Fast-StrucTexT: An Efficient Hourglass Transformer with Modality-guided Dynamic Token Merge for Document Understanding

Mingliang Zhai[1,2] , **Yulin Li**[2] , **Xiameng Qin**[2] , **Chen Yi**[2] , **Qunyi Xie**[2] , **Chengquan Zhang**[2]
Kun Yao[2] , **Yuwei Wu**[1,3,*] , **Yunde Jia**[1,3]

[1]Beijing Key Laboratory of Intelligent Information Technology, Beijing Institute of Technology, China
[2]Department of Computer Vision Technology (VIS), Baidu Inc.
[3]Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
{zhaimingliang, wuyuwei, jiayunde}@bit.edu.cn,
{liyulin03, qinxiameng, chenyi17, xiequnyi, zhangchengquan, yaokun01}@baidu.com

Abstract

Transformers achieve promising performance in document understanding because of their high effectiveness and still suffer from quadratic computational complexity dependency on the sequence length. General efficient transformers are challenging to be directly adapted to model document. They are unable to handle the layout representation in documents, *e.g.* word, line and paragraph, on different granularity levels and seem hard to achieve a good trade-off between efficiency and performance. To tackle the concerns, we propose **Fast-StrucTexT**, an efficient multi-modal framework based on the StrucTexT algorithm with an hourglass transformer architecture, for visual document understanding. Specifically, we design a modality-guided dynamic token merging block to make the model learn multi-granularity representation and prunes redundant tokens. Additionally, we present a multi-modal interaction module called Symmetry Cross-Attention (SCA) to consider multi-modal fusion and efficiently guide the token mergence. The SCA allows one modality input as query to calculate cross attention with another modality in a dual phase. Extensive experiments on FUNSD, SROIE, and CORD datasets demonstrate that our model achieves the state-of-the-art performance and almost $1.9\times$ faster inference time than the state-of-the-art methods.

1 Introduction

Visual document understanding technology aims to analyze visually rich documents (VRDs), such as document images or digital-born documents, enables to extract key-value pairs, tables, and other key structured data. At present, multi-modal pre-training transformer models [Lee *et al.*, 2022; Gu *et al.*, 2022; Huang *et al.*, 2022; Hong *et al.*, 2022] have
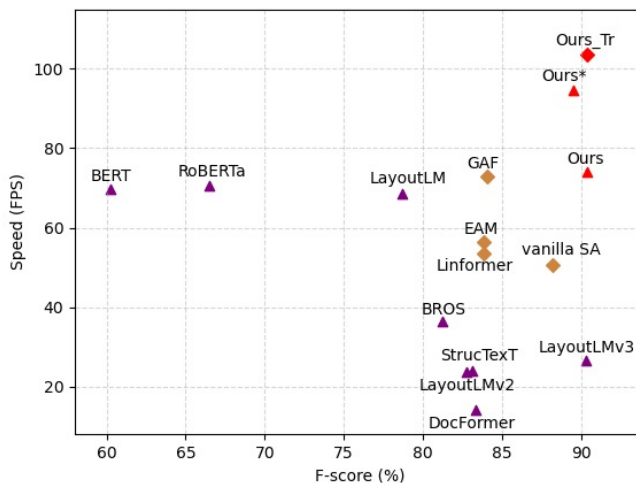
---

*Corresponding author

Figure 1: F-score (%) vs Inference Speed (FPS) on FUNSD test set. *Triangles* indicate the methods focus on document understanding, and the red triangle is our method. *Diamonds* mean that we utilize the efficient transformers in our framework, and the FPS is computed without image embeddings. The red diamond shows our method.

shown impressive performance in visual document understanding. The inside self-attention mechanism is crucial in modeling the long-range dependencies to capture contextual information. However, its quadratic computational complexity is the limitation of the transformer involved in visual document understanding with long sequences directly.

To improve the computational efficiency, there are three typical solutions. The first solution [Kim *et al.*, 2022; Hongxu *et al.*, 2022] simply reduces input sequence length by efficient sampling or new tokenization process. The second solution attempts to redesign transformer architectures, such as reformulating the self-attention mechanism [Verma, 2021; Wang *et al.*, 2022; Beltagy *et al.*, 2020] or producing a lightweight model [Jiao *et al.*, 2020; Wu *et al.*, 2022; Zhang *et al.*, 2022]. The third solution merges tokens by MLP [Ryoo *et al.*, 2021], grouping [Xu *et al.*, 2022a], or clustering [Marin *et al.*, 2023] to prune unnecessary infor-

