LoPE: Learnable Sinusoidal Positional Encoding for Improving Document Transformer Model

Anonymous ACL submission

Abstract

Positional encoding plays a key role in Transformer-based architecture, which is to indicate and embed token sequential order information. Understanding documents with unreliable reading order information is a real challenge for document Transformer model. This paper proposes a new and generic positional encoding method, learnable sinusoidal positional encoding (LoPE), by combining sinusoidal positional encoding function and a learnable feed-forward network. We apply LoPE to document Transformer model and pretrain the model on document datasets. Then we finetune and evaluate the model performance on document understanding tasks in form and receipt domains. Experimental results not only show our proposed method outperforms other baselines and state-of-the-arts, but also demonstrate its robustness and stability on handling noisy data with incorrect order information.

1 Introduction

Document understanding (or in some contexts known as Document intelligence, Document AI) aims to extract, recognize and understand information from document images. The performance of document understanding model is largely benefited from recent development of large scale pretraining technique on cross-modality data and effective transformer architectures (Cui et al., 2021). Document Transformer Model, e.g. LayoutLM (Xu et al., 2020b), is pretrained from visually-rich document data which consists of text, layout and visual information based on Transformer architecture (Shaw et al., 2018). Recently, (Xu et al., 2020a; Hong et al., 2021; Appalaraju et al., 2021; Li et al., 2021a) propose various approaches to further improve the performance of Transformer model on more challenging document understanding tasks.

Different from recurrent and convolutional based structures, Transformer based model does not encode relative or absolute position information explicitly since it is solely based on order-invariant attentional mechanism. In the original Transformer architecture (Vaswani et al., 2017), both learnable vector embedding and sinusoidal function are introduced as positional encoding methods for capturing positional information from input tokens. In order to improve positional representation ability, (Shaw et al., 2018; Huang et al., 2020; He et al., 2021; Chi et al., 2021) introduce several relative position strategies into attention computation steps in Transformer. Along with sequential reading order from text, visually-rich documents contain more spatial information of text block which poses a greater challenge to understand rich semantic and spatial relationship information at same time. To obtain text blocks from document image, current off-the-shelf method is borrowing results from existing Optical Character Recognition (OCR) engine while mostly the reading order of text blocks is just arranged by a heuristic manner, top-to-bottom and left-to-right (Clausner et al., 2013; Wang et al., 2021). For documents with complex layout, such as forms, invoices or receipts, the performance of reading order is not consistent which always leads to irrelevant or embarrassing predictions (Cui et al., 2021). Moreover, existing Document Transformer Models suffer from huge performance degradation on noisy data with unreliable reading order information (Hong et al., 2021). Therefore positional encoding plays an essential role in document Transformer models, which is to encode position embedding from data with inherent reading or spatial information. Thus, it's crucial to improve the robustness and learnability of position encoding method, and boost the model performance on noisy data with unreliable order and spatial information.

In this paper, we introduce a learnable sinusoidal position encoding method, $LoPE$, by combining the sinusoidal positional encoding function and a learnable fully connected feed-forward network. And we apply it to represent multidimensional position information in document Transformer model. Compared with current discrete embedding layer in Transformer model, our method is numeric continuous for position scales which improve positional representation of relative position or distances between spatial elements. We enhance the original sinusoidal positional function by adding a learnable network which allows pretrained language model to adapt to various downstream tasks effectively. It keeps the advantage of extrapolability from sinusoidal function which could extend to longer position than training cases. We pretrain transformer model on document datasets with our positional encoding and baseline methods. Then we evaluate the model performance on document understanding downstream tasks and compare model performance with various positional encoding methods with the same input modality and model size setting. Experimental results illustrate that our $LoPE$ method significantly outperforms baseline methods and recent pretrained document language models on both FUNSD and SROIE benchmarks. In addition, we evaluate the model robustness on noisy order data by utilizing global and local shuffling augmentation strategies. Our method shows stable performance than other positional encoding methods with unreliable order information. Furthermore, we visualize and analyze similarity of positional representation for each method from the 1D to 2D positional embeddings of our pretrained models.

