-DOC is pretrained with a document-aware segment-reordering objective to explicitly learn the relationship among segments of a long context. Experiments on various downstream tasks demonstrate that ERNIE-DOC outperforms existing strong pretraining models such as RoBERTa, Longformer, and BigBird and achieves SOTA results on several language modeling and language understanding benchmarks. In future studies, we will evaluate ERNIE-DOC on language generation tasks, such as generative question answering and text summarization. We will also investigate its potential applicability in other areas, such as computational biology. Another possibility is to incorporate graph neural networks into ERNIE-DOC to enhance its modeling capability

for tasks that require multi-hop reasoning and long-document modeling ability.

## Acknowledgments

This work was supported by the National Key Research and Development Project of China (No. 2018AAA0101900).

## References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284.

Alexei Baevski and Michael Auli. 2018. Adaptive input representations for neural language modeling. *arXiv preprint arXiv:1809.10853*.

He Bai, Peng Shi, Jimmy Lin, Yuqing Xie, Luchen Tan, Kun Xiong, Wen Gao, and Ming Li. 2020. Segatron: Segment-aware transformer for language modeling and understanding.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509.

Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020a. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020b. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2016. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*.

Dirk Groeneveld, Tushar Khot, Ashish Sabharwal, et al. 2020. A simple yet strong pipeline for hotpotqa. *arXiv preprint arXiv:2004.06753*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467*.

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 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.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An analysis of neural language modeling at multiple scales. *arXiv preprint arXiv:1803.08240*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *CoRR*, abs/1911.05507.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019a. Probing prior knowledge needed in challenging chinese machine reading comprehension. *CoRR*, abs/1904.09679.

M Sun, J Li, Z Guo, Z Yu, Y Zheng, X Si, and Z Liu. 2016. Thuctc: an efficient chinese text classifier. *GitHub Repository*.

Si Sun, Chenyan Xiong, Zhenghao Liu, Zhiyuan Liu, and Jie Bao. 2020a. Joint keyphrase chunking and salience ranking with bert. *arXiv preprint arXiv:2004.13639*.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020b. Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*, pages 8968–8975.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse sinkhorn attention. *arXiv preprint arXiv:2002.11296*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Heng Wang, Jianfeng Xu, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Lee Xiong, Chuan Hu, Chenyan Xiong, Daniel Campos, and Arnold Overwijk. 2019. Open domain web keyphrase extraction beyond language modeling. *arXiv preprint arXiv:1911.02671*.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. 2019. Bp-transformer: Modelling long-range context via binary partitioning. *arXiv preprint arXiv:1911.04070*.

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. *Advances in neural information processing systems*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

# A Appendices

## A.1 Tasks

Following previous work, we evaluate ERNIE-DOC on various tasks that require the ability to model a long document.

**Document-level Language Modeling Task.** We employ WikiText-103 (Merity et al., 2016) in language modeling experiments. WikiText-103 is the largest available word-level benchmark with long-term dependency for language modeling, which consists of 28K articles, where each article has 3.6K tokens on average, thus 103M training tokens in total.

**Long Text classification**. We consider two English datasets: IMDB reviews (Maas et al., 2011) and Hyperpartisan news detection (Kiesel et al., 2019) (see Tab. 8), and two Chinese datasets: IFLYTEK (Xu et al., 2020) and THUCNews (Sun et al., 2016) (see Tab. 9). IMDB is a widely used sentiment analysis dataset containing 50,000 movie reviews labeled as positive or negative. Training and dev dataset is equally split. Hyperpartisan contains news that takes an extreme left-wing or right-wing standpoint. Documents are extremely long in Hyperpartisan which makes it a good test for long text classification. We use the same split as Longformer by dividing 654 documents into train/dev/test sets. IFLYTEK contains 17,332 app descriptions. The task is to assign each description into one of 119 categories, such as food, car rental and education. THUCNews is generated by filtering historical data of Sina News RSS subscription channel from 2005 to 2011, including 740,000 news documents and 14 categories. In this paper, we employ the subset version instead of the full one [6], which contains 10 categories, each with 5,000 pieces of data.

For the above four long text classification datasets, we concatenate [CLS] token with each segment and takes as input multiple segments of a text sequentially. Each segment is generated by slicing the text with a sliding window of 128 tokens. We apply binary cross entropy loss on the [CLS] token of the last segment.