(a) Visualization of Co-Attention matrix   (b) A weight map of the red rectangle in (a)   (c) The correlation of words and segments
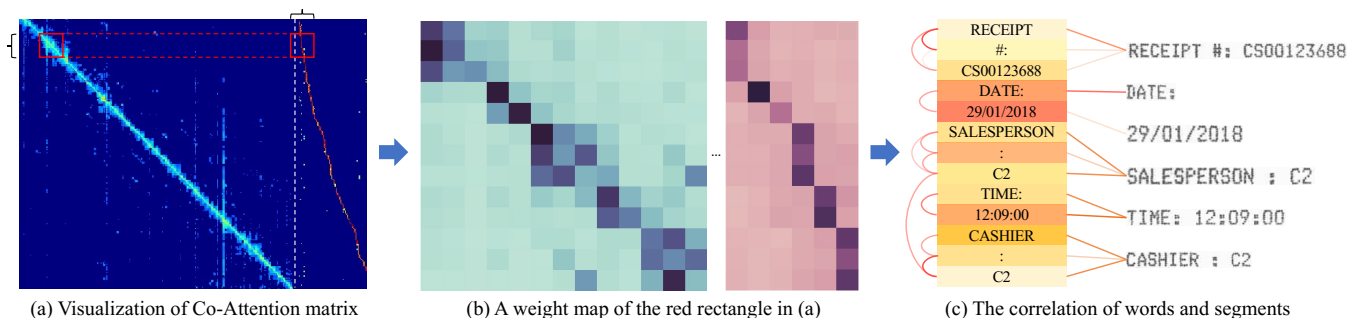
Figure 2: We accumulate the attention map of all layers to get (a), and then enlarge the part in the red box to get (b). The figure (c) is a semantic diagram according to the relationships in (b) and the orange line represents the correlation between words and segments.

mation related to several works [Pan *et al.*, 2022; Beltagy *et al.*, 2020] have proved the redundancies in attention maps.

Although those methods significantly reduce the computational complexity of the transformer model, they do not take into account the multiple granularity expressions in the visual documents, including words, lines, paragraphs, and so on. Furthermore, some token reduction methods learn multi-granularity information to a certain extent, but without considering the correlation between modality and granularity.

In Figure 2, we visualize the attention maps and token interactions of a standard transformer in the inference phase. In particular, Figure 2(a) displays the aggregated attention map of all Transformer layers. The left half shows a self-attention map inside words and the right half shows the cross-attention map between words and segments. We re-sample two regions with highlighted red boxes and zoomed in Figure 2(b). Moreover, for closer observation, Figure 2(c) gives the correlation visualization (curves and lines) of words and segments based on attention scores related to Figure 2(b). Almost the words and segments belonging to one semantic entity have significant correlations to their counterparts, indicating a high redundancy between irrelevant tokens in attention scores, which illustrates two key viewpoints:

1. Strong correlations are existed across granularities.

2. There is rich redundancy in attention computation.

In this paper, we propose an efficient multi-modal transformer called **Fast-StrucTexT** based on StrucTexT [Li *et al.*, 2021c], which is not only devoted to improving the model efficiency but also enhancing the expressiveness. We design an hourglass transformer that consists of merging and extension blocks, which receives the full-length sequence as input and compresses the redundant tokens progressively. A weighted correlation between two granularity features is produced by each merging block to dynamically guide the merging process. Since massive redundant information is eliminated, our model yields great efficiency gain and higher-level semantics. Further, we develop a Symmetry Cross-Attention (SCA) module, a dual cross-attention mechanism with multi-modal interaction, to conduct multi-modal feature fusion and provide modal semantic guidance for token merging. We pretrain Fast-StrucTexT with four task-agnostic self-supervised tasks for learning a good representation, and then fine-tune the pre-trained model in three benchmark datasets. Exper-

iment results demonstrate that our model achieves state-of-the-art performance and FPS.

The main contributions are summarized below:

- We propose Fast-StrucTexT with an efficient hourglass transformer by performing modal-guided dynamic token merging to reduce the number of tokens.

- We develop a dual multi-modal interaction mechanism named Symmetry Cross-Attention, which can enhance the multi-modal feature fusion from visual documents.

- Extensive experiments on four benchmarks show our model achieve state-of-the-art speed and performance.

2 Related Work

2.1 Multi-Modal Pre-training Model

As the first heuristic work, NLP-based approaches [Devlin *et al.*, 2019; Liu *et al.*, 2019] adopt the language model to extract the semantic structure. Various works [Hong *et al.*, 2022; Lee *et al.*, 2022; Li *et al.*, 2021a; Xu *et al.*, 2020] then jointly leverage layout information by spatial coordinates encoding, leading to better performance and extra computations simultaneously. After that, some researchers realize the effectiveness of deep fusion among textual, visual, and layout information from document images. A quantity of works [Gu *et al.*, 2021; Li *et al.*, 2021b; Li *et al.*, 2021c; Xu *et al.*, 2021; Gu *et al.*, 2022; Appalaraju *et al.*, 2021] rely on text spotting to extract semantic region features with a visual extractor. However, the transformer-based architectures are inefficient for long sequence modeling because of computationally expensive self-attention operations. To this end, we propose an efficient multi-modal model, Fast-StrucTexT, for visual document understanding.

2.2 Efficient Transformers

A well-known issue with self-attention is its quadratic time and memory complexity, which can impede the scalability of the transformer model in long sequence settings. Recently, there has been an overwhelming influx of model variants proposed to address this problem.