In summary, our contributions could be highlighted as follows: 1) We propose $LoPE$ as a new and generic learnable positional encoding method with better learnability and extrapolability to improve document Transformer model. 2) We pretrain document Transformer models with $LoPE$ and other baselines, and evaluate model performance on document understanding tasks. Experimental results show our proposed method outperforms other baselines and recent SOTA approaches on FUNSD and SROIE datasets.3) By ablation study of employing global and local block shuffling augmentation strategies, our method demonstrates optimal performance and robustness on noisy data with unreliable reading order information. Finally, our pretrained models with implementation of fine-tuning code will be open to public.[1]

[1]Our code will be made publicly available.

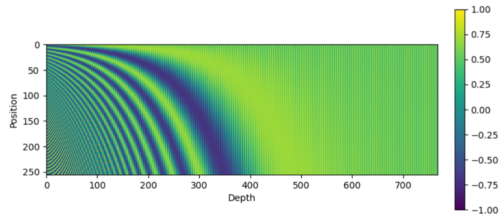


Figure 1: Visualization of 768-dimensional sinusoidal positional encoding for sequence with the maximum length of 256. Each position row p represents the embedding vector $PE_{sine}(p)$ as positional representation.

2 Background

Positional Encoding Methods in Transformer In the original proposal of Transformer architecture (Vaswani et al., 2017), both learnable vector and sinusoidal function are introduced as positional encoding methods and perform nearly identically in their downstream tasks. Although sinusoidal version with predefined wavelength has unique extrapolability which allows to encode longer sequential position than pre-training samples, it does not always perform well on downstream tasks (Shaw et al., 2018), due to the lack of learnability and flexibility. In practical, most pretrained language models, (e.g. (Devlin et al., 2018; Liu et al., 2019)), utilize learnable vector embedding (Gehring et al., 2017) as positional representation. Recently, several approaches are proposed to enhance positional representation by adding relative position information into attention score computation stage to improve performance of Transformer based models (Shaw et al., 2018; Huang et al., 2020; Dai et al., 2019). By leveraging relative positional encoding and other advanced pre-training techniques, (He et al., 2021), (Chi et al., 2021) achieve state-of-the-art performance on multiple nature language understanding tasks. (Li et al., 2021b) explore the position encoding method in vision domain and propose a learnable Fourier feature to enhance positional encoding in Transformer. It outperforms other methods on both accuracy and convergence speed with vision transformer (Dosovitskiy et al., 2020) based model. Since it is non-trivial to modify or replace backbone of model structure during fine-tuning stage, some research works propose auxiliary tasks (Wang et al., 2019; Pham et al., 2021) or data augmentation approaches (Wei and Zou, 2019; Dai and Adel, 2020) to leverage absolute or relative position information without modifying model structure.

Document Transformer Models In document understanding area, LayoutLM (Xu et al., 2020b) utilizes the pretrained language model to resolve document understanding tasks, and achieves state-of-the-art performance on multiple document understanding benchmarks. To represent 2D position embedding, it decouples the x- and y- axes of text bounding box and sums up positional representations from each dimension independently. LayoutLMv2(Xu et al., 2020a) introduces spatial-aware self-attention mechanism to enhance the layout representation from both 1d and 2d relative position bias. BROS(Hong et al., 2021) uses relative position information in attentional mechanism along with absolute positional encoding from sinusoidal function, which perceives more spatial layout information. (Li et al., 2021a) utilizes shared position information in the text block as position representation which further improves entity extraction performance by understanding cell information from layout. (Appalaraju et al., 2021) proposes an End-to-End Transformer based model with 1D relative position embedding in attentional mechanism.

Document Understanding Tasks RVL-CDIP (Harley et al., 2015) is a document classification dataset with 400K gray-scale English document images in 16 document categories. This dataset is a subset of IIT-CDIP (Lewis et al., 2006) and widely used for pre-training language model purpose. Entity extraction is a classic and essential task in nature language understanding. It is to locate the boundary of entities and assign predefined classes to them. There are several popular benchmarks, consisting of multi-modality information with text, layout, and visual, to evaluate the performance of visually-rich document understanding. FUNSD (Guillaume Jaume, 2019) is a form understanding dataset for key-value extraction research [2] from 199 English forms. SROIE (Huang et al., 2019) and CORD (Park et al., 2019) are receipt understanding datasets to extract related entity types in English. XFUND (Xu et al., 2021) is an extended multi-lingual FUNSD dataset, which contains visually-rich documents in seven commonly-used languages.

