DocFormer Supplemental Paper

Srikar Appalaraju
AWS AI
srikara@amazon.com

Bhavan Jasani
AWS AI
bjasani@amazon.com

Bhargava Urala Kota
AWS AI
bharkota@amazon.com

Yusheng Xie
AWS AI
yushx@amazon.com

R. Manmatha
AWS AI
manmatha@amazon.com

1. Supplemental

This is the supplemental material for the main DocFormer paper [1]. Please read the main paper for model formulation, performance numbers on various datasets and further analysis and ablation.

1.1. Implementation Details

We present all the hyper-parameters in Table 1 used for pre-training and fine-tuning DocFormer . We fine-tune on downstream tasks on the same number of epochs as prior art [9, 10, 5]: FUNSD [3], Kleister-NDA [2] datasets were fine-tuned for 100 epochs. CORD [7] for 200 epochs. RVL-CDIP [4] for 30 epochs. For Key, Query 1-D relative local attention we choose a span of 8 i.e. for a particular multi-modal feature, DocFormer gives more attention 8 tokens to its left and right.

Hyper-Parameter	Pre-training	Fine-tuning
Epochs	5	varies
Learning rate	5E-05	2.5E-05
Warm-up	10% iters	0
Gradient Clipping	1.0	1.0
Gradient agg.	False	False
Optimizer	AdamW[6]	AdamW[6]
Lower case	True	True
Sequence length	512	512
Encoder layers	12	12
32-bit mixed precision	True	True
Batch size	9 per GPU	4 per GPU
GPU hardware	A100 (40GB)	V100 (16GB)
Training Num. Samples	5M	varies
Training time	17 hours/epoch	varies

Table 1: **Implementation Details**: Hyper-parameters used for pre-training DocFormer and fine-tuning for downstream tasks. Training epochs vary for down-stream tasks.

1.2. Run-time Complexity Analysis

Since we propose a variant of the self-attention [8] operation, we compute the train and inference run-time analysis in big-o notation.

Layer Type	Run-time Complexity	Seq. Complexity
Convolution	$O\left(k \cdot n \cdot d^2\right)$	$O(1)$
Recurrent	$O\left(n \cdot d^2\right)$	$O(n)$
Self-Attention	$O\left(n^2 \cdot d\right)$	$O(1)$
Self-Attention (relative)	$O\left(r \cdot n \cdot d\right)$	$O(1)$
DocFormer MMSA	$2 \cdot \left[O\left(n^2 \cdot d\right)\right]$	$O(1)$

Table 2: **Complexity analysis**: Here n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention. Omitting number of attention heads h for brevity. Here assume $h = 1$. In addition, MMSA: multi-modal self-attention.

Please note that the full run-time complexity for DocFormer has been abridged as the self-attention is the most significant operation (keeping in line with big-O notation). In addition, the presence of 2 is to signify the unique MMSA operation proposed in this paper, where multi-modal feature from each layer is added with image and spatial features (see Section 3.1). We see that DocFormer's multi-modal self-attention (Section 3.1) is an efficient way to do multi-modal learning.

1.3. Pseudo-code

We present a rough pseudo-code for our novel multi-modal self-attention (MMSA) as described in section 3.1. We believe the pseudo-code would aid an independent researcher to better replicate our proposed novelty. Please note omitting dropout and layer norm at the end for brevity.

```
##Multi Modal Self Attention

#text kqv embeddings
key1 = Linear(d_model, n_head * d_k)
query1 = Linear(d_model, n_head * d_k)
value1 = Linear(d_model, n_head * d_v)

#image kqv embeddings
key2 = Linear(d_model, n_head * d_k)
query2 = Linear(d_model, n_head * d_k)
value2 = Linear(d_model, n_head * d_v)