Fixed Patterns. The earliest modification to self-attention simply specifies the attention matrix by limiting the field of view, such as local windows and block patterns of fixed
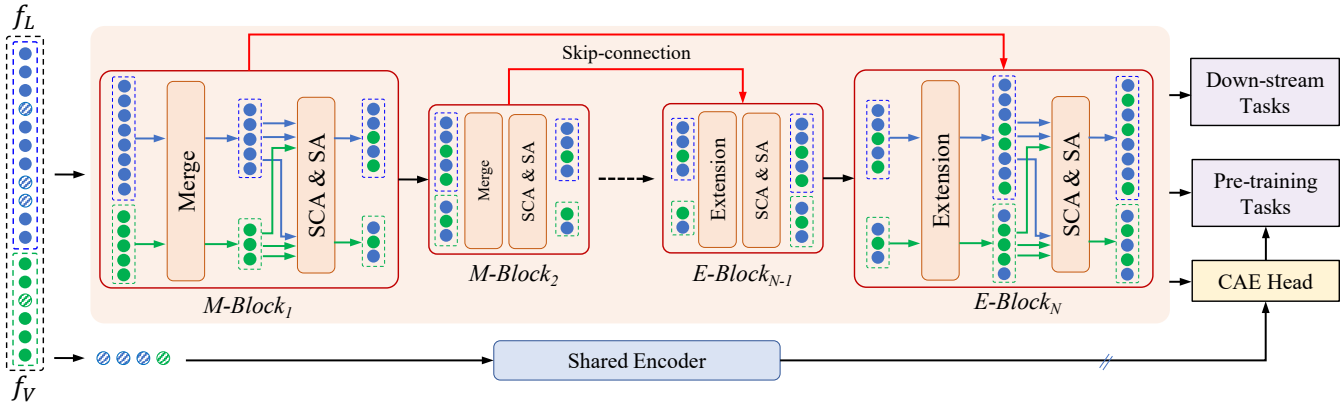
Figure 3: The overall Fast-StrucText architecture. Encoder is an hourglass architecture with the input of language and visual features. The merging operation can effectively reduce redundancy computations, and multi-modal interaction information is obtained through SCA & SA. The skip-connection exits from before merging to after extension.

strides. Sparse Transformer [Child *et al.*, 2019] converts the dense attention matrix to a sparse version by only computing attention on a sparse number of q_i, k_j pairs. GroupingViT [Xu *et al.*, 2022a] divides tokens into multiple groups, and then group-wised aggregates these tokens. Chunking input sequences into blocks that reduces the complexity from N^2 to B^2 (block size) with $B \ll N$, significantly decreasing the cost.

Down-sampling. Down-sampling methods that narrow the resolution of a sequence can effectively reduce the computation costs by a commensurate factor. Zihang Dai *et al.* [Dai *et al.*, 2020a] have highlighted the much-overlooked redundancy in maintaining a full-length token-level representation. To solve this problem, they compress the sequence of hidden states to a shorter length, thereby reducing the computation cost. TokenLearner [Ryoo *et al.*, 2021] uses MLP to project the tokens to low-rank space. The recent Nyströmformer [Xiong *et al.*, 2021] is a down-sampling method in which the "landmarks" are simply strided-based pooling in a similar spirit to Set Transformer [Lee Juho, 2019], Funnel Transformer [Dai *et al.*, 2020a], or Perceiver [Jaegle Andrew, 2021]. Inspired by those works, we design the hourglass transformer composed of modality-guided dynamic token merging and extension.

3 Method

In this section, we provide the framework of the proposed Fast-StrucTexT. First, we introduce the model architecture and describe the approach for generating multi-modal input features. Next, we present the details of the hourglass transformer with a hierarchical architecture. Finally, we explain the pre-training objectives and fine-tuning tasks. The overall Fast-StrucTexT architecture is depicted in Figure 3.

3.1 Model Architecture

Given a visual document and its text content extracted from OCR toolkits, the feature extractor first extracts both textual and visual features from the region proposals of text segments. These features are then fed into a transformer-based

encoder to learn multi-modal contextualized representations via alternating self-attention and cross-attention mechanisms. By leveraging redundancies across input tokens, the proposed encoder consisting of several Merging-Blocks and Extension-Blocks is designed as a hierarchical structure for shortening the sequence length progressively. In particular, the Merging-Block dynamically merges nearby tokens and the Extension-Block recovers the shortened sequence to the original scale according to the merging information to support token-level downstream tasks. Besides, we further adopt the alignment constraint strategy to improve the model ability by introducing a CAE [Chen *et al.*, 2022] head in the pre-training. The generated contextual representations can be used for fine-tuning downstream tasks of visual document understanding.

3.2 Feature Extraction

We employ an off-the-shelf OCR toolkit to a document image $I \in \mathbb{R}^{W \times H}$ to obtain text segments with a list of sentences S and corresponding 2D coordinates B. We then extract both segment-level visual features and word-level textual features through a ConvNet and a word embedding layer. For visual features f_V, the pooled RoI features of each text segment extracted by RoI-Align are projected to a vector. For textual features f_L, we utilize the WordPiece to tokenize text as sub-words and convert them to the ids.

We add a special start tag [CLS] with coordinates $B_0 = [0, 0, W, H]$ at the beginning of the input to describe the whole image. Besides, several [PAD] tags with zero bounding boxes append to the end of f_V, f_L to a fixed length. We ensure the length of f_L is an integral multiple of f_V.