[2] More license and term of use information at https://guillaumejaume.github.io/FUNSD/work/

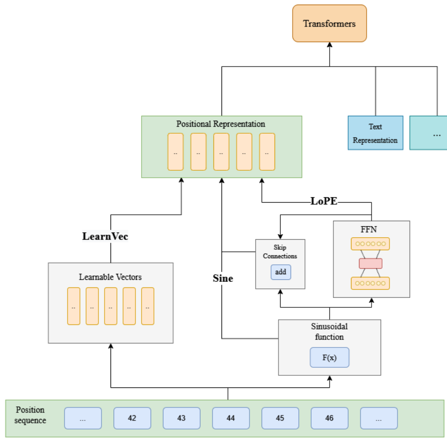


Figure 2: Flow of four positional encoding methods in Transformer based architecture: learnable vector embedding ($LearnVec$), sinusoidal positional encoding ($Sine$), learnable sinusoidal positional encoding ($LoPE$) and $LoPE_{SC}$ with skip connection structure.

3 Methodology

In this section, we formulate our positional encoding method $LoPE$ and introduce its applications on document transformer based language model. In order to evaluate its robustness and stability on noisy data with unreliable order information, we introduce two augmentation strategies: global and local text-block shuffling during fine-tuning stage.

3.1 Learnable Sinusoidal Positional Encoding

Positional representation is utilized as an inductive bias of positional relevance information by positional encoding function (PE) in Transformer model (Vaswani et al., 2017). Sinusoidal positional encoding is originally proposed and employed in attentional mechanism as better extrapolability and spatial correlation from the clean mathematical definition. Figure 1 shows the heatmap of sinusoidal positional encoding method. The hidden representation of position p in a sequence could be computed as Equation 1:

$$\begin{aligned} PE_{sine}(p, 2d) &= sin\frac{p}{10000^{2d/D}} \\ PE_{sine}(p, 2d+1) &= cos\frac{p}{10000^{2d/D}} \end{aligned} \quad (1)$$

In practical applications, some pretrained Transformer language models (Gehring et al., 2017; Devlin et al., 2018; Liu et al., 2019; Xu et al., 2020b; Dosovitskiy et al., 2020) treat each position index p as a discrete learnable embedding vector ($LearnVec$) by learning from pre-training and fine-tuning data. This approach is generic and effective to adapt pretrained Transformer models to specific domains and tasks with various behavior of spatial sensitivity. However, for more challenging tasks, such as document understanding tasks, the performance of document Transfomer model with existing positional encoding approach drops significantly on noisy data with unreliable order information (Hong et al., 2021).

We propose a learnable sinusoidal positional encoding (LoPE) method by combining sinusoidal position encoding function with a fully connected feed-forward network, which consists of two linear transformations with $GeLU$ (Hendrycks and Gimpel, 2020) as activation function σ in between as:

$$\begin{aligned} FFN(x) &= \sigma(xW_1 + b_1)W_2 + b_2 \\ PE_{LoPE}(p) &= FFN(PE_{sine}(p)) \end{aligned} \tag{2}$$

Skip connection is a generic strategy to sum the input and output representation from a computational unit with a skip edge. In transformer based models, (He et al., 2020) has proposed a residual attention layer and shown some regularization effects that could stabilize training and benefit fine-tuning stages. Inspired by this, we conduct the skip connection strategy in $LoPE$ module as a variant of our method. It could be formulated as eq.3.

$$PE_{LoPEsc}(p) = PE_{sine}(p) + PE_{LoPE}(p) \tag{3}$$

Figure 2 visualizes the flow of our proposed method and baselines in this paper. Compared with discrete embedding, our method extends from sinusoidal function and treats position index as a continuous-valued vector which allows the model to extrapolate to longer length from training cases. Meanwhile, the learnable FFN component boosts the learnability and flexibility of positional representation for multidimensional spatial information.

