

# DocFormer: End-to-End Transformer for Document Understanding

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

## Abstract

We present *DocFormer* - a multi-modal transformer based architecture for the task of Visual Document Understanding (VDU). VDU is a challenging problem which aims to understand documents in their varied formats (forms, receipts etc.) and layouts. In addition, *DocFormer* is pre-trained in an unsupervised fashion using carefully designed tasks which encourage multi-modal interaction. *DocFormer* uses text, vision and spatial features and combines them using a novel multi-modal self-attention layer. *DocFormer* also shares learned spatial embeddings across modalities which makes it easy for the model to correlate text to visual tokens and vice versa. *DocFormer* is evaluated on 4 different datasets each with strong baselines. *DocFormer* achieves state-of-the-art results on all of them, sometimes beating models 4x its size (in no. of parameters).

## 1. Introduction

The task of Visual Document Understanding (VDU) aims at understanding digital documents either born as PDF's or as images. VDU focuses on varied document related tasks like entity grouping, sequence labeling, document classification. While modern OCR engines [34] have become good at predicting text from documents, VDU often requires understanding both the structure and layout of documents. The use of text or even text and spatial features alone is not sufficient for this purpose. For the best results, one needs to exploit the text, spatial features and the image. One way to exploit all these features is using transformer models [5, 15, 53]. Transformers have recently been used for VDU [26, 56, 57]. These models differ in how the unsupervised pre-training is done, the way self-attention is modified for the VDU domain or how they fuse modalities (text and/or image and spatial). There have been text only [15], text plus spatial features only [26, 56] approaches for VDU. However, the holy-grail is to fuse all three modalities (text,

FILING FORM		TO BE FILED BY: POLITICAL COMMITTEES NOT DOMICILED IN WASHINGTON STATE (Sec. 9)	
TO THE STATE OF WASHINGTON PUBLIC DISCLOSURE COMMISSION CHAP. 1, LAWS OF 1973		THIS SPACE FOR OFFICE USE	
See completion instructions at bottom of page. (Type or print clearly)		P.M. DATE	DATE RECD.
NAME AND ADDRESS OF POLITICAL COMMITTEES Tobacco People's Public Affairs Comm. 1776 K Street, N. W. Washington, D. C. 20006		DATE PREPARED 1/29/74	THIS FORM <input type="checkbox"/> REPLACES <input type="checkbox"/> AMENDS PREVIOUS FILING PREPARED: (Mo.) (Day) (Yr.)
ITEM 1 PURPOSE(S) OF THE POLITICAL COMMITTEE support candidates for U. S. House and Senate			
ITEM 2 POLITICAL COMMITTEE'S OFFICERS OR RESPONSIBLE LEADERS			
NAME	ADDRESS	TITLE	
Earle C. Clements, Chairman	1776 K Street, N. W., DC	Chairman	
John F. Mills, Treasurer	1776 K Street, N. W. DC	Treasurer	

Figure 1: **Snippet of a Document:** Various VDU tasks on this document may include labeling each text token into fixed classes or grouping tokens into a semantic class and finding relationships between tokens e.g. (“DATE PREPARED” → Key and “1/29/74” → Value) or classifying the document into different categories. Note a document could have “other” text e.g. “C-5” which the model should ignore or classify as “other” depending on the task.

visual and spatial features). This is desirable since there is some information in text that visual features miss out (language semantics), and there is some information in visual features that text misses out (text font and visual layout for example).

Multi-modal training in general is difficult since one has to map a piece of text to an arbitrary span of visual content. For example in Figure 1, “ITEM 1” needs to be mapped to the visual region. Said a different way, text describes semantic high-level concept(s) e.g. the word “person” whereas visual features map to the pixels (of a person) in the image. It is not easy to enforce feature correlation across modalities from text ↔ image. We term this issue as *cross-modality feature correlation* and reference it later to show how *DocFormer* presents an approach to address this.