**Long Text Semantic Similarity.** Considering that there is no available long text semantic similarity dataset in English, we evaluate the effectiveness of ERNIE-DOC on semantic similarity task only depending on Chinese dataset CAIL2019-SCM. According to Xiao et al. (2019), CAIL2019-SCM is a sub-task of the **C**hinese **AI** and **L**aw Challenge (CAIL) competition in 2019, which contains 8,964 triplets of legal documents collected from China Judgments Online. Every document in a majority of triplet has more than 512 characters, therefore, the total length of a triplet is quite long. CAIL2019-SCM requires researchers to decide which two cases are more similar in a triplet. Specifically, given a triplet $(A, B, C)$, where A, B, C are fact descriptions of three cases. The model needs to predict whether $sim(A, B) > sim(A, C)$ or $sim(A, C) > sim(A, B)$, in which $sim$ denotes the similarity between two cases. Instead of separately feeding the document A, B, C into the model to get the feature $h$, we use the combinations of $(A, B)$ and $(A, C)$ as input. We generate multiple segments for $(A, B)$ or $(A, C)$ with a sliding window of 128 tokens and feed them as input sequentially. The binary cross entropy loss is applied to the difference of [CLS] token output of each segment.

**Document-level Question answering**. We utilize two English question answering datasets (TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018)) (see Tab. 8) and four Chinese question answering datasets (CMRC2018 (Cui et al., 2018), DRCD (Shao et al., 2018), DuReader (He et al., 2017), $C^3$ (Sun et al., 2019a)) (see Tab. 9) to evaluate models' reasoning ability over long documents.

TriviaQA is a large scale QA dataset that contains over 650K question-answer pairs. We evaluate models on its Wikipedia setting where documents are Wikipedia articles, and answers are named entities mentioned in multiple documents. The dataset is distantly supervised meaning that there is no golden span, thus we find all superficial identical answers in provided documents[7]. We use the following input format for each segment: "`[CLS] context [q] question [/q]`" where context is generated by slicing multi-documents input with a sliding window of 128 tokens. We take as input multiple segments of a sample sequentially and attach a linear layer to each token in a segment to predict the answer span. We

[6] The subset version is also released and can be downloaded from the official website of THUCTC.

[7] We use the same preprocessing code for TriviaQA dataset as BigBird, see https://github.com/tensorflow/models/blob/master/official/nlp/projects/triviaqa/preprocess.py

| Datasets | IMDB | | Hyperpartisan | | | TriviaQA | | HotpotQA | | OpenKP | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| split | train | dev | train | dev | test | train | dev | train | dev | train | dev |
| # samples | 25,000 | 2,000 | 516 | 64 | 65 | 61,888 | 7,993 | 90,432 | 7,404 | 134,894 | 6,616 |
| # tokens of context length in each percentile using RoBERTa wordpiece tokenizer | | | | | | | | | | | |
| 50% | 215 | 212 | 537 | 521 | 639 | 8,685 | 8,586 | 1,279 | 1,325 | 894 | 681 |
| 90% | 569 | 550 | 1,519 | 1,539 | 1,772 | 25,207 | 24,825 | 1,725 | 1,785 | 3,451 | 2,734 |
| 95% | 745 | 724 | 1,997 | 1,979 | 1,994 | 32,018 | 32,132 | 1,888 | 1,943 | 5,340 | 4,130 |
| max | 3,084 | 2,778 | 5,566 | 2,643 | 5,566 | 173,302 | 146,012 | 3,733 | 3,618 | 105,548 | 43,609 |

Table 8: English Datasets statistics.

| Datasets | IFLYTEK | | THUCNews | | CAIL | | CMRC2018 | | DuReader | | $C^3$ | | DRCD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| split | train | dev | train | dev | train | dev | train | dev | train | dev | train | dev | train | dev | test |
| # samples | 12,133 | 2,599 | 50,000 | 5,000 | 5,102 | 1,500 | 10,121 | 3,219 | 15,763 | 1,628 | 11,869 | 3,816 | 26,936 | 3,524 | 3,493 |
| # tokens of context length in each percentile using BERT tokenizer | | | | | | | | | | | | | | | |
| 50% | 243 | 242 | 656 | 579 | 1,837 | 1,834 | 423 | 426 | 163 | 182 | 96 | 89 | 397 | 421 | 405 |
| 90% | 507 | 508 | 1,821 | 1,599 | 1,965 | 1,962 | 745 | 771 | 550 | 567 | 591 | 554 | 616 | 666 | 626 |
| 95% | 563 | 560 | 2,455 | 2,245 | 2,008 | 1,995 | 827 | 840 | 652 | 667 | 697 | 692 | 709 | 740 | 736 |
| max | 3,153 | 1,698 | 26,659 | 9,128 | 2,400 | 2,310 | 970 | 961 | 1,021 | 854 | 1,534 | 1,167 | 1,678 | 989 | 950 |

Table 9: Chinese Datasets statistics.

use a modified cross entropy loss (Clark and Gardner, 2017) assuming that each segment contains at least one correct answer span. The final prediction for each question is a span with the maximum sum of start and end logit across multiple segments.