3.2 Positional Representation in Document Transformer Language Model

Distinct from nature language data which only consist of 1D order information, visually-rich document data require more model capacity to represent both 1D and 2D positional information from individual element. Given token x_i series from a document D, let p_i donate 1D position index and b_i as $((x_0, y_0), (x_1, y_1))$ present the bounding box in normalized 2D coordinate system.

As a general and commonly used pre-trained model for Document AI, LayoutLM (Xu et al., 2020b) utilizes independent 2D spatial embedding layers along with 1D position embedding initialized from pretrained BERT (Devlin et al., 2018) to represent positional information. Its composed positional representation R_i is computed via:

$$\begin{aligned} \mathcal{R}_i^{2D} &= \sum_{j=0}^{k}(PE_x(x_j) + PE_y(y_j)) \\ \mathcal{R}_i &= PE_{1d}(p_i) + \mathcal{R}_i^{2D} \end{aligned} \tag{4}$$

Where k donates the count of points in bounding box, and PE_{1d}, PE_x, PE_y are the positional encoding methods for 1D order and 2D spatial information separately. The original positional encoding of LayoutLM is a learnable embedding which is identical to $PE_{LearnVec}$3.1 in this paper. The composed positional representation will be summed up with text embedding and token type embedding vectors as input of Transformer.

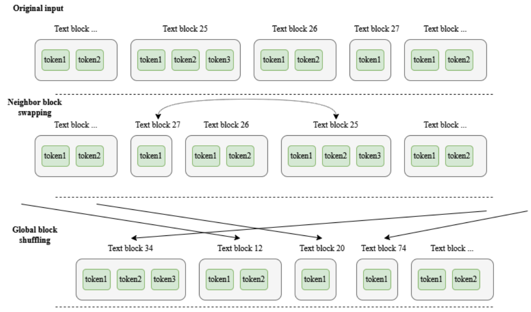


Figure 3: An example of text block shuffling augmentation methods, Neighbor Block Swapping and Global Block Shuffling.

3.3 Text Block Shuffling Augmentations

In practical, understanding documents with incorrect reading order is a real challenge for document Transformer model which always leads to irrelevant or embarrassing error results. We introduce two text block shuffling augmentation methods: **Global Block Shuffling** and **Neighbor Block Swapping**, to simulate the noisy reading order scenario as shown in Figure 3. We apply these methods on text block level to a document, and keep the relative word order in the same text block. The text block is defined as a group of continual words in a spatial region (or a line of words).

For global block shuffling process, we first obtain the block information for each token, and shuffle the order of block index but keep the relative token order in the input sequence. For neighbor block swapping method, each text block is swapped

Parameter Name	Value
max_steps	500K
per_device_train_batch_size	12
gradient_accumulation_steps	4
max_seq_length	512
max_2d_position_embeddings	1024
learning_rate	7e-5
warmup_ratio	0.1

Table 1: Pretraining hyperpameters for document Transformer model with our positional encoding methods.

to neighbor block randomly, and the distance d of swapped block pair follows a normal distribution function $\mathcal{N}(0, \sigma^2)$.

The intuition of applying augmentation method on text block level is that we observe it is closed to error cases from document understanding application in real word, and the text block information could be obtained from existing OCR engines.

4 Experiments

4.1 Pretraining

In order to verify the effectiveness of our positional encoding approach, we employ LayoutLM frame and exclude the visual feature related structure. We reproduce the pretraining experiments with our positional encoding method as well as baseline methods on a 1M random subset of IIT-CDIP (Lewis et al., 2006) pretraining data set.

All pretraining jobs run on 8 NVIDIA Tesla V100 32GB GPUs server with approximately 150 hours for each job. The pretraining hyperparameters are shown in Table 1. The pretrain models are initialized from Bert-base-uncased except for specified positional encoding weights.

We obtain our pretrained models with four positional encoding methods ($LearnVec$, $Sine$, $LoPEsc$, $LoPE$) for next fine-tuning experiments. The name of positional encoding method is used to indicate the pretrained model in the result table.