```

```
#spatial embeddings. note! shared by text, image
key3 = Linear(d_model, n_head * d_k)
query3 = Linear(d_model, n_head * d_k)

#See Eq. 6 and 7 in main paper for formulation
def multi_modal_self_attention(emb, img_feat, spatial_feat):

  #self-attention of text (and prev. layers subseq.)
  k1,q1,v1 = emb,emb,emb
  k1 = rearr(key1(k1), 'b t (head k) -> head b t k')
  q1 = rearr(query1(q1), 'b l (head k) -> head b l k')
  v1 = rearr(value1(v1), 'b t (head v) -> head b t v')
  attn1 = einsum('hblk,hbtk->hblt', [q1,k1])/sqrt(q1.shape[-1])

  #1D relative pos. (query, key)
  #note rel_pos_embed1 is learnt relative pos emb. nxn
  rel_pos_key1 = einsum('bhrd,lrd->bhlr', k1, rel_pos_embed1)
  rel_pos_query1 = einsum('bhld,lrd->bhlr', q1, rel_pos_embed1)

  #shared spatial - text/hidden features
  sp_k1, sp_q1 = spatial_feat, spatial_feat
  sp_k1=rearr(key3(sp_k1),'b t (head k) -> head b t k')
  sp_q1=rearr(query3(sp_q1),'b l (head k)->head b l k')
  text_only_spatial_scores = einsum('hblk,hbtk->hblt', [sp_q1,sp_k1])/sqrt(sp_q1.shape[-1])

  text_attn_scores = attn1 + rel_pos_key1 + rel_pos_query1 + text_only_spatial_scores

  #-----
  ##Self-attn of image (repeat of above for img feat)
  k2,q2,v2 = img_feat,img_feat,img_feat
  k2 = rearr(key2(k2), 'b t (head k) -> head b t k')
  q2 = rearr(query2(q2), 'b l (head k) -> head b l k')
  v2 = rearr(value2(v2), 'b t (head v) -> head b t v')
  attn2 = einsum('hblk,hbtk->hblt', [q2,k2])/sqrt(q2.shape[-1])

  #1D relative pos. (query, key)
  #note rel_pos_embed1 is learnt relative pos emb. nxn
  rel_pos_key2 = einsum('bhrd,lrd->bhlr', k2, rel_pos_embed2)
  rel_pos_query2 = einsum('bhld,lrd->bhlr', q2, rel_pos_embed2)

  #shared spatial - img features
  sp_k2, sp_q2 = spatial_feat, spatial_feat
  sp_k2=rearr(key3(sp_k2),'b t (head k) -> head b t k')
  sp_q2=rearr(query3(sp_q2),'b l (head k)->head b l k')
  img_only_spatial_scores = einsum('hblk,hbtk->hblt', [sp_q2,sp_k2])/sqrt(sp_q2.shape[-1])

  img_attn_scores = attn2 + rel_pos_key2 + rel_pos_query2 + img_only_spatial_scores

  #---- attended output: multi-modal
  text_attn_probs = dropout(softmax(dim=-1)(text_attn_scores))
  img_attn_probs = dropout(softmax(dim=-1)(img_attn_scores))

  text_cntx = einsum('hblt,hbtv->hblv', [text_attn_probs, v1])
  img_cntx = einsum('hblt,hbtv->hblv', [img_attn_probs, v2])
  context = text_cntx + img_cntx
  return context
```

1.4. DocFormer Architecture for Downstream Tasks

DocFormer is pre-trained as mentioned in section 3.2. After training it for 5 epochs, we remove the pre-training multi-task heads and use DocFormer (including the visual branch) as a backbone. We simply add a trainable linear-head which predicts the appropriate number of classes which is dataset specific. Please see Figure 1 for architecture modifications for downstream tasks.

1.5. DocFormer Multi-Modal Self-Attention

In Figure 2 we show a more detailed visual representation of the novel multi-modal self-attention introduced in this paper. For reference we also show the original self-attention used by Vaswani et al. [8].

1.6. FUNSD Vizualizations

DocFormer achieves state-of-the-art performance of 83.34% F1-score (see Section 4.1) on FUNSD [3] dataset amongst other multi-modal models its size. In this sub-section we look at more visualizations by DocFormer on the test-set. One important aspect of this VDU we would like to mention is the OCR is not in human reading-order.

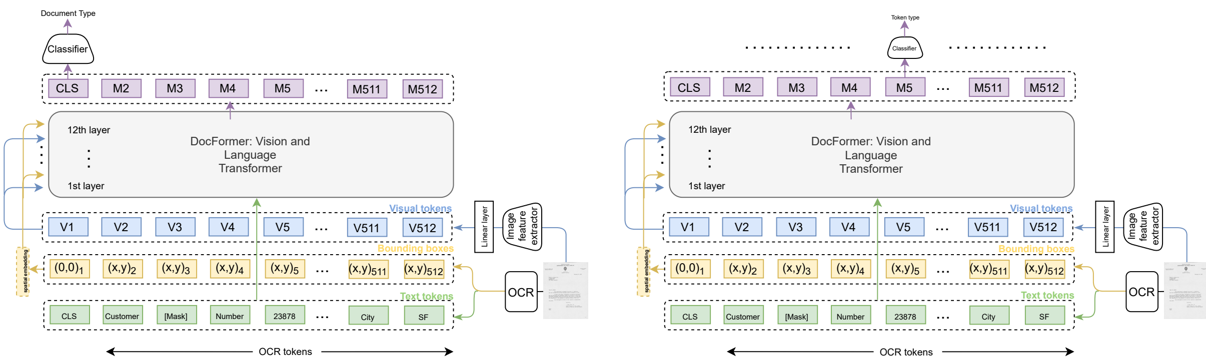
Please note that, we search for and present cases where mistakes were made by DocFormer with the aim of understanding mistakes. Legend for the colors used in images is, Header-label: **Red**, Question: **Blue**, Answer: **Green**, Other: Grey color. Please see Figures 3, 4, 5.

In Figure 6, we show one specific pattern that DocFormer learns through its novel multi-modal self-attention. We show that DocFormer automatically learns repetitive local patterns even though it was not explicitly taught this.