*DocFormer* follows the now common, pre-training and fine-tuning strategy. *DocFormer* incorporates a novel multi-modal self-attention with shared spatial embeddings in an encoder only transformer architecture. In addition, we pro-

pose three pre-training tasks of which two are novel unsupervised multi-modal tasks: *learning-to-reconstruct* and *multi-modal masked language modeling* task. Details are provided in Section 3. To the best of our knowledge, this is the first approach for doing VDU which does not use bulky pre-trained object-detection networks for visual feature extraction. DocFormer instead uses plain ResNet50 [22] features along with shared spatial (between text and image) embeddings which not only saves memory but also makes it easy for DocFormer to correlate text, visual features via spatial features. DocFormer is trained end-to-end with the visual branch trained from scratch. We now highlight the contributions of our paper:

- A novel multi-modal attention layer capable of fusing text, vision and spatial features in a document.
- Three unsupervised pre-training tasks which encourage multi-modal feature collaboration. Two of these are novel unsupervised multi-modal tasks: *learning-to-reconstruct* task and a *multi-modal masked language modeling* task.
- DocFormer is end-to-end trainable and it does not rely on a pre-trained object detection network for visual features simplifying its architecture. On four varied downstream VDU tasks, DocFormer achieves state of the art results. On some tasks it out-performs large variants of other transformer almost 4x its size (in the number of parameters). In addition, DocFormer does not use custom OCR unlike some of the recent papers [57, 26].

## 2. Background

Document understanding methods in the literature have used various combinations of image, spatial and text features in order to understand and extract information from structurally rich documents such as forms [19, 59, 13], tables [46, 58, 25], receipts [28, 27] and invoices [36, 44, 39]. Finding the optimal way to combine these multi-modal features is an active area of research.

Grid based methods [30, 14] were proposed for invoice images where text pixels are encoded using character or word vector representations and classified into field types such as Invoice Number, Date, Vendor Name and Address etc. using a convolutional neural network.

BERT [15] is a transformer-encoder [53] based neural network that has been shown to work well on language understanding tasks. LayoutLM [56] modified the BERT architecture by adding 2D spatial coordinate embeddings along with 1D position and text token embeddings. They also added visual features for each word token, obtained using a Faster-RCNN and its bounding box coordinates.

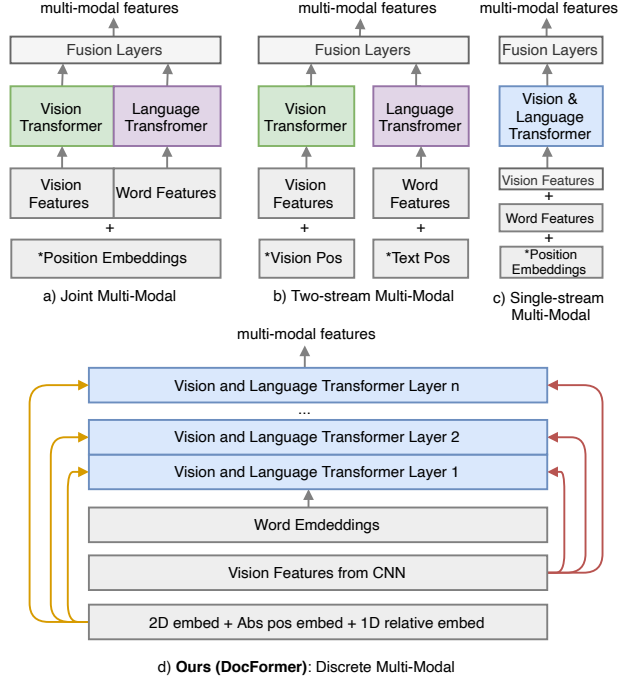


Figure 2: Conceptual Comparisons of **Transformer Multi-Modal Encoder Architectures**: The mechanisms differ in how the modalities are combined. **Type A)** Joint Multi-Modal: like VL-BERT[48], LayoutLMv2[57], VisualBERT [33], MMBT[31], UNITER [9] **Type B)** Two-stream Multi-Modal: CLIP[42], ViBERT[38], **Type C)** Single-stream Multi-Modal, **Type D)** Ours: Discrete Multi-modal. e.g. DocFormer . Note: in each transformer layer, each input modality is self-attended separately. Best viewed in color.

LayoutLM was pre-trained on 11 million unlabeled pages and was then finetuned on several document understanding tasks - form processing, classification and receipt processing. This idea of pre-training on large datasets and then finetuning on several related downstream tasks is also seen in general vision and language understanding work [48, 38, 31, 33] etc. Figure 2 shows a comparison of multi-modal transformer encoder architectures.

Recently, LayoutLMv2 [57] improved over LayoutLM by changing the way visual features are input to the model - treating them as separate tokens as opposed to adding visual features to the corresponding text tokens. Further, additional pre-training tasks were explored to make use of unlabeled document data.

BROS [27] also uses a BERT based encoder, with a graph-based classifier based on SPADE [29], which is used to predict entity relations between text tokens in a document. They also use 2D spatial embeddings added along with text tokens and evaluate their network on forms, receipts document images. Multi-modal transformer encoder-decoder architectures based on T5 [43] have been proposed