4.2 Experimental Settings

We fine-tune and evaluate the performance of our pretrained models on two datasets: FUNSD (Guillaume Jaume, 2019) and SROIE (Huang et al., 2019), which are two popular benchmark datasets for entity extraction in form and receipt domains.

FUNSD [3] consists of noisy scanned documents. There are 149 scanned forms for training and 50 scanned forms for testing with more than 31K words, 9.7K entities, and 5.3K relations in combination. For more fair comparison, we refer the evaluation results from LayoutLM, DocFormer, and BROS with the same text and spatial features as input and similar model size architecture.

SROIE [4] attracts a lot of attention from both research and industry community as an open-source OCR and information extraction benchmark for receipt understanding. The dataset consists of 626 receipt images for training and 347 receipt images for testing with four predefined entities which are $company$, $date$, $address$, and $total$. There is no post-processing strategy before evaluation as we tend to compare the performance gap only from positional encoding differences. We also experiment with official pretrained LayoutLM[5] with the same fine-tuning hyper-parameters for a fair comparison purpose.

We use entity recognition evaluation metrics including entity-level precision, recall, and F1-score for each experiment by default settings of seqeval package (Nakayama, 2018). The learning rate is set to 3e-5 with linear decay, and 10% of total steps are used for warm-up purpose. We use max_steps as $2k$, and report the evaluation metrics on the final fine-tuned model. Other environment settings or hyper-parameters are same as pretraining experiments 4.1. We average evaluation results with different initial seeds to eliminate bias of shuffling augmentations.

4.3 Experimental Results

As shown in Table 2, on FUNSD dataset, our $LoPE$ model achieves 82.04 F1-score and outperforms other baseline methods. The $Sine$ model achieves low performance and $LoPE_{SC}$ is worse than $LoPE$ which indicates the sinusoidal function cannot represent layout positional information with skip connection structure. The small performance gap between our $LearnVec$ and official LayoutLM model with shared model structure might be from different pretraining data and settings since our pretraining experiments run on a 1M subset training data and fewer pretraining steps.

We observe similar trend on SROIE experiment from Table 3. $LoPE$ model achieves F1 score of 93.87 with text and spatial features. With larger scale of training size on SROIE, the performance gap is narrowed down between $LearnVec$ and $LoPE$ in testing data set.

[3] https://guillaumejaume.github.io/FUNSD

[4] https://github.com/zzzDavid/ICDAR-2019-SROIE

[5] https://github.com/microsoft/unilm/tree/master/layoutlm

These results illustrate the effectiveness of our $LoPE$ on document understanding tasks with different data scale. The ability of positional representation affects the final performance significantly on document understanding models.

Method	P(%)	R(%)	F1(%)
$LayoutLM$(2020b)	75.97	81.55	78.66
$DocFormer$(2021)	77.63	83.69	80.54
$BROS$(2021)	80.56	81.88	81.21
$LearnVec$	75.97	80.04	77.95
$Sine$	72.8	77.24	74.95
$LoPE_{SC}$	78.25	82.79	80.46
$LoPE$	80.4	83.74	**82.04**

Table 2: Entity level evaluation results on FUNSD dataset. All models utilize input features of text and spatial information with "Base" model size architecture.

Method	P(%)	R(%)	F1(%)
$LayoutLM_{base}$	91.98	94.16	93.06
$LearnVec$	92.57	94.31	93.43
$Sine$	87.72	90.06	88.87
$LoPE_{SC}$	89.89	92.87	91.35
$LoPE$	92.94	94.81	**93.87**

Table 3: Results on SROIE datasets. All above experiments are fine-tuned with same hyper-parameter setting. We evaluate the performance on official $LayoutLM_{base}$ model for reference.

4.4 Ablation Study

In real-world application, the reading order of text blocks is not always reliable and consistent. The incorrect reading order harms the performance of existing document language models and leads to embarrassing error of predictions in downstream tasks. We conduct three ablation experiments to simulate the impact of such error with the above augmentation methods 3.3.