HotpotQA is a QA dataset where golden spans of an answer and sentence-level supporting facts are provided. Thus, it contains two tasks namely, answer span prediction and supporting facts prediction. In the distractor setting, each question is associated with 10 documents where only 2 documents contain supporting facts. It requires the model to find and reason over multiple documents to find answers, and explain the predicted answers using predicted supporting facts. Following Groeneveld et al. (2020), we implemented a two-stage model based on ERNIE-DOC and use the following input format for each segment: "`[CLS]` $\text{title}_1$ `[p]` $\text{sent}_{1,1}$ `[SEP]` $\text{sent}_{1,2}$ `[SEP]` ... $\text{title}_2$ `[p]` $\text{sent}_{2,1}$ `[SEP]` $\text{sent}_{2,2}$ `[SEP]` ... `[q]` question `[/q]`" For evidence prediction, we apply 2 layer feedforward networks over the special token [SEP] and [p] representing a sentence and a paragraph separately. Then we use binary cross entropy loss to do binary classification. For answer span prediction, we train the model with a multi-task objective: 1) question type (yes/no/span) classification on the [CLS] token. 2) supporting evidence prediction on [SEP] and [p]. 3) span prediction on the start and end token of a golden span.

CMRC2018, DRCD and DuReader are common Chinese QA datasets with same format, which have been evaluated in numerous popular pretraining models, such as BERT (Devlin et al., 2018), ERNIE 1.0 (Sun et al., 2019b), ERNIE 2.0 (Sun et al., 2020b) and etc. The detailed descriptions of three datasets can refer to Cui et al. (2018), Shao et al. (2018) and He et al. (2017). We adopt the same input format as TriviaQA for each segment, denotes as "`[CLS] context [SEP] question [SEP]`" where context is generated by slicing multi-documents input with a sliding window of 128 tokens. We take as input multiple segments of a sample sequentially and attach a linear layer to each token in a segment to predict the answer span. Then, we apply a softmax and use the cross entropy loss with the correct answer. The final prediction for each question is a span with the maximum sum of start and end logit across multiple segments.

The multiple **C**hoice **C**hinese machine reading **C**omprehension dataset ($C^3$) (Sun et al., 2019a) is the first Chinese free-form multi-choice dataset where each question is associated with at most four choices and a single document. According to (Sun et al., 2019a), $m$ segments are constructed for a question, in which $m$ denotes the number of choice for that question. We use the following input format for each segment: "`[CLS] context [SEP] question [SEP]` $\text{choice}_i$ `[SEP]` " where context is generated by slicing document input with a sliding window of 128 tokens stride. We take as input multiple segments of a sample in a single batch and attach a linear layer to `[CLS]` that outputs an unnormalized logit. Then we obtain the final prediction for a question by applying a softmax layer over the unnormalized logits of all choices

| | QA | | Classification | | |
|---|---|---|---|---|---|
| **Models\Dataset** | **TriviaQA** | **HotpotQA** | **IMDB** | **Hyperpartisan** | **Avg.** |
| #0 ERNIE-DOC | 64.56 | 50.85 | 93.14 | 86.10 | **73.66** |
| #1 w/o so | 63.59 | 50.04 | 93.15 | 84.60 | 72.85 |
| #2 w/o so&retro | 63.38 | 49.87 | 92.56 | 83.27 | 72.27 |
| #3 w/o so&retro&en-rec | 61.09 | 44.05 | 92.07 | 81.67 | 69.72 |
| #4 w/o so&retro&recur | 58.35 | 31.54 | 91.60 | 77.72 | 64.80 |

Table 10: Performance of ERNIE-DOC-small after ablating each proposed component. (**so** denotes the segment-reordering objective, **re** denotes the retrospective feed mechanism, **en-rec** denotes the enhanced recurrence mechanism, and **recur** denotes the segment-level recurrence module. We used the Acc. metric for IMDB, F1 metric for TriviaQA and Hyperpartisan, Joint-F1 for HotpotQA.)

associated with it.