1.7. CORD Visualizations

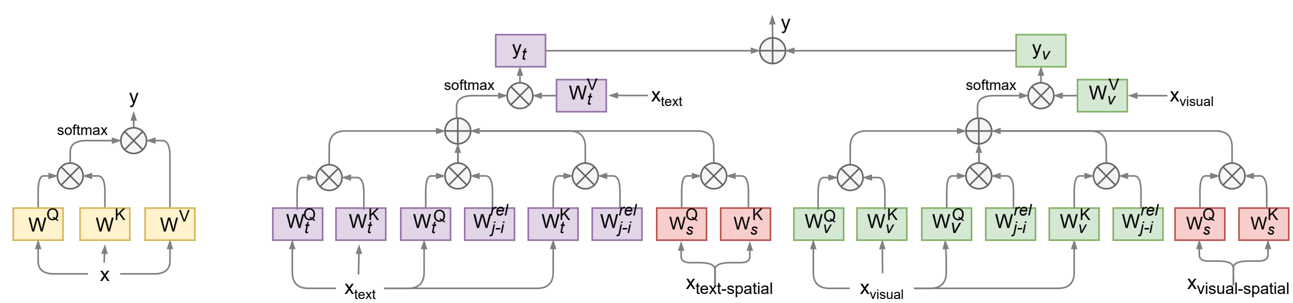
DocFormer matches the state-of-the-art performance of 96.33% F1-score on CORD [7] dataset (previous state-of-the-art model TILT-large consists of 780M parameters almost 4x the size of DocFormer). Please see Section 4.3 in the main paper.

In this sub-section we look at CORD [7] visualizations by DocFormer . We explicitly show hard-cases where DocFormer does well, see Figures 7, 8, 9, 10, 11. In order to be transparent, we also show an error scenario in Figure 12. Legend for the colors in images is, Menu items: **Red**, Total: **Blue**, Sub-total (pre-tax): **Green**, Void-menu: Cyan color, Other: grey.



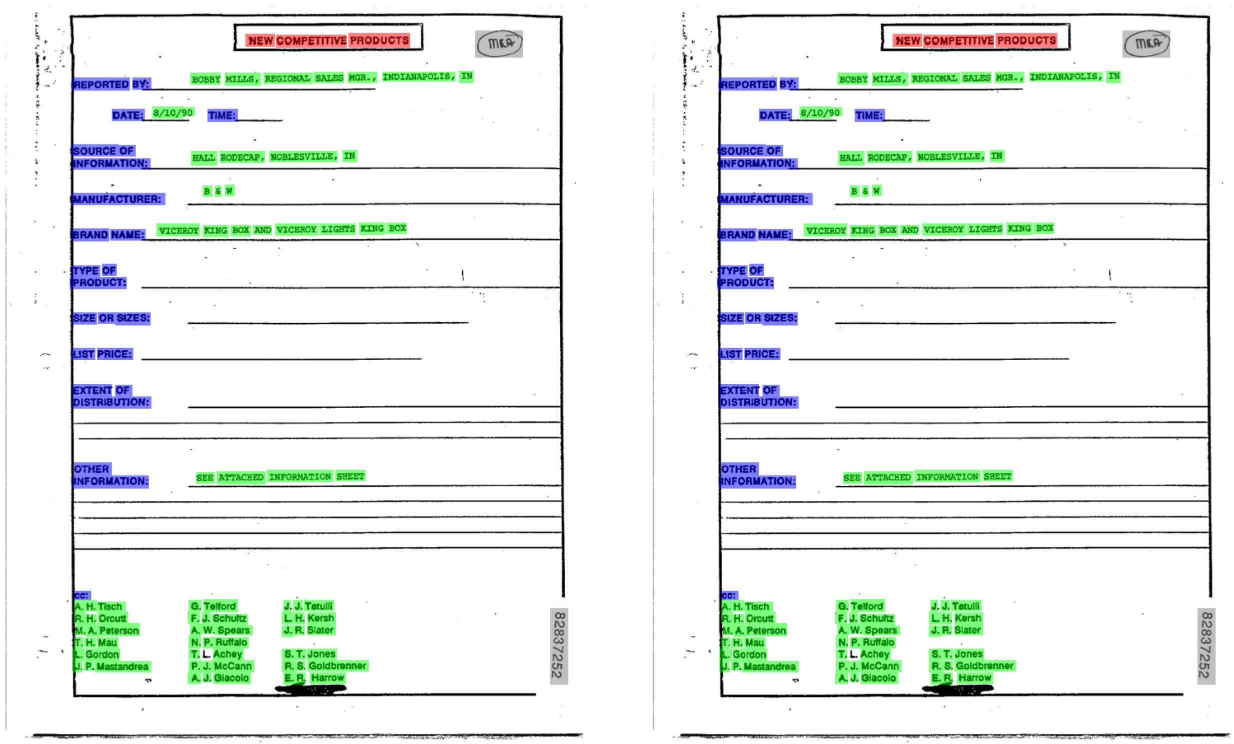
(a) Architecture for downstream Document Classification Task (b) Architecture for downstream Sequence and Entity Labeling Tasks

Figure 1: **DocFormer architecture for various downstream tasks**: Image on **Left** (a) is the architecture for document classification [CLS] is a pooling layer (fn → ReLU → fn) to get a pooled representation used for document classification task. Image on **Right** (b) is the architecture used for entity and sequence labeling tasks. Note, only a single linear layer is added for all downstream tasks. Also, all components of DocFormer are fine-tuned for each of the downstream tasks.



a) orignal self-attention b) multi-modal spatial relative self-attention in each layer of DocFormer

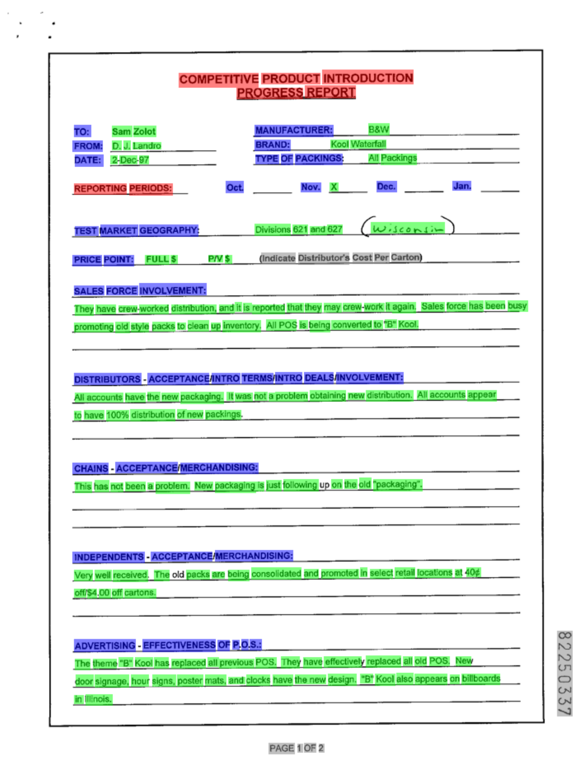
Figure 2: **Multi-Modal Self-Attention Layer**: the image **a) Left** shows the traditional self-attention proposed in Vaswani et al [8]. Note the multi-head attention and feed-forward layers are omitted for brevity. Cross (X) is matrix-multiplication and (+) is element-wise addition. **b) Right** shows the proposed multi-modal self-attention layer. This comprises each layer of DocFormer . Notice, the spatial weights across text, vision are shared (RED color), thus helping DocFormer address the *cross-modality feature correlation* issue commonly faced in multi-modal training. The notation is consistent with Equations 1-7 in the main paper. Best viewed in color.



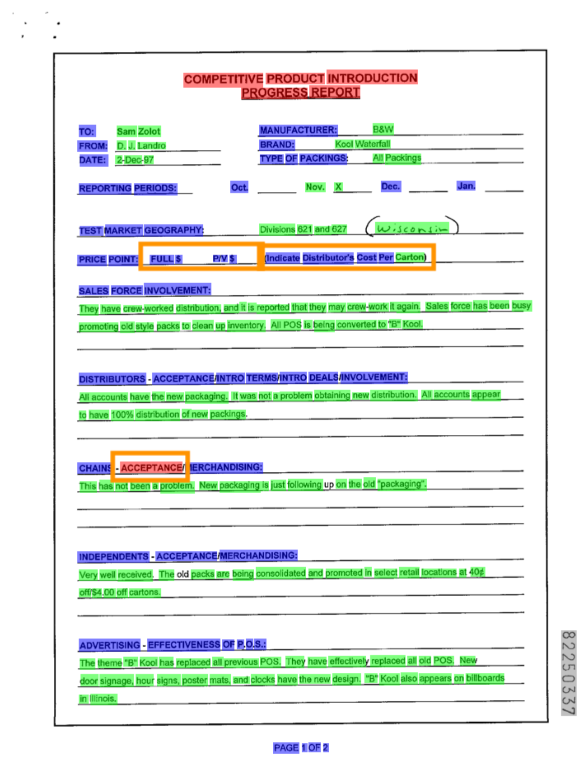
(a) Ground Truth

(b) DocFormer predictions

Figure 3: **DocFormer perfect predictions for 82837252 testfile of FUNSD dataset**: Left image shows GT and right image is the prediction made by DocFormer which perfectly matches with GT. Best viewed in color.