Neighbor Block Swapping and Global Block Shuffling We apply these methods to training data only during fine-tuning which simulates impact of incorrect block order data. The testing set is kept as original which allows us to compare the performance with 2 fairly. The σ of neighbor block swapping is set to 1 in all experiments. Note that the augmentation methods in this paper require block information of each token, and that might cause leaking of block boundary information during the model training indirectly. Besides of data impact, the model receives inconsistent reading order during training and it might benefit the evaluation performance by eliminating the over-fitting from 1D positional embedding, and tent to learn more information of relative token order inside block and 2D spatial information.

In Table 4, with these noisy data by adding these two augmentation methods, our $LoPE$ methods show better performance than existing discrete $LearnVec$ embedding or sinusoidal function $Sine$ consistently on FUNSD data. The global block shuffling is harmful for all pretrained models while the performance impact of neighbor block swapping is marginally. The discrete positional encoding method shows more sensitive with significant performance drop by global block shuffling augmentation.

Removing 1D Position Input We throw the 1D positional embedding and only consider the 2D positional representation $\mathcal{R}^{2D}$ in eq. 4 in composed positional representation for both training and testing data sets. The model does not receive word order information on both text block and sub-token level. We refer the performance result from BROS (Hong et al., 2021) with similar setting for comparison.[6]

On FUNSD dataset, we observe a significant performance degradation across all positional methods in Table 5. The $LearnVec$ leads a huge drop from approximately 79% to 49% F1 score which indicates the discrete 2D embedding is not well represented without optimal order information. The continuous 2D positional encoding methods perform better relatively. $LoPE_{SC}$ performs best with 2.67% F1 drops in absolute from Table 2, and keeps a reasonable mode even with none order information.

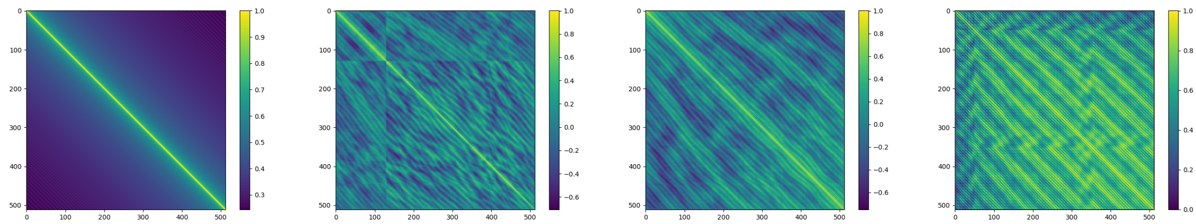
From Table 6, we observe our $LoPE$ model achieves 89.98 F1 score with 3.89% absolute drop (4.14% relatively) from Table 3. The performance of $LoPE_{SC}$ drops 3.2% relatively which shows better robustness on such extreme condition. There is significant performance regression with discrete $LearnVec$ method on this receipt understanding data set. The $LoPE_{SC}$ performs better with global block shuffling method on FUNSD data set which might be beneficial from regularization advantage of skip connection structure.

Ablation study results further prove that better learnability and spatial correlation of positional rep-

[6] Result from text line in their ablation study paragraph

Method	Neighbor Block Swapping			Global Block Shuffling		
	P(%)	**R**(%)	**F1**(%)	**P** (%)	**R**(%)	**F1** (%)
$LearnVec$	76.43	79.49	77.93	72.32	69.78	71.03
$Sine$	73.77	78.24	75.94	74.1	74.99	74.54
$LoPE_{SC}$	78.72	81.79	80.23	77.09	80.14	**78.59**
$LoPE$	79.9	82.14	**81.01**	78.03	78.34	78.18

Table 4: Comparison on **FUNSD** dataset for four positional encoding methods by applying **Neighbor** Block Swapping and **Global** Block Shuffling on training data set, evaluation results clearly shows our methods perform stable and robustness with unreliable order information.



(a) $Sine$ (b) official $BERT$ (c) $LearnVec$ (d) $LoPE$

Figure 4: Similarity of 1D position embedding from our pretrained $Sine$, official BERT, $LearnVec$, $LoPE$ models.

resentation are essential factors to improve existing document Transformer model. By comparing with baseline positional encoding methods and other recent pretrained Transformer based solutions, our methods demonstrate optimal performance and robustness on noisy data with unreliable order information.