**Keyphrase Extraction**. We include OpenKP (Xiong et al., 2019) dataset [8] to evaluate ERNIE-DOC's ability to extract keyphrases from a long document. Each document contains up to three short keyphrases and we follow the model setting of JointKPE (Sun et al., 2020a) and ETC (Ainslie et al., 2020) by applying CNNs on BERT's output to compose n-gram embeddings for classification. We clean the dataset by removing some nonsense words such as the HTTP links. In detail, we apply five CNNs on BERT's output with the kernel size ranging from 1 to 5. Since each word is composed of several sub-tokens, we take the first token's embedding as the input for CNNs. Finally, we use the binary cross entropy loss as the optimization objective.

### A.2 Ablation Studies

Tab. 10 shows the performance of ERNIE-DOC-Small on English tasks after ablating each proposed component. All models were pretrained and fine-tuned with the same experimental setup, and we report the mean results of five runs. In the last column in Tab. 10, we see that the segment-reordering objective is improved ERNIE-DOC by 0.81% on average (#1 - #0), the retrospective feed mechanism is improved ERNIE-DOC by an average of 0.58% (#2 - #1), and the enhanced recurrence mechanism makes a large contribution of 2.55 percentage points on average (#3 - #2). By comparing #3 with #4, we see that segment-level recurrence is necessary for modeling long documents and produces a 4.92 percentage point improvement on average. Considering different types of tasks, we observe that on Hyperpartisan, an extremely long text classification dataset, a substantial improvement is achieved using the segment-reordering objective (1.5% point). This indicates that the `[CLS]` token, pretrained using the segment-reordering objective, is more adaptable to the document-level text classification task. Moreover, we observed a stable performance gain across all tasks using the enhanced recurrence mechanism.

[8] The dataset can be downloaded from https://github.com/thunlp/BERT-KPE

### A.3 Hyperparameters for Language Modeling

In Tab. 11, we present the detailed hyperparameters used for our experiments, which are the same as the hyperparameters employed in Transformer-XL (Dai et al., 2019).

| **Hyperparameters** | **WikiText-103 Base** | **WikiText-103 Large** |
|---|---|---|
| Layers | 16 | 18 |
| Hidden size | 410 | 1,024 |
| Attention heads | 10 | 16 |
| Training sequence length | 150 | 384 |
| Training memory length | 150 | 384 |
| Testing sequence length | 64 | 128 |
| Testing sequence length | 640 | 1,600 |
| Batch size | 64 | 128 |
| Learning rate | 2.5e-4 | 2.5e-4 |
| Warmup steps | 0 | 16,000 |
| Training steps | 200k | 400k |

Table 11: Hyperparameters used for WikiText-103.

### A.4 Hyperparameters for Pre-Training

As shown in Tab. 12, we present the detailed hyperparameters adopted to pretraining ERNIE-DOC on English text corpora and Chinese text corpora. For comparisons, we follow the same optimization hyperparameters of $\text{RoBERTa}_{\text{BASE}}$ or $\text{RoBERTa}_{\text{LARGE}}$ (Liu et al., 2019) for base-size or large-size model in English domain. As for Chinese ERNIE-DOC, we follow the same optimization hyperparameters of $\text{ERNIE 2.0}_{\text{BASE}}$.

### A.5 Hyperparameters for Fine-Tuning

#### A.5.1 Long Text Classification tasks

The finetuning hyperparameters for IMDB (Maas et al., 2011) and Hyperpartisan (Kiesel et al., 2019)

| Hyperparameters | English | | Chinese |
|---|---|---|---|
| | **BASE** | **LARGE** | **BASE** |
| Layers | 12 | 24 | 12 |
| Hidden size | 768 | 1,024 | 768 |
| Attention heads | 12 | 16 | 12 |
| Training steps | 400K | 100K | 300K |
| Batch size | 2,560 | 3,920 | 2,560 |
| Learning rate | 1e-4 | 1e-4 | 1e-4 |
| Warmup steps | 4,000 | 4,000 | 4,000 |
| Adam (beta1,beta2) | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) |
| Adam (epsilon) | 1e-6 | 1e-6 | 1e-6 |
| Learning rate schedule | Linear | Linear | Linear |
| Weight decay | 0.01 | 0.01 | 0.01 |
| Dropout | 0.1 | 0.1 | 0 |
| GPU (Nvidia V100) | 40 | 80 | 40 |

Table 12: Hyperparameters used for ERNIE-DOC pretraining.

are presented in Tab. 13.

| Hyperparameters | BASE | | LARGE | |
|---|---|---|---|---|
| | **IMDB** | **HYP** | **IMDB** | **HYP** |
| Batch size | 32 | 32 | 32 | 16 |
| Learning rate | 7e-5 | 1e-4 | 1e-5 | 4e-6 |
| Epochs | 3 | 15 | 3 | 15 |
| LR schedule | linear | linear | linear | linear |
| Layerwise LR decay | 1 | 0.7 | 0.9 | 1 |
| Warmup proportion | 0.1 | 0.1 | 0.1 | 0.1 |
| Weight decay | 0.01 | 0.01 | 0.01 | 0.01 |

Table 13: Hyperparameters used for finetuning on IMDB and Hyperpartisan (HYP).

### A.5.2 Document-level Question answering tasks

The finetuning hyperparameters for TriviaQA (Welbl et al., 2018) and HotpotQA (Yang et al., 2018) are presented in Tab. 14. HQA-sent. is the model for coarse-grained evidence prediction, and we choose the evidence with the probability larger than a pre-defined threshold 1e-3 and 1e-5 for base and large models, respectively. HQA-span. is the model for span prediction.

### A.5.3 Keyphrase Extraction task

The finetuning hyperparameters for the OpenKP (Xiong et al., 2019) dataset are presented in Tab. 15.

| Hyper. | BASE | | | LARGE | | |
|---|---|---|---|---|---|---|
| | **TQA** | **HQA-sent.** | **HQA-span.** | **TQA** | **HQA-sent.** | **HQA-span.** |
| Batch size | 64 | 128 | 128 | 64 | 32 | 32 |
| Learning rate | 3e-5 | 3e-5 | 1.5e-4 | 5e-6 | 5e-6 | 1.5e-5 |
| Epochs | 5 | 6 | 6 | 3 | 5 | 5 |
| LR schedule | linear | linear | linear | linear | linear | linear |
| Layer-decay | 0.8 | 1 | 0.8 | 0.9 | 0.9 | 0.9 |
| Warmup prop. | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Weight decay | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

Table 14: Finetuning hyperparameters on the TQA and HQA for base- and large-size ERNIE-DOC.

| Hyperparameters | OpenKP |
|---|---|
| Batch size | 32 |
| Learning rate | 1.5e-4 |
| Epochs | 5 |
| LR schedule | linear |
| Layerwise LR decay | 0.8 |
| Warmup proportion | 0.1 |
| Weight decay | 0.01 |

Table 15: Finetuning hyperparameters on the OpenKP for base-size ERNIE-DOC.

### A.5.4 Chinese NLU tasks

Tab. 16 lists the finetuning hyperparameters for Chinese NLU tasks including IFLYTEK (Xu et al., 2020), THUCNews (Sun et al., 2016), CMRC2018 (Cui et al., 2018), DRCD (Shao et al., 2018), DuReader He et al. (2017), $C^3$ (Sun et al., 2019a) and CAIL2019-SCM (Xiao et al., 2019).

| Tasks | Batch size | Learning rate | Epochs | Dropout |
|---|---|---|---|---|
| DRCD | 64 | 2.25-4 | 5 | 0.1 |
| CMRC2018 | 64 | 1.75e-4 | 5 | 0.2 |
| DuReader | 64 | 2.75e-4 | 5 | 0.1 |
| C3 | 24 | 1e-4 | 8 | 0.1 |
| CAIL | 48 | 5e-5 | 15 | 0.1 |
| THU | 16 | 1.5e-4 | 16 | 0.1 |
| IFK | 16 | 1.5e-4 | 5 | 0.1 |

Table 16: Hyperparameters used for finetuning on Chinese NLU tasks. Note that the warmup proportion are set to 0.1 and the layerwise learning rate decay rate are set to 0.8 for all tasks.

## B Attention Complexity

Given a long document with length $L$, Longformer and BigBird usually applies a local attention with a window size of 512 tokens on the entire input resulting in $L * 512$ token-to-token calculations. While the long document is fed twice as input and each input is sliced with a sliding window size of 512 tokens in ERNIE-DOC, which resulting in $2 * \frac{L}{512} * 512 * (512 + m)$ token-to-token calculations where $m$ is the memory length. Since $512 \ll L$ and $m \ll L$, the attention complexity of ERNIE-DOC is comparable to Longformer and BigBird which scales linearly with respect to the input length $L$, i.e., $O(L)$. Notably, the segments produced from the long document are fed one by one in ERNIE-DOC, leading to the lower spatial complexity.