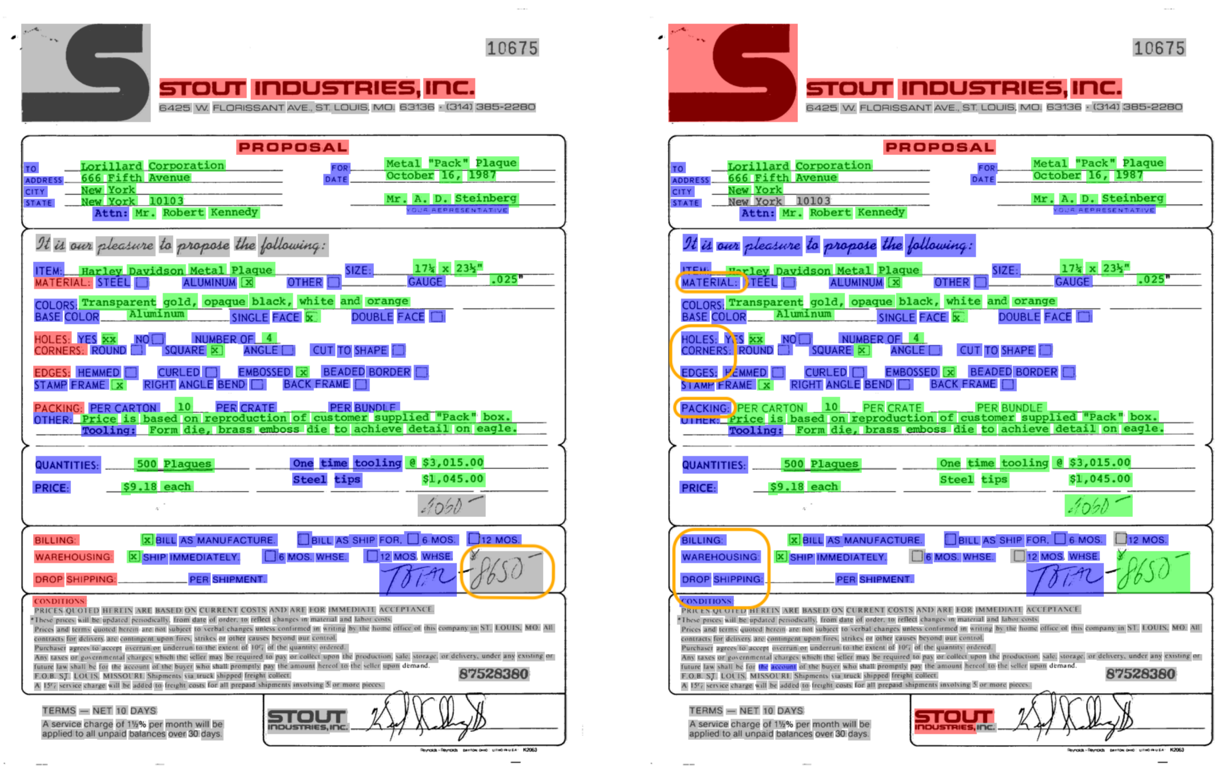


(a) Ground Truth



(b) DocFormer predictions

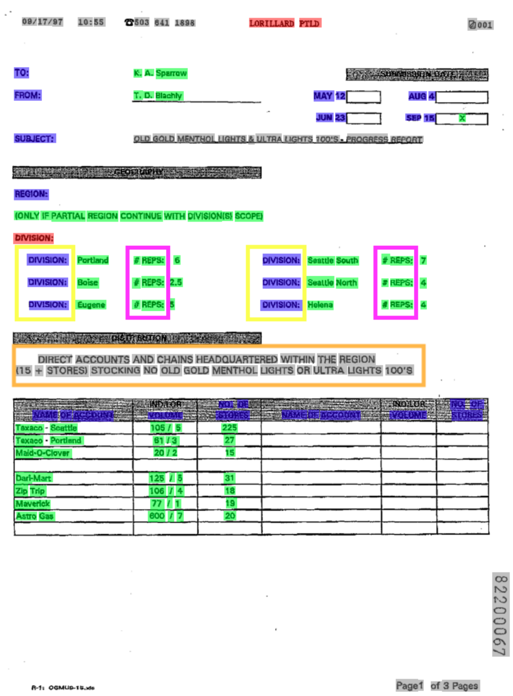
Figure 4: **DocFormer slightly bad predictions for 82250337_0338 testfile on FUNSD dataset**: Based on the predictions on the right (b), we can see that DocFormer was able to classify most of the sequence correctly. However, if we look at the orange bounding boxes we can spot the errors. "(Indicate Distributor's Cost per Carton)" is tagged as Other text in ground-truth but DocFormer incorrectly classified part of the tokens as **Question**. Best if viewed digitally and in color.



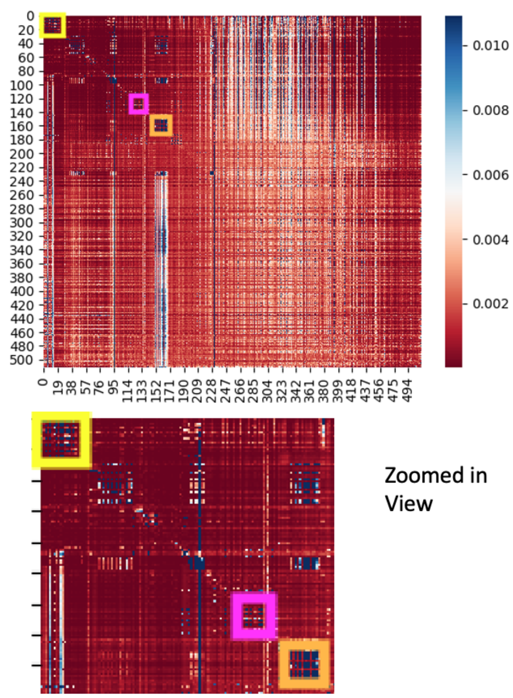
(a) Ground Truth

(b) DocFormer predictions

Figure 5: **DocFormer slightly bad predictions for 87528380 testfile on FUNSD dataset**: Here, we focus the readers attention on two specific scenarios: FUNSD dataset has been known to have ground-truth annotation issues. We find on the left image the orange highlighted box "8650" is incorrectly annotated in GT as "$other$" text, however DocFormer correctly predicts it as "answer" token for the question "$total$". **Scenario 2:** The orange highlighted boxes on the right image are tokens which are actually sub-headers but DocFormer mis-classifies as "$question$" tokens. In this case, DocFormer likely gave more weight-age to language features and not so much to visual features and so ended up mis-classifying. We would like to point out that this is an ambiguous example as the language in mis-classified regions do look like "$questions$". Best viewed in color.



(a) Example FUNSD document with Ground Truth overlays

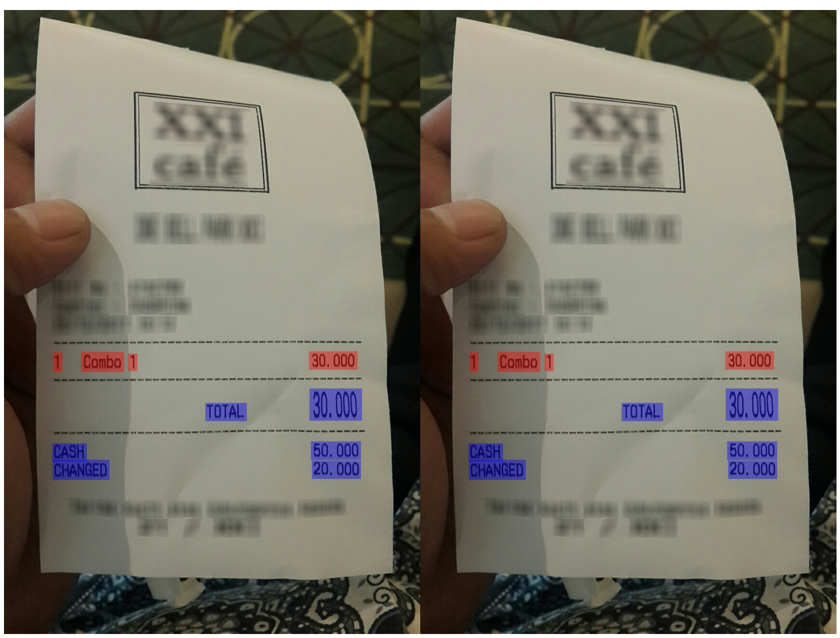


(b) DocFormer prediction self-attention heatmap (last encoder layer, 2nd head). DocFormer has up to 512 tokens in each layer. Each point on the image shows the strength of attention from a token on the y-axis to a token on the x-axis. The bluer colors show more attention and the red less attention.

Figure 6: **DocFormer learns repetition and regularity**: the yellow and **purple** boxes in the left figure matches the yellow and **purple** boxes in the right figure. The OCR is not in reading order. Hence the six occurrences of "DIVISION" appear together in front (among the top 25) - yellow box in Figure b) and they correspond to the yellow boxes in Figure a). Similarly, the **purple** box in Figure b) corresponds to the **purple** boxes in Figure a). DocFormer is able to pick up such repetitions as strong self-attention signals (blue colored pixels in the right self-attention figure) that help the model solve the task. This example shows that regular indentation and spacing help DocFormer understand the form better just as they would help humans parse a form. The **orange** boxed region in the heatmap also shows strong self-attention. We think that is due to DocFormer representing the blob of text as a single paragraph (in this case, as background text). Best viewed digitally and in color.

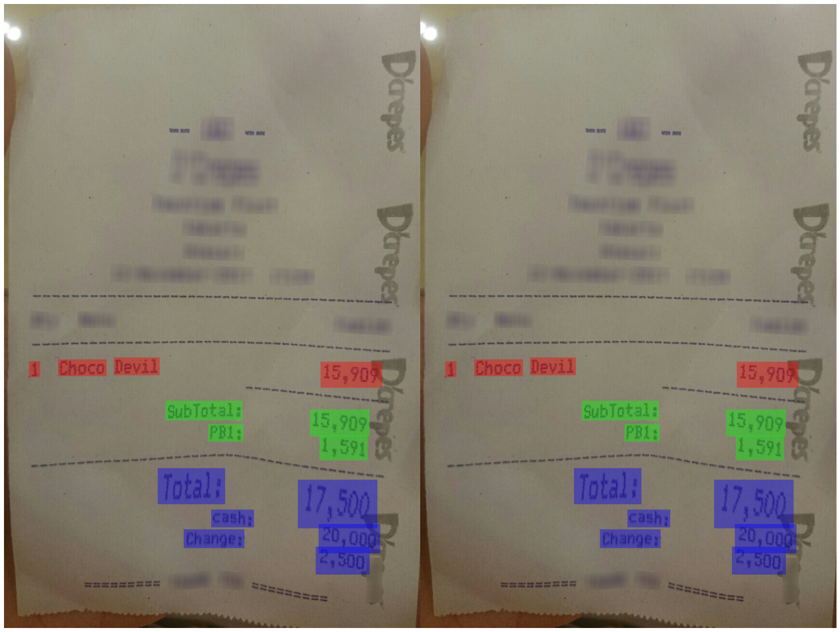
(a) Ground Truth (left) and DocFormer predictions (right)

Figure 7: **DocFormer predictions on CORD**: For file receipt_00053 (a) shows both ground-truth and predictions. DocFormer predicted correctly all the entity regions in the image. Best if viewed digitally and in color.



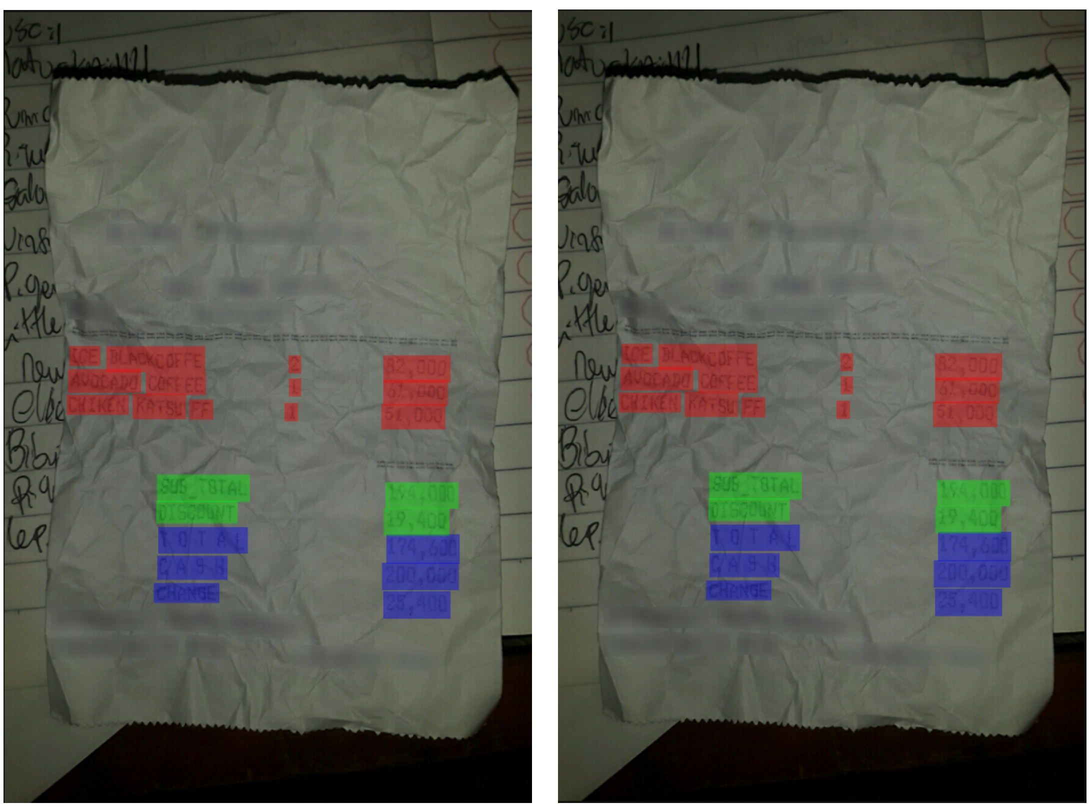
(a) Ground Truth (left) and DocFormer predictions (right)

Figure 8: **DocFormer predictions on CORD**: For file receipt_00044. Best if viewed digitally and in color.



(a) Ground Truth (left) and DocFormer predictions (right)

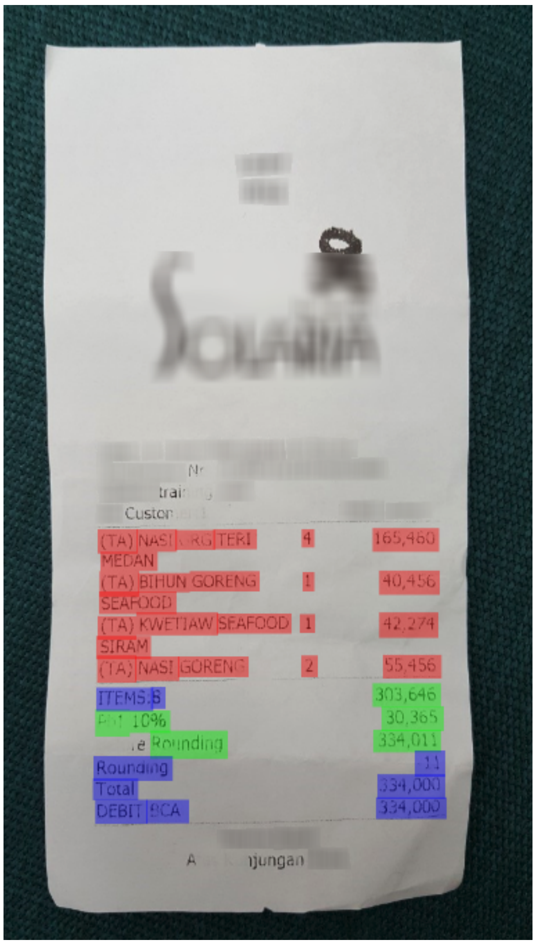
Figure 9: **DocFormer predictions on CORD**: For file receipt_00072. Best if viewed digitally and in color.



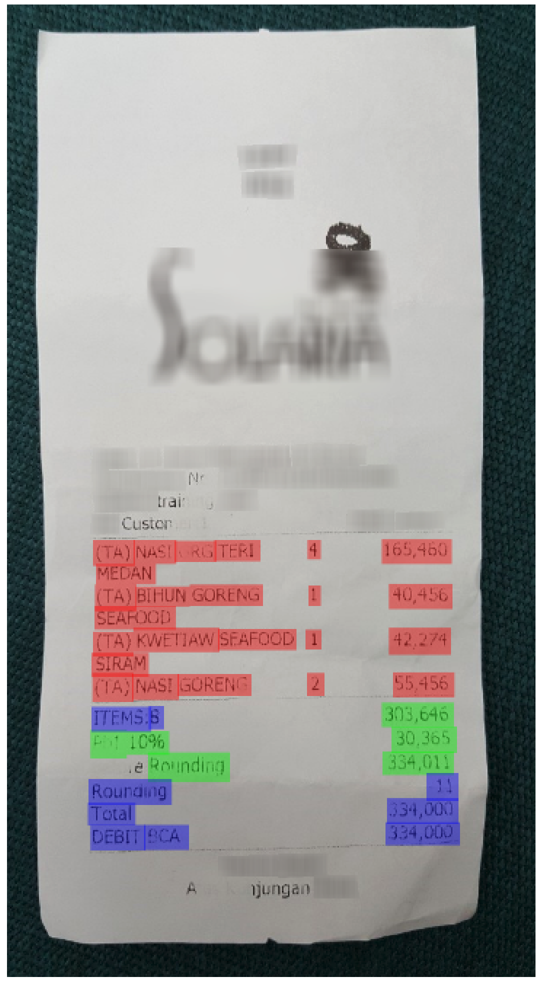
(a) Ground Truth

(b) DocFormer predictions

Figure 10: **DocFormer perfect predictions on CORD dataset**: Left image shows GT and right image is the prediction for file receipt_00004 made by DocFormer which perfectly matches with the GT despite the presence of distortion and background text.

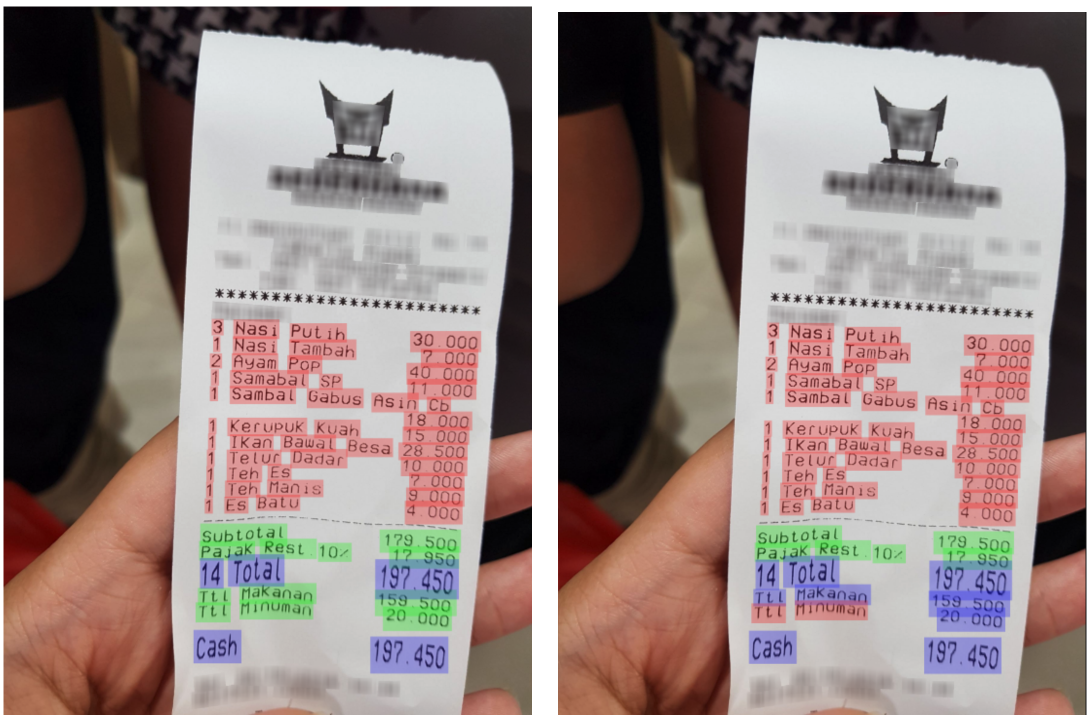


(a) Ground Truth



(b) DocFormer predictions

Figure 11: **DocFormer perfect predictions on CORD dataset**: Left image shows GT and right image is the prediction for file receipt_00051 made by DocFormer which perfectly matches with GT. Note that the faded out text which is hard to OCR is correctly classified due to multi-modal self-attention features.



(a) Ground Truth

(b) DocFormer predictions

Figure 12: **DocFormer Partially correct predictions on CORD dataset**: Left image shows GT and right image is the prediction for file receipt_00085 made by DocFormer with a misclassification of tokens of category SUBTOTAL with TOTAL items. This could be due to the rarity of SUBTOTAL tokens appearing below TOTAL tokens which DocFormer may not have encountered during training.

References

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

[2] Filip Graliński, Tomasz Stanisławek, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: A novel task for information extraction involving long documents with complex layout. *arXiv preprint arXiv:2003.02356*, 2020.

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

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

[5] Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained language model for understanding texts in document. *under review https://openreview.net/references/pdf?id=uCz3OR6CJT*, 2020.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[7] Park Seunghyun, Shin Seung, Lee Bado, Lee Junyeop, Surh Jaeheung, Seo Minjoon, and Lee Hwalsuk. Cord: A consolidated receipt dataset for post-ocr parsing. 2019.

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

[9] 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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020.

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