Method	P(%)	R(%)	F1(%)
$BROS$(2021)	–	–	70.07
$LearnVec$	44.66	54.63	49.14
$Sine$	69.4	73.74	71.5
$LoPE_{SC}$	75.71	79.99	**77.79**
$LoPE$	72.2	77.19	74.61

Table 5: Experimental results by removing 1D position inputs on training and testing sets of FUNSD. The BROS performance is referenced from their ablation study with similar experimental setting.

Method	P(%)	R(%)	F1(%)
$LearnVec$	75.12	79.18	77.1
$Sine$	83.71	87.03	85.34
$LoPE_{SC}$	87.46	89.41	88.42
$LoPE$	87.9	92.15	**89.98**

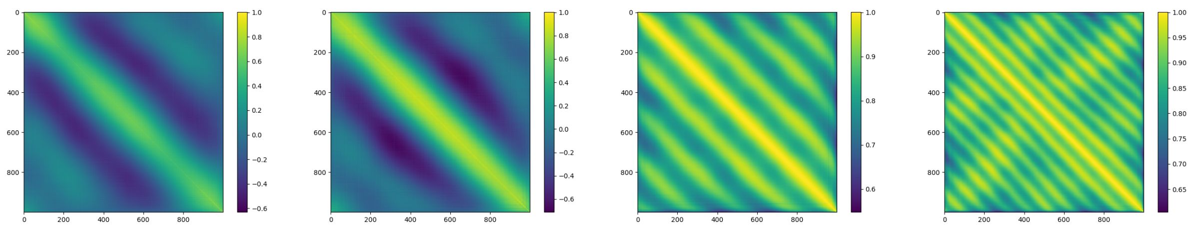
Table 6: Experimental results by removing 1D position inputs on training and testing sets of SROIE. The $LoPE$ achieves best performance and $LoPE_{SC}$ keeps lowest relative performance drop with this extra settings.

5 Position Embedding Similarity Analysis

In this section, we visualize the similarity of positional representation from our pretrained models, official BERT, and LayoutLM as reference. We obtain our pretrained models with different positional encoding methods from 4.1. The positional representation could be computed from the specific position embedding layer and a range of position inputs. We use $Cosine\ similarity$ to measure the similarity between two positional representations.

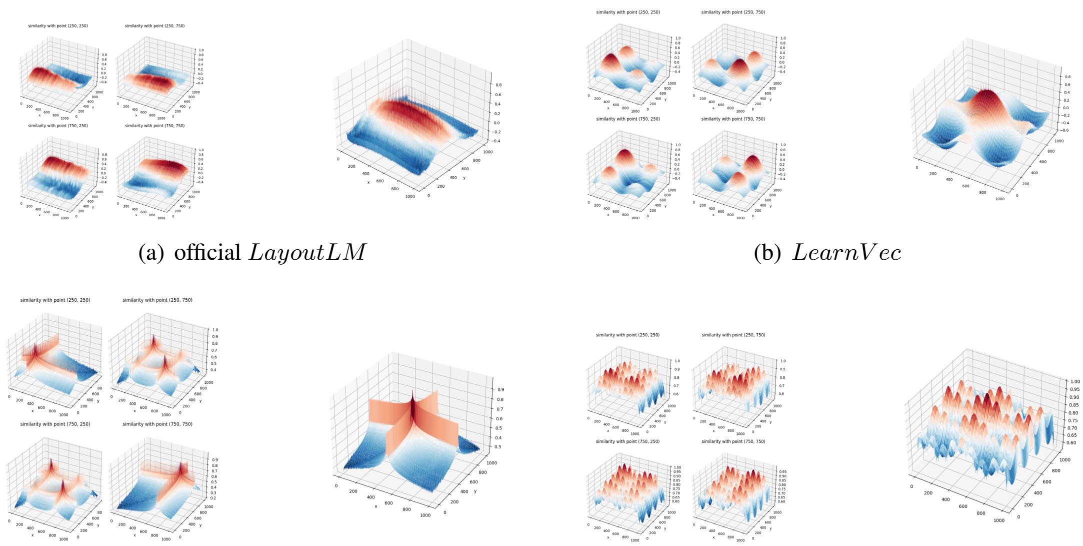
In Figure 4, we obtain the 1D positional representation from our pretrained model with $Sine$, $LearnVec$, and $LoPE$ methods in range 0 to 512. The position embedding of official BERT model is also computed as reference. The points which are closer to diagonal tend to have higher similarity on each positional encoding method. Meanwhile, with learnable structures, the similarity heatmap shows different texture patterns which might be learned from pretraining data. The length of text input from document data set is usually longer than samples from NLP data which might lead different attention distribution on 1D position embedding. Our $LoPE$ method shows clear volatility from heatmap of 1D positional similarity.