3.3 Hourglass Encoder

We propose an hourglass transformer as the encoder module and progressively reduce the redundancy tokens. Instead of shortening the sequence length by a fixed pattern of merging tokens, the model performs a dynamic pooling operation under the guidance of semantic units (text sentence/segment), which leads to a shorter hidden states without harming model capability. The encoder consists of several Merging- and
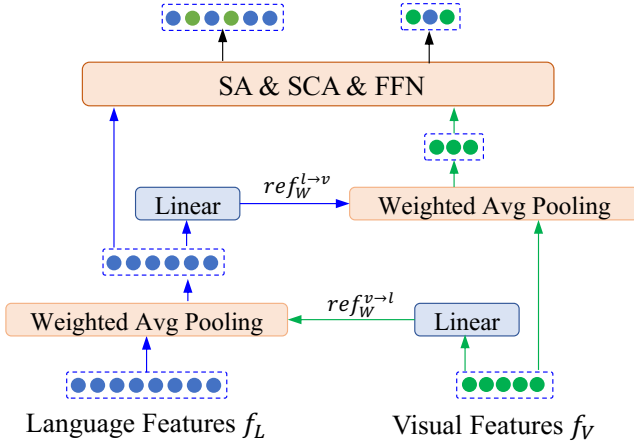
Figure 4: Proposed modality-guided dynamic token merge block.

Extension-blocks. The Merging-block (M-Block) merges the tokens of hidden states and the Extension-block (E-Block) conducts up-sampling to make up for the original length.

Merging. The M-Block suggests merging nearby k tokens with weighted 1D average pooling, where k is the shortening factor. In view of the multi-model hidden states, a referred weighting is predicted from another modality by a linear layer. The process is denoted in Figure 4.

$$f_n^i = \text{AVG\_POOL}(f_n^{i-1} \cdot \text{Linear}^{m \to n}(f_m^{i-1})), \quad (1)$$

where i is the stage index of M-Block, m and n denotes textual and visual modality ($m, n \in \{V, L\}$). Notably, Linear$^{* \to *}$ share their parameters for all M-Blocks.

Extension. The E-Block is required to transform shortened sequence of hidden states back to the entire token-level state. In detail, we simply apply repeat up-sampling method [Dai *et al.*, 2020b] to duplicate the vector of each merged token for k times. This method is computationally efficient. For maintaining the distinct information of tokens, we fed the hidden states from the corresponding M-Block into the E-Block through a skip-connection.

Symmetry Cross-Attention. Cross-Attention has shown the effectiveness of multi-modal learning on the vision-and-language tasks. To enhance the interactions among textual and visual modalities, we introduce Symmetry Cross-Attention (SCA) module to model the cross-modality relationships, which consists of two dual cross-attentions to handle the text and visual features in this work. We also leverage SCA to provide cross-modality guidance for token merging in the M-Blocks. The SCA is defined as follows:

$$\text{CA}(f_n|f_m) = W_o \sigma \left(\frac{F_q(f_n) F_k^T(f_m)}{\sqrt{d}} \right) F_v(f_m)$$

$$\text{SCA}(f_n, f_m) = \{\text{CA}(f_n|f_m), \text{CA}(f_m|f_n)\} \quad (2)$$

where F_q, F_k, F_v are the linear project for query, key and value. σ is the *Softmax* function, d is the hidden size and W_o is the output weight. The multi-head settings of attention are omitted for brevity.

Yet it's worth noting that our SCA incorporates semantic embedding as an additional input that gives the identical index for each segment and its corresponding words. SCA can provide multi-granularity interaction information for the subsequent token merging in addition to ensure the multi-modal interaction. Furthermore, a global context information be taken into account by Self-Attention (SA). Therefore, SCA and SA are adopted in turn to build transformer layers.

3.4 Pre-training Objectives

We adopt four self-supervised tasks simultaneously during the pre-training stage, which are described as follows.

Masked Visual-Language Modeling (MVLM) is the same as LayoutLMv2 [Xu *et al.*, 2021] and StrucTexT [Li *et al.*, 2021c]. Moreover, a CAE [Chen *et al.*, 2022] head is introduced to eliminate masked tokens in feature encoding and keep the consistency of document representation between pre-training and fine-tuning.

Graph-base Token Relation (GTR) constructs a ground truth matrix G with 0~9 to express the spatial relationship between each pairwise text segments. We give G_{ij} a layout knowledge, *i.e.*, 0 means the long distance (exceeding half the document size) between text segment i and j, and 1~9 indicate eight buckets of positional relations (up, bottom, left, right, top-left, top-right, bottom-left, bottom-right). We apply a bilinear layer in the segment-level visual features to obtain the pairwise features and fed them into a linear classifier driven by a cross-entropy loss.

Sentence Order Prediction (SOP) uses two normal-order adjacent sentences as positive examples, and others as negative examples. SOP aims to learn fine-grained distinctions about discourse-level coherence properties. Hence, the encoder is able to focus on learning semantic knowledge, and avoids the influence of the decoder.

Text-Image Alignment (TIA) is a fine-grained cross-modal alignment task. In the TIA task, some image patches are randomly masked with zero pixel values and then the model is pre-trained to identity the masked image patches according to the corresponding textual information. It enable the combine information between visual and text. Masked text token is not participating when estimating TIA loss.

3.5 Fine-tuning

We fine-tune our model on visual information extraction tasks: entity labeling and entity linking.

Entity Labeling. The entity labeling task aims to assign each identified entity a semantic label from a set of predefined categories. We perform an arithmetic average on the text tokens which belong to the same entity field and get the segment-level features of the text part. To yield richer semantic representations of entities containing the multi-modal information, we fuse the textual and visual features by the *Hadamard* product operation to get the final context features. Finally, a fully-connected layer with *Softmax* is built above the fused features to predict the category for each entity field.

Model	Modality	Image Embedding	FPS[†]	FLOPs (G)	Param. (M)	FUNSD F1	CORD F1	SROIE F1
BERT$_{BASE}$ [Devlin *et al.*, 2019]	T	None	69.77	48.36	**110**	60.26	89.68	93.67
RoBERTa$_{BASE}$ [Liu *et al.*, 2019]	T	None	70.47	48.36	125	66.48	93.54	-
LayoutLM$_{BASE}$ [Xu *et al.*, 2020]	T+L	None	68.68	48.36	113	78.66	94.72	94.38
BROS$_{BASE}$ [Hong *et al.*, 2022]	T+L	None	36.29	54.00	**110**	81.21	95.36	95.48
FormNet$_{A2}$ [Lee *et al.*, 2022]	T+L	None	-	-	217	84.69	97.10	-
StructuralLM$_{BASE}$ [Li *et al.*, 2021a]	T+L	None	-	-	113	78.66	-	-
UniDoc [Gu *et al.*, 2021]	T+L+I	ResNet-50	-	-	272	87.96	96.64	-
StrucTexT$_{BASE}$ [Li *et al.*, 2021c]	T+L+I	ResNet-50	24.11	82.21	107[‡]	83.09	-	96.88
DocFormer$_{BASE}$ [Appalaraju *et al.*, 2021]	T+L+I	ResNet-50	14.32	93.13	183	83.34	96.33	-
SelfDoc [Li *et al.*, 2021b]	T+L+I	ResNeXt-101	-	-	137	83.36	-	-
LayoutLMv2$_{BASE}$ [Xu *et al.*, 2021]	T+L+I	ResNeXt-101	15.50	91.45	200	82.76	94.95	96.25
LayoutLMv3$_{BASE}$ [Huang *et al.*, 2022]	T+L+I	Linear	39.55	55.95	133	<u>90.29</u>	96.56	-
Fast-StrucTexT*	T+L+I	Linear	**94.64**	**19.85**	<u>111</u>	89.50	<u>96.65</u>	97.12
Fast-StrucTexT	T+L+I	ResNet-18	<u>74.12</u>	<u>44.91</u>	116	**90.35**	**97.15**	**97.55**

Table 1: Entity labeling performance and model efficiency comparison on FUNSD, CORD, and SROIE datasets. "T/L/I" denotes "text/layout/image" modality. FPS[†] computation excludes the heads of downstream tasks. [‡] indicates the parameter number of StrucTexT [Li *et al.*, 2021c] is calculated without visual backbone of image embedding.

The Expression is shown as follow,

$$Z_F^i = (\frac{1}{k} \sum_k Z_L^{i,k}) * Z_V^i$$
$$P_i = \sigma(W^c Z_F^i) \tag{3}$$

where $W^c \in \mathbb{R}^{d \times d}$ is the weight matrix of the MLP layer and σ is the *Softmax* function. For entity e^i, Z_F^i indicates the final fused contextual features, k is the token length and P_i is the probability vector.

Entity Linking. The entity linking task desires to extract the relation between any two semantic entities. We use the bi-affine attention [Zhang *et al.*, 2021] for linking prediction, which calculates a score for each relation decision. In particular, for entity e^i and e^j, two MLPs are used to project their corresponding features Z_F^i and Z_F^j, respectively. After that, a bilinear layer W^b is utilized for the relation score $Score^{i,j}$.

$$X_k^i = W^k Z_F^i + b^k$$
$$X_v^j = W^v Z_F^i + b^v \tag{4}$$
$$Score^{i,j} = \sigma(X_k^i W^b X_v^j)$$

where W^k, W^v, W^b is the parameter weights that $\in \mathbb{R}^{d \times d}$, b^k and b^v is the bias. σ denotes the sigmoid activation function.

4 Experiment

In this section, we introduce several datasets used for visual document understanding. We then provide implementation details, including our pre-training and fine-tuning strategies for downstream tasks. We conclude with evaluations of Fast-StrucTexT on four benchmarks, as well as ablation studies.

4.1 Datasets

IIT-CDIP [Harley *et al.*, 2015] is a large resource for various document-related tasks and includes approximately 11 mil-lion scanned document pages. We pre-trained our model on this dataset.

FUNSD [Jaume *et al.*, 2019] is a form understanding dataset designed to address the challenges presented by noisy scanned documents. It comprises fully annotated training samples (149) and testing samples (50), and we focused on the semantic entity labeling and linking tasks.

CORD [Park *et al.*, 2019] is typically utilized for receipt key information extraction, consisting of 800 training receipts, 100 validation receipts, and 100 test receipts. We used the official OCR annotations and an entity-level F1 score to evaluate our model's performance.

SROIE [Huang *et al.*, 2021] is a scanned receipt dataset consisting of 626 training images and 347 testing images. Each receipt contains four predefined values: company, date, address, and total. We used the official evaluation tools to evaluate our model's performance.

EPHOIE [Wang *et al.*, 2021] is a collection of 1,494 actual Chinese examination papers with a rich text and layout distribution. The dataset is divided into 1,183 training images and 311 testing images, and each character in the document is annotated with a label from ten predefined categories.

4.2 Implementation

We followed the typical pre-training and fine-tuning strategies to train our model. For all pre-training and downstream tasks, we resized the images along their longer side and padded them to a size of 512×512. The input sequence is set to a maximum length of 640, with text tokens padded to a length of 512 and image tokens padded to a length of 128. We apply the PaddleOCR to extract the text segment.

Model Configurations. The hourglass encoder in our model comprises three M-Blocks and three E-Blocks. Each block consists of a SA and a SCA layer, with 12 heads, in addition to Merging or Extension operations. The encoder ar-

Model	Subject	Test Time	Name	School	#Exam	#Seat	Class	#Student	Grade	Score	Mean
GraphIE	94.00	100	95.84	97.06	82.19	84.44	93.07	85.33	94.44	76.19	90.26
TRIE [Zhang *et al.*, 2020]	98.79	100	99.46	99.64	88.64	85.92	97.94	84.32	97.02	80.39	93.21
VIES	99.39	100	99.67	99.28	91.81	88.73	99.29	89.47	98.35	86.27	95.23
WatchVIE [Tang *et al.*, 2021]	**99.78**	100	**99.88**	98.57	94.21	93.48	**99.54**	92.44	98.35	92.45	96.87
StrucTexT [Li *et al.*, 2021c]	99.25	100	99.47	**99.83**	**97.98**	95.43	98.29	97.33	**99.25**	93.73	97.95
Fast-StrucTexT	98.39	**100**	99.34	99.55	96.07	**97.22**	96.73	**100**	95.09	**99.41**	**98.18**

Table 2: Entity labeling performance comparison in Chinese on the EPHOIE test set.

chitecture is a 12-layer transformer with a hidden size of 768 and an intermediate size of the feed-forward networks, D_f, of 3072. The shortening factor k is set to 2 for all stages. The input sequence lengths of the three blocks in M-Blocks are $256, 128, 64$, successively, and vice versa in E-Blocks. We tokenize segment-level text by BERT [Devlin *et al.*, 2019], and trans to text sequence using One-Hot embedding with the vocabulary size L_v=30522. To pre-process the image sequence, we use ResNet18 [He *et al.*, 2016] pre-trained on ImageNet [Deng *et al.*, 2009] to extract ROI features with the size of $C \times H \times W = 128 \times 4 \times 64$, followed by a linear project of D_t=1024.

Pre-training. We pre-train our model on overall IIT-CDIP dataset. The model parameters are randomly initialized. While pre-training, we apply AdamX optimizer with a batch size of 64 for 1 epoch on 8 NVIDIA Tesla A100 80GB GPUs. The learning rate is set as 1×10^{-5} during the warm-up for 1000 iterations and the keep as 1×10^{-4}. We set weight decay as 10^{-4} and $(\beta_1 = 0.9, \beta_2 = 0.999)$.

Fine-tuning on Downstream Tasks. We fine-tune two downstream tasks: entity labeling and entity linking. For entity labeling task on FUNSD, CORD, and SROIE, we set the training epoch as 100 with a batch size of 8 and learning rate of $5 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-4}$, respectively.

4.3 Comparison with the State-of-the-Arts

We compare Fast-StrucTexT with BASE scale multi-modal transformer pre-trained models on public benchmarks. We evaluate our model on three benchmark datasets for entity labeling and entity linking tasks with metrics such as Frames Per Second (FPS), Floating Point Operations Per Second (FLOPs), Parameters, and F1 score.

Entity Labeling. The comparison results are exhibited in Table 1. Our Fast-StrucTexT achieves state-of-the-art FPS and outperforms the performance of the current state-of-the-art model by 156%. Fast-StrucTexT achieves a $1.9\times$ throughput gain and comparable F1 score on the CORD and SROIE entity labeling tasks. To demonstrate the effectiveness of Fast-StrucTexT in Chinese, we pre-train the model in a self-built dataset which consists of 8 million document images in Chinese, and fine-tune the pre-trained model on EPHOIE. Table 2 illustrates the overall performance of the EPHOIE dataset, where our model obtains the best result with 98.18% F1-score.

Model	FUNSD
BERT [Devlin *et al.*, 2019]	27.65
SPADE [Hwang *et al.*, 2021]	41.70
StrucTexT$_{BASE}$ [Li *et al.*, 2021c]	44.10
LayoutLM$^{\dagger}_{BASE}$ [Xu *et al.*, 2020]	48.00
LayoutXLM$_{BASE}$ [Xu *et al.*, 2022b]	54.83
BROS$_{BASE}$ [Hong *et al.*, 2022]	**67.63**
Fast-StrucTexT	<u>67.36</u>

Table 3: Entity linking performance comparison on FUNSD dataset. LayoutLM$^{\dagger}_{BASE}$ is implemented by [Hong *et al.*, 2022]. It's worth noticing that our method is $2\times$ faster than BROS.

Entity linking. As shown in Table 3, we compare Fast-StrucTexT with several state-of-the-art methods on FUNSD for entity linking. Compared with BROS [Hong *et al.*, 2022], our method achieves a comparable performance with $2\times$ speed in Table 1.

4.4 Ablation Studies

We conduct ablation studies on each component of the model on the FUNSD and SROIE datasets, including the backbone, pooling strategy, and shorten factor. At last, we evaluate the cost of our proposed hourglass transform with various sequence lengths.

Backbone. To prove the pre-trained Fast-StrucTexT can obtain state-of-the-art efficiency and performance. We replaced a variety of popular lightweight backbones [Guo *et al.*, 2021; Verma, 2021; Wang *et al.*, 2022] for evaluation. As shown in Table 4, Fast-StrucTexT can achieve the highest performance and FPS. The point here is that transformer takes responsibility for token feature representation, which could benefit from large model size. It is the reason why those lightweight architectures lead to worse performance. Particularly, we believe that those methods do not take into multi-modal interaction for visual document understanding. In Table 4, the model shows better efficiency and performance than other lightweight encoders and achieves FPS = 103.50, F1 = 90.35%. Comparing "Ours w/o M&E" and "Ours", we can obverse that merging and extension operations are not only efficient but also can improve performance. Specifically, it obtains $1.5\times$ the throughput of "w/o M&E" settings, and 0.41% improvement in the labeling task.

Model	FPS‡	Labeling
Vanilla SA	70.70	88.20
Linformer [Verma, 2021]	73.58	83.83
EAM [Guo *et al.*, 2021]	76.38	83.85
+M&E	91.42	86.98
GFA [Wang *et al.*, 2022]	72.82	84.08
+M&E	103.38	88.45
Fast-StrucTexT w/ only SA	88.88	89.59
Fast-StrucTexT w/o M&E	81.85	89.94
Fast-StrucTexT	**103.50**	**90.35**

Table 4: Using different transformer design in FUNSD on entity labeling tasks. FPS‡ only calculate the part of the encoder. "M&E" is merging and extension operations for the token.

Method	FPS‡	FUNSD
GlobalPool	**120.33**	86.62
AvgPool	107.46	88.01
DeformableAttention	83.14	88.57
Fast-StrucTexT	103.50	**90.35**

Table 5: Ablation of token merging strategies.

k	SROIE-F1	FUNSD-F1	CORD-F1	FPS	FLOPs
1	97.46	**90.54**	97.15	81.85	46.48G
2	**97.55**	90.35	**97.15**	103.50	19.85G
4	95.88	88.69	96.85	117.69	12.46G
8	92.22	83.07	93.85	133.48	10.14G

Table 6: Ablation study of the various k.

Pooling. Table 5 gives the results of different merging strategies for entity labeling task on FUNSD. GlobalPool is a form to directly merge all token-level features into segment-level before encoding. AvgPool is our merging strategy without cross-modal guidance. DeformableAttention [Xia *et al.*, 2022] attempts to learn several reference points to sample the sequence. The experimental results show the effectiveness of our token merging method.

Shorten factor. We study the hyper-parameter k as shown in Table 6. We have evaluated multiple benchmarks and have established that the setting k=2 is the best trade-off for document understanding. In addition, we can adjust the k in the fine-tuning stage. Nevertheless, ensuring consistency of k during pre-training and fine-tuning can take full advantage of knowledge from pre-trained data. Our framework supports multi-scale pre-training with a list of k factors to handle different shortening tasks.

Sequence length. Our token merging can adapt to the arbitrary length of the input. The ratio of merging is adjustable and determined by the value of k and the number of M-blocks and E-blocks. Referring to Table 7, we investigate the ability of our method with various sequence lengths. The experimental results show that the speed and computation gains become

Seq. Len.	V.T. (FPS)	Ours (FPS)	V.T. (FLOPs)	Ours (FLOPs)
1024	30.25	52.86 (+74%)	106.39G	41.57G (-60%)
2048	13.46	24.76 (+84%)	251.44G	90.12G (-64%)
4096	4.68	10.59 (+126%)	657.50G	208.16G (-68%)
8192	1.22	3.82 (+213%)	1933.49G	527.95G (-72%)

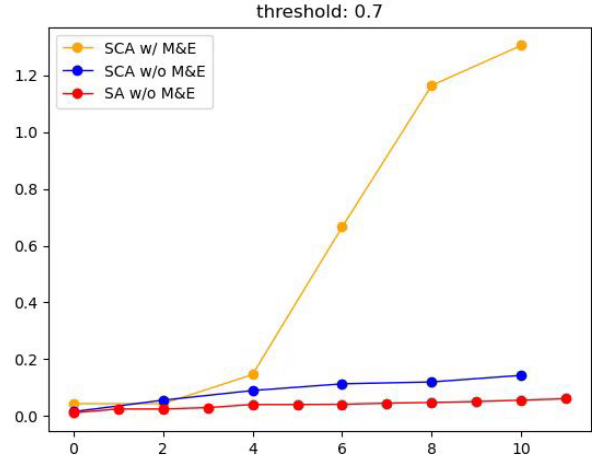Table 7: Ablation study of sequence length, where V.T. represents Vanilla Transformer.

Figure 5: The cumulative proportions of the filted tokens which probability beyond 0.7 in attention maps of transformers. SCA has better modal interaction ability.

more pronounced with the sequence length increasing.

Modal interaction ability. As shown in Fig 5, we conducted an ablation study on the hourglass architecture and SCA module. We calculated the number of more than 0.7 in each layer of the attention map and then accumulated it layer by layer to observe the utilization of each token. Compared with the yellow line and the blue line, the number of highly responsive areas in the later layer is significantly increased, which can show that after aggregating the multi-granularity semantics, each token has sufficient ability to express more information. The blue line is slightly higher than the red line, which indicates that our SCA has better modal interaction ability than SA.

5 Conclusion

In this paper, we present Fast-StrucTexT, an efficient transformer for document understanding task. Fast-StrucTexT significantly reduces the computing cost through hourglass transformer architecture, and utilizes multi-granularity information through modality-guided dynamic token merging operation. Besides, we propose the Symmetry Cross-Attention module to enhance the multi-modal interaction and reduce the computational complexity. Our model shows state-of-the-art performance and efficiency on four benchmark datasets.

Acknowledgments

References

[Appalaraju *et al.*, 2021] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. In *ICCV*, pages 973–983, 2021.

[Beltagy *et al.*, 2020] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, 2020.

[Chen *et al.*, 2022] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint:2202.03026*, 2022.

[Child *et al.*, 2019] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *CoRR*, 2019.

[Dai *et al.*, 2020a] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. In *NeurIPS*, pages 4271–4282, 2020.

[Dai *et al.*, 2020b] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. In *NeurIPS*, pages 4271–4282, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Gu *et al.*, 2021] Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. In *NeurIPS*, pages 39–50, 2021.

[Gu *et al.*, 2022] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *CVPR*, pages 4573–4582, 2022.

[Guo *et al.*, 2021] Menghao Guo, Zhengning Liu, Taijiang Mu, and Shimin Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *TPAMI*, pages 5436–5447, 2021.

[Harley *et al.*, 2015] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, pages 991–995, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hong *et al.*, 2022] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *AAAI*, 2022.

[Hongxu *et al.*, 2022] Yin Hongxu, Vahdat Arash, Alvarez Jose M, Mallya Arun, Kautz Jan, and Molchanov Pavlo. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, pages 10809–10818, 2022.

[Huang *et al.*, 2021] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar. *CoRR*, 2021.

[Huang *et al.*, 2022] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *ACM Multimedia*, 2022.

[Hwang *et al.*, 2021] Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. Spatial dependency parsing for semi-structured document information extraction. In *ACL*, pages 330–343, 2021.

[Jaegle Andrew, 2021] Alayrac Jean-Baptiste Doersch Carl Ionescu Catalin Ding David Koppula Skanda Zoran Daniel Brock Andrew Shelhamer Evan others Jaegle Andrew, Borgeaud Sebastian. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*, 2021.

[Jaume *et al.*, 2019] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDAR*, pages 1–6, 2019.

[Jiao *et al.*, 2020] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. In *EMNLP*, pages 4163–4174, 2020.

[Kim *et al.*, 2022] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *SIGKDD*, pages 784–794, 2022.

[Lee *et al.*, 2022] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. Formnet: Structural encoding beyond sequential modeling in form document information extraction. In *ACL*, pages 3735–3754, 2022.

[Lee Juho, 2019] Kim Jungtaek-Kosiorek Adam Choi Seungjin Teh Yee Whye Lee Juho, Lee Yoonho. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, pages 3744–3753, 2019.

[Li *et al.*, 2021a] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding. In *ACL*, pages 6309–6318, 2021.

[Li *et al.*, 2021b] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *CVPR*, pages 5652–5660, 2021.

[Li *et al.*, 2021c] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *ACM Multimedia*, pages 1912–1920, 2021.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

[Marin *et al.*, 2023] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In *WACV*, pages 12–21, 2023.

[Pan *et al.*, 2022] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martínez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *ECCV*, pages 294–311, 2022.

[Park *et al.*, 2019] Seunghyun Park, Seung Hyun Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. In *NIPS Workshop*, 2019.

[Ryoo *et al.*, 2021] Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *CoRR*, 2021.

[Tang *et al.*, 2021] Guozhi Tang, Lele Xie, Lianwen Jin, Jiapeng Wang, Jingdong Chen, Zhen Xu, Qianying Wang, Yaqiang Wu, and Hui Li. Matchvie: Exploiting match relevancy between entities for visual information extraction. In *IJCAI*, pages 1039–1045, 2021.

[Verma, 2021] Madhusudan Verma. Revisiting linformer with a modified self-attention with linear complexity. *CoRR*, 2021.

[Wang *et al.*, 2021] Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. Towards robust visual information extraction in real world: new dataset and novel solution. In *AAAI*, pages 2738–2745, 2021.

[Wang *et al.*, 2022] Jian Wang, Chenhui Gou, Qiman Wu, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Rtformer: Efficient design for real-time semantic segmentation with transformer. In *NeurIPS*, 2022.

[Wu *et al.*, 2022] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*, pages 68–85, 2022.

[Xia *et al.*, 2022] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, pages 4784–4793, 2022.

[Xiong *et al.*, 2021] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *AAAI*, pages 14138–14148, 2021.

[Xu *et al.*, 2020] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *ACM SIGKDD*, pages 1192–1200, 2020.

[Xu *et al.*, 2021] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*, pages 2579–2591, 2021.

[Xu *et al.*, 2022a] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18113–18123, 2022.

[Xu *et al.*, 2022b] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, and Furu Wei. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *ACL*, pages 3214–3224, 2022.

[Zhang *et al.*, 2020] Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. TRIE: end-to-end text reading and information extraction for document understanding. In *ACM Multimedia*, pages 1413–1422, 2020.

[Zhang *et al.*, 2021] Yue Zhang, Bo Zhang, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. Entity relation extraction as dependency parsing in visually rich documents. In *EMNLP*, pages 2759–2768, 2021.

[Zhang *et al.*, 2022] Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Minivit: Compressing vision transformers with weight multiplexing. In *CVPR*, pages 12135–12144, 2022.