Figure 5 shows similarity heatmap of x- and y-axes 2D positional embedding from our pretrained $LearnVec$ and $LoPE$ models. Figure 6 demonstrates the similarity of $\mathcal{R}^{2D}$ representation from



(a) PE_x of $LearnVec$ (b) PE_y of $LearnVec$ (c) PE_x of $LoPE$ (d) PE_y of $LoPE$

Figure 5: Similarity of x and y axes in 2D positional embedding from our pretrained $LearnVec$ and $LoPE$ models.



(a) official $LayoutLM$ (b) $LearnVec$

(c) $Sine$ (d) $LoPE$

Figure 6: Similarity of 2D positional representation on 5 fixed points ((250, 250), (250, 750), (750, 250), (750, 750), (500, 500)) to rest position from official LayoutLM, $LearnVec$, $Sine$, $LoPE$ based pretrained model.

five specific points to rest position from our pretrained models and official LayoutLM model. We observe slightly different distribution of heatmap between our pretrained $LearnVec$ and official LayoutLM model, which might arise from distinct pretraining dataset and settings. The official LayoutLM model shows boarder vision horizontally with proper spatial correlation. The similarity of $Sine$ is decaying rapidly from central point and shows sharp edge on the 2D heatmap. Our $LoPE$ shows higher wave frequency on both x- and y-axes which tend to capture the long distance signals with speckled pattern.

6 Conclusions

In this paper, we propose a new and generic learnable positional encoding method $LoPE$ to improve the positional representation in Transformer based model. By combining sinusoidal positional function and learnable feed-forward network, our method takes advantages of better learnability and extrapolability. Experimental results on both FUNSD and SROIE data sets clearly illustrate the effectiveness of our proposed method on document understanding tasks. By leveraging global and local shuffling augmentation methods or removing order information from inputs, we demonstrate our methods substantially outperform other positional encoding methods on noisy data with unreliable order information.

The conclusion of this paper is made from limited tasks, datasets and linguistic terms which might be bias from the task definition, annotation guidance or imbalanced data distribution. Meanwhile, it is unclear if our method is effective on other domain, modality, and area. For future research, we will evaluate our method on other tasks and transfer to other area such as image related tasks with Vision Transformer (Dosovitskiy et al., 2020) architecture.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. *arXiv preprint arXiv:2106.11539*.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021. Xlm-e: Cross-lingual language model pre-training via electra.

Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. 2013. The significance of reading order in document recognition and its evaluation. In *2013 12th International Conference on Document Analysis and Recognition*, pages 688–692. IEEE.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition.

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

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*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Ruining He, Anirudh Ravula, Bhargav Kanagal, and Joshua Ainslie. 2020. Realformer: Transformer likes residual attention. *arXiv preprint arXiv:2012.11747*.

Dan Hendrycks and Kevin Gimpel. 2020. Gaussian error linear units (gelus).

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2021. {BROS}: A pre-trained language model for understanding texts in document.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. 2020. Improve transformer models with better relative position embeddings.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. Structurallm: Structural pre-training for form understanding. *arXiv preprint arXiv:2105.11210*.

Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. 2021b. Learnable fourier features for multi-dimensional spatial positional encoding.

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*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: A consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Thang M. Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks?

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. Layoutreader: Pre-training of text and layout for reading order detection.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding.