FormNetV2: Multimodal Graph Contrastive Learning for Form Document Information Extraction

Chen-Yu Lee[1][*] **Chun-Liang Li**[1]**, Hao Zhang**[2]**, Timothy Dozat**[2]**, Vincent Perot**[2]**,**
Guolong Su[2]**, Xiang Zhang**[1]**, Kihyuk Sohn**[2]**, Nikolai Glushnev**[3]**, Renshen Wang**[2]**,**
Joshua Ainslie[2]**, Shangbang Long**[2]**, Siyang Qin**[2]**, Yasuhisa Fujii**[2]**, Nan Hua**[2]**, Tomas Pfister**[1]
[1]Google Cloud AI Research, [2]Google Research, [3]Google Cloud AI

Abstract

The recent advent of self-supervised pre-training techniques has led to a surge in the use of multimodal learning in form document understanding. However, existing approaches that extend the mask language modeling to other modalities require careful multi-task tuning, complex reconstruction target designs, or additional pre-training data. In FormNetV2, we introduce a centralized multimodal graph contrastive learning strategy to unify self-supervised pre-training for all modalities in one loss. The graph contrastive objective maximizes the agreement of multimodal representations, providing a natural interplay for all modalities without special customization. In addition, we extract image features within the bounding box that joins a pair of tokens connected by a graph edge, capturing more targeted visual cues without loading a sophisticated and separately pre-trained image embedder. FormNetV2 establishes new state-of-the-art performance on FUNSD, CORD, SROIE and Payment benchmarks with a more compact model size.

1 Introduction

Automated information extraction is essential for many practical applications, with form-like documents posing unique challenges compared to article-like documents, which have led to an abundance of recent research in the area. In particular, form-like documents often have complex layouts that contain structured objects like tables, columns, and fillable regions. Layout-aware language modeling has been critical for many successes (Xu et al., 2020; Majumder et al., 2020; Lee et al., 2022).

To further boost the performance, many recent approaches adopt multiple modalities (Xu et al., 2021; Huang et al., 2022; Appalaraju et al., 2021). Specifically, the image modality adds more structural information and visual cues to the existing layout and text modalities. They therefore extend the masked language modeling (MLM) from text to masked image modeling (MIM) for image and text-image alignment (TIA) for cross-modal learning. The alignment objective may also help to prime the layout modality, though it does not directly involve text layouts or document structures.

In this work, we propose FormNetV2, a multimodal transformer model for form information extraction. Unlike existing works – which may use the whole image as one representation (Appalaraju et al., 2021), or image patches (Xu et al., 2021), or image features of token bounding boxes (Xu et al., 2020) – we propose using image features extracted from the region bounded by a *pair* of tokens connected in the constructed graph. This allows us to capture a richer and more targeted visual component of the intra- and inter-entity information. Furthermore, instead of using multiple self-supervised objectives for each individual modality, we introduce graph contrastive learning (Li et al., 2019; You et al., 2020; Zhu et al., 2021) to learn multimodal embeddings jointly. These two additions to FormNetV1 (Lee et al., 2022) enable the graph convolutions to produce better super-tokens, resulting in both improved performance and a smaller model size.

In experiments, FormNetV2 outperforms its predecessor FormNetV1 as well as the existing multimodal approaches on four standard benchmarks. In particular, compared with FormNetV1, FormNetV2 outperforms it by a large margin on FUNSD (86.35 v.s. 84.69) and Payment (94.90 v.s. 92.19); compared with DocFormer (Appalaraju et al., 2021), FormNetV2 outperforms it on FUNSD and CORD with nearly 2.5x less number of parameters.

---

[*]All work done at Google. Correspondence to: Chen-Yu Lee <chenyulee@google.com>, Chun-Liang Li <chunliang@google.com>

2 Related Work

Early works on form document information extraction are based on rule-based models or learning-based models with handcrafted features (Lebourgeois et al., 1992; O'Gorman, 1993; Ha et al., 1995; Simon et al., 1997; Marinai et al., 2005; Chiticariu et al., 2013). Later on, various deep neural models have been proposed, including methods based on recurrent nets (Palm et al., 2017; Aggarwal et al., 2020), convolutional nets (Katti et al., 2018; Zhao et al., 2019; Denk and Reisswig, 2019), and transformers (Majumder et al., 2020; Garncarek et al., 2020).

Recently, in addition to the text, researchers have explored the layout attribute in form document modeling, such as the OCR word reading order (Lee et al., 2021; Gu et al., 2022b), text coordinates (Majumder et al., 2020; Xu et al., 2020; Garncarek et al., 2020; Li et al., 2021a; Lee et al., 2022), layout grids (Lin et al., 2021), and layout graphs (Lee et al., 2022). The image attribute also provides essential visual cues such as fonts, colors, and sizes. Other visual signals can be useful as well, including logos and separating lines from form tables. Xu et al. (2020) uses Faster R-CNN (Ren et al., 2015) to extract token image features; Appalaraju et al. (2021) uses ResNet50 (He et al., 2016) to extract full document image features; Li et al. (2022) use ViT (Dosovitskiy et al., 2020) with FPN (Lin et al., 2017) to extract non-overlapping patch image features. These sophisticated image embedders require a separate pre-training step using external image datasets (e.g. ImageNet (Russakovsky et al., 2015) or PubLayNet (Zhong et al., 2019)), and sometimes depend upon a visual codebook pre-trained by a discrete variational autoencoder (dVAE).

When multiple modalities come into play, different supervised or self-supervised multimodal pre-training techniques have been proposed. They include mask prediction, reconstruction, and matching for one or more modalities (Xu et al., 2020, 2021; Appalaraju et al., 2021; Li et al., 2021b; Gu et al., 2022a; Huang et al., 2022; Li et al., 2022; Pramanik et al., 2020). Next-word prediction (Kim et al., 2022) or length prediction (Li et al., 2021c) have been studied to bridge text and image modalities. Direct and relative position predictions (Cosma et al., 2020; Wei et al., 2020; Li et al., 2021a; Wang et al., 2022a; Li et al., 2021c) have been proposed to explore the underlying layout semantics of documents. Nevertheless, these pre-training objectives require strong domain expertise, specialized designs, and multi-task tuning between involved modalities. In this work, our proposed graph contrastive learning performs multimodal pre-training in a centralized design, unifying the interplay between all involved modalities without the need for prior domain knowledge.

3 FormNetV2

We briefly review the backbone architecture FormNetV1 (Lee et al., 2022) in Sec 3.1, introduce the multimodal input design in Sec 3.2, and detail the multimodal graph contrastive learning in Sec 3.3.

3.1 Preliminaries

ETC. FormNetV1 (Lee et al., 2022) uses Extended Transformer Construction (ETC; Ainslie et al., 2020) as the backbone to work around the quadratic memory cost of attention for long form documents. ETC permits only a few special tokens to attend to every token in the sequence (global attention); all other tokens may only attend to k local neighbors within a small window, in addition to these special tokens (local attention). This reduces the computational complexity from $O(n^2)$ query-key pairs that need scoring to $O(kn)$. Eq. (2) formalizes the computation of the attention vector $\mathbf{a}_0$ for a model with one global token at index 0, and Eq. (2) formalizes computation of the attention vector $\mathbf{a}_{i>0}$ for the rest of the tokens in the model.

$$\mathbf{a}_0 = \texttt{attend}(\mathbf{h}_0, [\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_n]) \qquad (1)$$
$$\mathbf{a}_{i>0} = \texttt{attend}(\mathbf{h}_i, [\mathbf{h}_0, \mathbf{h}_{i-k}, \ldots, \mathbf{h}_{i+k}]) \quad (2)$$

Rich Attention. To address the distorted semantic relatedness of tokens created by imperfect OCR serialization, FormNetV1 adapts the attention mechanism to model spatial relationships between tokens by proposing Rich Attention, a mathematically sound way of conditioning attention on low-level spatial features without resorting to quantizing the document into regions associated with distinct embeddings in a lookup table. In Rich Attention, the model constructs the (pre-softmax) attention score (Eq. 10) from multiple components: the usual transformer attention score (Eq. 7); the order of tokens along the x-axis and the y-axis (Eq. 8); and the log distance (in number of pixels) between tokens, again along both axes (Eq. 9). The expression for a transformer head with Rich Attention on the x-axis is provided in Eqs. (3–10); we
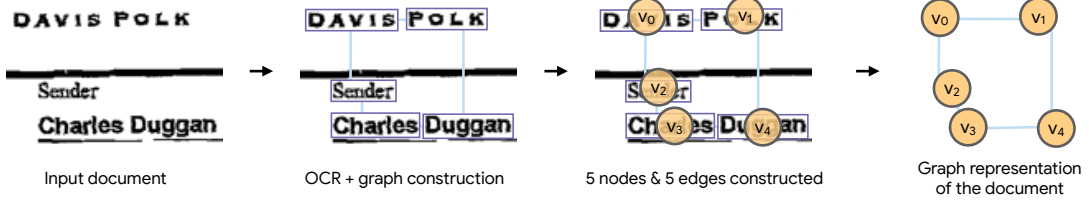
Figure 1: Graph of a sample region from a form. Token bounding boxes are identified, and from them the graph is constructed. Nodes are labeled and the graph structure is shown abstracted away from its content.
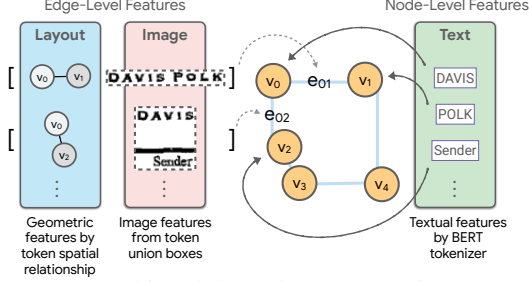

Figure 2: Multimodal graph representations are composed from three modalities: text at node-level; concatenation of layout and image at edge-level.
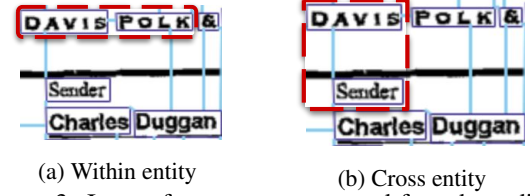

(a) Within entity   (b) Cross entity

Figure 3: Image features are extracted from bounding boxes (red) that join pairs of tokens connected by edges to capture (a) similar patterns within an entity, or (b) dissimilar patterns or separating lines between entities.

refer the interested reader to Lee et al. (2022) for further details.

$$o_{ij} = \texttt{int}(x_i < x_j) \qquad (3)$$

$$d_{ij} = \ln(1 + |x_i - x_j|) \qquad (4)$$

$$p_{ij} = \texttt{Sigmoid}(\text{affine}^{(p)}([\mathbf{q}_i; \mathbf{k}_j])) \qquad (5)$$

$$\mu_{ij} = \text{affine}^{(\mu)}([\mathbf{q}_i; \mathbf{k}_j]) \qquad (6)$$

$$s_{ij}^{(t)} = \mathbf{q}_i^\top \mathbf{k}_j \qquad (7)$$

$$s_{ij}^{(o)} = o_{ij} \ln(p_{ij}) + (1 - o_{ij}) \ln(1 - p_{ij}) \qquad (8)$$

$$s_{ij}^{(d)} = -\frac{\theta^2 (d_{ij} - \mu_{ij})^2}{2} \qquad (9)$$

$$s_{ij} = s_{ij}^{(t)} + s_{ij}^{(o)} + s_{ij}^{(d)} \qquad (10)$$

GCN. Finally, FormNetV1 includes a graph convolutional network (GCN) contextualization step *before* serializing the text to send to the ETC transformer component. The graph for the GCN locates up to K neighbors for each token – defined broadly by geographic "nearness" – before convolving their token embeddings to build up supertoken representations as shown in Figure 1. This allows the network to build a weaker but more complete picture of the layout modality than Rich Attention, which is constrained by local attention.

The final system was pretrained end-to-end with a standard masked language modeling (MLM) objective. See Sec A.3 in Appendix for more details.

3.2 Multimodal Input

In FormNetV2, we propose adding the image modality to the model in addition to the text and layout modalities that are already used in FormNetV1 (Sec 3.3 in Lee et al. (2022)). We expect that image features from documents contain information absent from the text or the layout, such as fonts, colors, and sizes of OCR words.

To do this, we run a ConvNet to extract dense image features on the whole document image, and then use Region-of-Interest (RoI) pooling (He et al., 2017) to pool the features within the bounding box that joins a pair of tokens connected by a GCN edge. Finally, the RoI pooled features go through another small ConvNet for refinement. After the image features are extracted, they are injected into the network through concatenation with the existing layout features at edges of the GCN. Figure 2 illustrates how all three modalities are utilized in this work and Sec 4.2 details the architecture.

Most of the recent approaches (Table 1) that incorporate image modality extract features from either (a) the whole image as one vector, (b) non-overlapping image patches as extra input tokens to transformers, or (c) token bounding boxes that are added to the text features for all tokens.

However, form document images often contain OCR words that are relatively small individually and are densely distributed in text blocks. They also contain a large portion of the background region without any texts. Therefore, the aforementioned method (a) only generates global visual representations with large noisy background regions but not
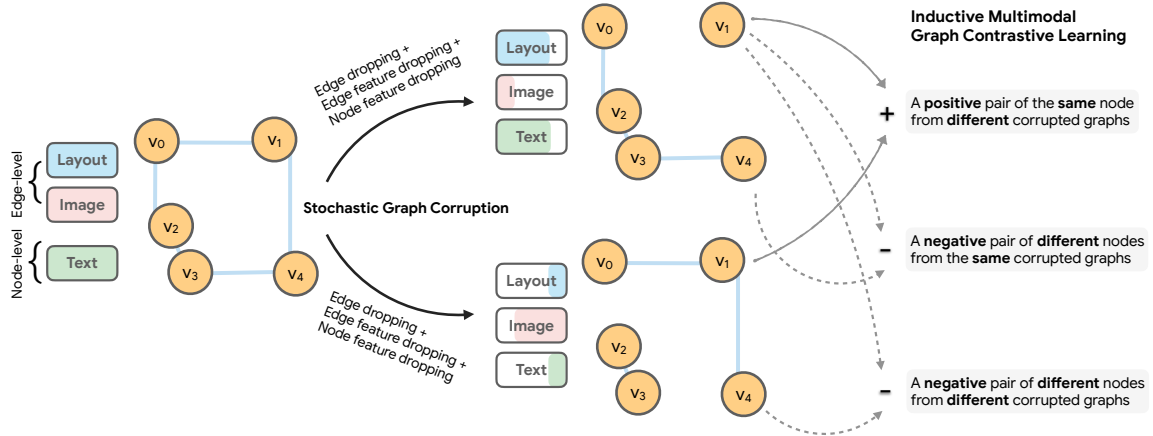
Figure 4: Multimodal graph contrastive learning. Two corrupted graphs are sampled from an input graph by corruption of graph topology (edges) and attributes (multimodal features). The system is trained to identify which pair of nodes across all pairs of corrupted nodes (including within the same graph) came from the same node.

targeted entity representations; method (b) tends to be sensitive to the patch size and often chops OCR words or long entities to different patches, while also increasing computational cost due to the increased token length; and method (c) only sees regions within each token's bounding box and lacks context between or outside of tokens.

On the other hand, the proposed edge-level image feature representation can precisely model the relationship between two nearby, potentially related "neighbor" tokens and the surrounding region, while ignoring all irrelevant or distracting regions. Figure 3 demonstrates that the targeted RoI image feature pooling through the union bounding box can capture any similar patterns (e.g. font, color, size) within an entity (left) or dissimilar patterns or separating lines between entities (right). See Sec 4.4 for detailed discussion.

3.3 Multimodal Graph Contrastive Learning

Previous work in multimodal document understanding requires manipulating multiple supervised or self-supervised objectives to learn embeddings from one or multiple modalities during pre-training. By contrast, in FormNetV2, we propose utilizing the graph representation of a document to learn multimodal embeddings with a contrastive loss.

Specifically, we first perform stochastic graph corruption to sample two corrupted graphs from the original input graph of each training instance. This step generates node embeddings based on partial contexts. Then, we apply a contrastive objective by maximizing agreement between tokens at node-level. That is, the model is asked to identify which pairs of nodes across all pairs of nodes – within the same graph and across graphs – came from the

same original node. We adopt the standard normalized temperature-scaled cross entropy (NT-Xent) loss formulation (Chen et al., 2020; Wu et al., 2018; Oord et al., 2018; Sohn, 2016) with temperature 0.1 in all experiments.

To build a centralized contrastive loss that unifies the interactions between multiple input modalities, we corrupt the original graph at both graph topology level and graph feature level. Topology corruption includes edge dropping by randomly removing edges in the original graph. Feature corruption includes applying dropping to all three modalities: dropping layout and image features from edges and dropping text features from nodes. Note that we only corrupt the graph in the GCN encoder and keep the ETC decoder intact to leverage the semantically meaningful graph representation of the document during graph contrastive learning.

To further diversify the contexts in two corrupted graphs and reduce the risk of training the model to over-rely on certain modalities, we further design an inductive graph feature dropping mechanism by adopting imbalanced drop-rates of modalities between the two corrupted graphs. Precisely, for a given modality, we discard p percent of the features in the first corrupted graph and discard $1-p$ percent of the features in the second corrupted graph. Experiments in Sec 4.4 show that $p = 0.8$ works best empirically and the inductive feature dropping mechanism provides further performance boost over the vanilla version. We stipulate that this boom-and-bust approach to regularization allows the model to learn rich, complex representations that take full advantage of the model's capacity without becoming overly dependent on specific feature interactions. Figure 4 illustrates the overall

process.

The proposed graph contrastive objective is also general enough in principle to adopt other corruption mechanisms (Zhu et al., 2020; Hassani and Khasahmadi, 2020; You et al., 2020; Velickovic et al., 2019). The multimodal feature dropping provides a natural playground to consume and allow interactions between multiple input modalities in one single loss design. It is straightforward to extend the framework to include more modalities without the need for hand crafting specialized loss by domain experts. To the best of our knowledge, we are the first to use graph contrastive learning during pre-training for form document understanding.

4 Evaluation

4.1 Datasets

FUNSD. FUNSD (Jaume et al., 2019) contains a collection of research, marketing, and advertising forms that vary extensively in their structure and appearance. The dataset consists of 199 annotated forms with 9,707 entities and 31,485 word-level annotations for 4 entity types: header, question, answer, and other. We use the official 75-25 split for the training and test sets.

CORD. CORD (Park et al., 2019) contains over 11,000 Indonesian receipts from shops and restaurants. The annotations are provided in 30 fine-grained semantic entities such as store name, quantity of menu, tax amount, discounted price, etc. We use the official 800-100-100 split for training, validation, and test sets.

SROIE. The ICDAR 2019 Challenge on Scanned Receipts OCR and key Information Extraction (SROIE) (Huang et al., 2019) offers 1,000 whole scanned receipt images and annotations. 626 samples are for training and 347 samples are for testing. The task is to extract four predefined entities: company, date, address, or total.

Payment. We use the large-scale payment data (Majumder et al., 2020) that consists of roughly 10,000 documents and 7 semantic entity labels from human annotators. We follow the same evaluation protocol and dataset splits used in Majumder et al. (2020).

4.2 Experimental Setup

We follow the FormNetV1 (Lee et al., 2022) architecture with a slight modification to incorporate multiple modalities used in the proposed method.

Our backbone model consists of a 6-layer GCN encoder to generate structure-aware super-tokens, followed by a 12-layer ETC transformer decoder equipped with Rich Attention for document entity extraction. The number of hidden units is set to 768 for both GCN and ETC. The number of attention heads is set to 1 in GCN and 12 in ETC. The maximum sequence length is set to 1024. We follow Ainslie et al. (2020); Lee et al. (2022) for other hyper-parameter settings. For the image embedder architecture, see Sec A.1 in Appendix.

Pre-training. We pre-train FormNetV2 using two unsupervised objectives: Masked Language Modeling (MLM) (Taylor, 1953; Devlin et al., 2019) and the proposed multimodal Graph Contrastive Learning (GCL).

Different from BERT (Devlin et al., 2019), here MLM has access to layout and image modalities during pre-training similar to Appalaraju et al. (2021); Xu et al. (2021, 2020). Nevertheless, the layout and image features are constructed at edge level instead of at node level, supplementing the text features for better underlying representation learning without directly leaking the trivial information.

GCL provides a natural playground for effective interactions between all three modalities from a document in a contrastive fashion. For each graph representation of a document, we generate two corrupted views by edge dropping, edge feature dropping, and node feature dropping with dropping rates {0.3, 0.8, 0.8}, respectively. The weight matrices in both GCN and ETC are shared across the two views.

We follow Appalaraju et al. (2021); Xu et al. (2021, 2020) and use the large-scale IIT-CDIP document collection (Lewis et al., 2006) for pre-training, which contains 11 million document images. We train the models from scratch using Adam optimizer with batch size of 512. The learning rate is set to 0.0002 with a warm-up proportion of 0.01. We find that GCL generally converges faster than MLM, therefore we set the loss weightings to 1 and 0.5 for MLM and GCL, respectively.

Note that we do not separately pre-train or load a pre-trained checkpoint for the image embedder as done in other recent approaches shown in Table 1. In fact, in our implementation, we find that using sophisticated image embedders or pre-training with natural images, such as ImageNet (Russakovsky et al., 2015), do not improve the final downstream

Dataset	Method	P	R	F1	F1[†]	Modality	Image Embedder	#Params
FUNSD	SPADE (Hwang et al., 2021)	-	-	70.5	-	T+L	-	110M
	UniLMv2 (Bao et al., 2020)	67.80	73.91	70.72	-	T	-	355M
	LayoutLMv1 (Xu et al., 2020)	75.36	80.61	77.89	-	T+L	-	343M
	DocFormer (Appalaraju et al., 2021)	81.33	85.44	83.33	-	T+L+I	ResNet50	502M
	FormNetV1 (Lee et al., 2022)	85.21	84.18	84.69	-	T+L	-	217M
	LayoutLMv1 (Xu et al., 2020)	76.77	81.95	79.27	-	T+L+I	ResNet101	160M
	LayoutLMv2 (Xu et al., 2021)	83.24	85.19	84.20	-	T+L+I	ResNeXt101-FPN	426M
	DocFormer (Appalaraju et al., 2021)	82.29	86.94	84.55	-	T+L+I	ResNet50	536M
	StructuralLM (Li et al., 2021a)	-	-	-	85.14	T+L	-	355M
	LayoutLMv3 (Huang et al., 2022)	81.35	83.75	82.53	92.08	T+L+I	Tokenization	368M
	FormNetV2 (ours)	85.78	86.94	**86.35**	**92.51**	T+L+I	3-layer ConvNet	204M
CORD	SPADE (Hwang et al., 2021)	-	-	91.5	-	T+L	-	110M
	UniLMv2 (Bao et al., 2020)	91.23	92.89	92.05	-	T	-	355M
	LayoutLMv1 (Xu et al., 2021)	94.32	95.54	94.93	-	T+L	-	343M
	DocFormer (Appalaraju et al., 2021)	96.46	96.14	96.30	-	T+L+I	ResNet50	502M
	FormNetV1 (Lee et al., 2022)	98.02	96.55	97.28	-	T+L	-	345M
	LayoutLMv2 (Xu et al., 2021)	95.65	96.37	96.01	-	T+L+I	ResNeXt101-FPN	426M
	TILT (Powalski et al., 2021)	-	-	96.33	-	T+L+I	U-Net	780M
	DocFormer (Appalaraju et al., 2021)	97.25	96.74	96.99	-	T+L+I	ResNet50	536M
	LayoutLMv3 (Huang et al., 2022)	95.82	96.03	95.92	97.46	T+L+I	Tokenization	368M
	FormNetV2 (ours)	97.74	97.00	**97.37**	**97.70**	T+L+I	3-layer ConvNet	204M
SROIE	UniLMv2 (Bao et al., 2020)	-	-	94.88	-	T	-	355M
	LayoutLMv1 (Xu et al., 2021)	95.24	95.24	95.24	-	T+L	-	343M
	LayoutLMv2 (Xu et al., 2021)	99.04	96.61	97.81	-	T+L+I	ResNeXt101-FPN	426M
	FormNetV2 (ours)	98.56	98.05	**98.31**	-	T+L+I	3-layer ConvNet	204M
Payment	NeuralScoring (Majumder et al., 2020)	-	-	87.80	-	T+L	-	-
	FormNetV1 (Lee et al., 2022)	92.70	91.69	92.19	-	T+L	-	217M
	FormNetV2 (ours)	94.11	95.71	**94.90**	-	T+L+I	3-layer ConvNet	204M

Table 1: Entity-level precision, recall, and F1 score comparisons on four standard benchmarks. "T/L/I" denotes "text/layout/image" modality. The proposed FormNetV2 establishes new state-of-the-art results on all four datasets. FormNetV2 significantly outperforms the most recent DocFormer (Appalaraju et al., 2021) and LayoutLMv3 (Huang et al., 2022) while using a 38% and 55% sized model, respectively. Note that LayoutLMv3 (Huang et al., 2022) and StructuralLM (Li et al., 2021a) use segment-level layout positions that incorporate ground truth entity bounding boxes, which is less practical for real-world applications. We nevertheless report our results under the same protocol in column F1[†]. See Sec 4.3 and Sec A.2 in Appendix for details.

entity extraction F1 scores, and they sometimes even degrade the performance. This might be because the visual patterns presented in form documents are drastically different from natural images that have multiple real objects. The best practice for conventional vision tasks (classification, detection, segmentation) might not be optimal for form document understanding.

Fine-tuning. We fine-tune all models for the downstream entity extraction tasks in the experiments using Adam optimizer with batch size of 8. The learning rate is set to 0.0001 without warm-up. The fine-tuning is conducted on Tesla V100 GPUs for approximately 10 hours on the largest corpus. Other hyper-parameters follow the settings in Lee et al. (2022).

4.3 Benchmark Results

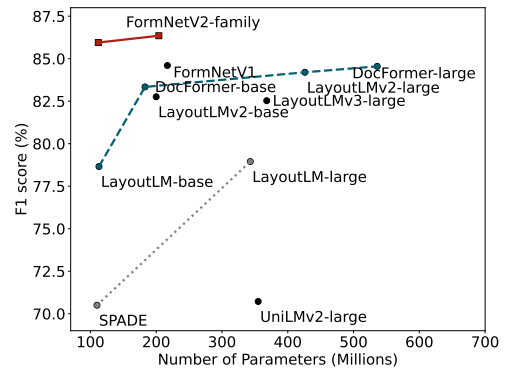Table 1 lists the results that are based on the same evaluation protocal[1].

Figure 5: **Model Size vs. Entity Extraction F1 Score** on FUNSD benchmark. The FormNetV2 family significantly outperforms other recent approaches – FormNetV2 achieves highest F1 score (86.35%) while using a 2.6x smaller model than DocFormer (84.55%; Appalaraju et al., 2021). FormNetV2 also outperforms FormNetV1 (Lee et al., 2022) by a large margin (1.66 F1) while using fewer parameters.

As the field is actively growing, researchers have started to explore incorporating additional

---

[1]Micro-F1 for FUNSD, CORD, and SROIE by following the implementation in Xu et al. (2021); macro-F1 for Pay-

ment (Majumder et al., 2020).

information into the system. For example, LayoutLMv3 (Huang et al., 2022) and StructuralLM (Li et al., 2021a) use segment-level layout positions derived from ground truth entity bounding boxes – the {Begin, Inside, Outside, End, Single} schema information (Ratinov and Roth, 2009) that determine the spans of entities are given to the model, which is less practical for real-world applications. We nevertheless report our results under the same protocol in column $\textbf{F1}^{\dagger}$ in Table 1. We also report LayoutLMv3 results without ground-truth entity segments for comparisons.

Furthermore, UDoc (Gu et al., 2022a) uses additional paragraph-level supervision returned by a third-party OCR engine EasyOCR[2]. Additional PubLayNet (Zhong et al., 2019) dataset is used to pre-train the vision backbone. UDoc also uses different training/test splits (626/247) on CORD instead of the official one (800/100) adopted by other works. ERNIE-mmLayout (Wang et al., 2022b) utilizes a third-party library spaCy[3] to provide external knowledge for the Common Sense Enhancement module in the system. The F1 scores on FUNSD and CORD are 85.74% and 96.31% without the external knowledge. We hope the above discussion can help clarify the standard evaluation protocol and decouple the performance improvement from modeling design vs. additional information.

Figure 5 shows model size vs. F1 score for the recent approaches that are directly comparable. The proposed method significantly outperforms other approaches in both F1 score and parameter efficiency: FormNetV2 achieves highest F1 score (86.35%) while using a 38% sized model than DocFormer (84.55%; Appalaraju et al., 2021). FormNetV2 also outperforms FormNetV1 (Lee et al., 2022) by a large margin (1.66 F1) while using fewer parameters. Table 1 shows that FormNetV2 outperforms LayoutLMv3 (Huang et al., 2022) and StructuralLM (Li et al., 2021a) with a considerable performance leap while using a 55% and 57% sized model, respectively. From Table 1 we also observe that using all three modalities (text+layout+image) generally outperforms using two modalities (text+layout), and using two modalities (text+layout) outperforms using one modality (text) only across different approaches.

4.4 Ablation Studies

We perform studies over the effect of image modality, graph contrastive learning, and decoupled graph corruption. The backbone for these studies is a 4-layer 1-attention-head GCN encoder followed by a 4-layer 8-attention-head ETC transformers decoder with 512 hidden units. The model is pre-trained on the 1M IIT-CDIP subset. All other hyperparameters follow Sec 4.2.

Effect of Image Modality and Image Embedder. Table 2 lists results of FormNetV1 (a) backbone only, (b) with additional tokens constructed from image patches[4], and (c) with the proposed image feature extracted from edges of a graph. The networks are pre-trained with MLM only to showcase the impact of input with image modality.

We observe that while (b) provides slight F1 score improvement, it requires 32% additional parameters over baseline (a). The proposed (c) approach achieves a significant F1 boost with less than 1% additional parameters over baseline (a). Secondly, we find the performance of more advanced image embedders (He et al., 2016) is inferior to the 3-layer ConvNet used here, which suggests that these methods may be ineffective in utilizing image modality. Nevertheless, the results demonstrate the importance of image modality as part of the multimodal input. Next we will validate the importance of an effective multimodal pre-training mechanism through graph contrastive learning.

Method	FUNSD	CORD	#Params
FormNetV1	82.53	95.16	81.7M
FormNetV1+Image Patch	82.65	95.43	107.0M
FormNetV1+Edge Image (ours)	83.13	95.85	82.3M

Table 2: F1 with different image modality setups.

Effect of Graph Contrastive Learning. The graph corruption step (Figure 4) in the proposed multimodal graph contrastive learning requires corruption of the original graph at both topology and feature levels. Considering the corruption happens in multiple places: edges, edge features, and node features, a naive graph corruption implementation would be to use the same drop-rate value everywhere. In Figure 6(a)(b), we show the downstream entity extraction F1 scores on FUNSD and CORD datasets by varying the dropping rate value during the graph contrastive pre-training. The selected

---

[2]https://github.com/JaidedAI/EasyOCR
[3]spacy.io

[4]We experiment with 32x32 image patch size, resulting in additional 256 image tokens to the model.
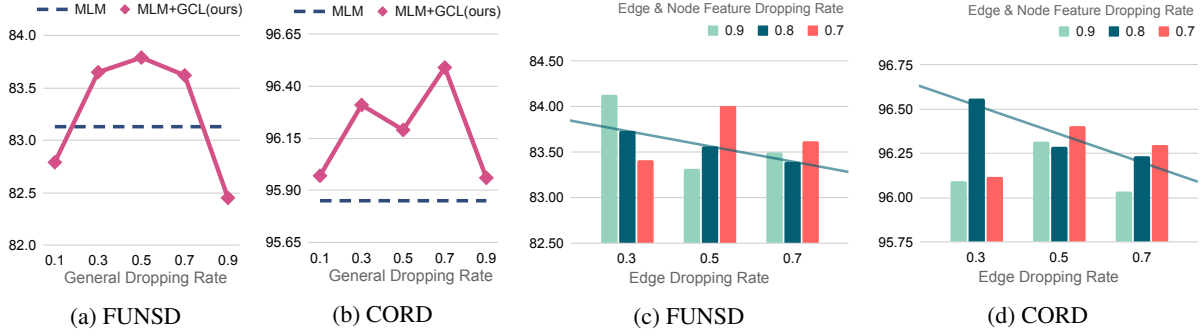
Figure 6: **Entity Extraction F1 Score vs. Graph Corruption Mechanism** on FUNSD and CORD benchmarks. (a)(b) show results using the same drop-rate across modalities. The proposed multimodal graph contrastive learning improves MLM pretraining at almost all drop-rates; (c)(d) show results using different drop-rates across modalities. The decoupled dropping mechanism permits further boosts to the F1 scores over non-decoupled counterparts. See Sec 4.4 for discussion.

dropping rate is shared across all aforementioned places.

Results show that the proposed multimodal graph contrastive learning works out of the box across a wide range of dropping rates. It demonstrates the necessity of multimodal corruption at both topology level and feature level – it brings up to 0.66% and 0.64% F1 boost on FUNSD and CORD respectively, when the model is pre-trained on MLM plus the proposed graph contrastive learning over MLM only. Our method is also stable to perturbation of different drop-rates.

We observe less or no performance improvement when extreme drop-rates are used; for example, dropping 10% edges and features or dropping 90% edges and features. Intuitively, dropping too few or too much information provides either no node context changes or too few remaining node contexts in different corrupted graphs for effective contrastive learning.

Effect of Decoupled Graph Corruption. In this study, we investigate whether decoupling the drop-rate in different places of graph corruption can learn better representations during pre-training and bring further improvement to the downstream entity extraction tasks. Specifically, we select different dropping rates for all four different places: edge, layout and image features at edge level, and text features at node level. At feature level (layout, image, text), when one of the corrupted graphs selects dropping rate p for a certain feature, the other corrupted graph will use the complement of the selected dropping rate $1-p$ for the same feature as introduced in Sec 3.3. This inductive multimodal contrastive design creates stochastically imbalanced information access to the features between two cor-

rupted views. It provides more diverse contexts at node level in different views and makes the optimization of the contrastive objective harder, ideally generating more semantically meaningful representations between the three modalities.

Figure 6(c)(d) show the downstream entity extraction F1 scores on FUNSD and CORD datasets by pre-training with three different edge dropping rates and three different feature dropping rates. We observe that decoupling the dropping rate at various levels further boosts the performance on both datasets – it brings another 0.34% and 0.07% F1 boost on FUNSD and CORD respectively, when decoupled dropping rates are used over the non-decoupled ones.

We also observe nonlinear interactions between different dropping rates at edge level and feature level. The best performing feature dropping rate might be sub-optimal when a different edge dropping rate is applied. This is noteworthy but not surprising behavior, since different edge dropping rates would drastically change the graph topology (and therefore the node embeddings). We expect the amount of information needed for maximizing the agreement of node contexts between two corrupted graphs to be different when the graph topology is altered. Nevertheless, we find that low edge dropping rates (e.g. 0.3) generally perform better than high edge dropping rates, and therefore select a low edge dropping rate in our final design.

Visualization. We visualize (Vig, 2019) the local-to-local attention scores of a CORD example for model pre-trained with MLM only and MLM+GCL but before fine-tuning in Figure 7(a). We observe that with GCL, the model can identify more meaningful token clusterings, leveraging

(a) Attention scores w/ and w/o GCL

(b) Model outputs for difficult cases.

Figure 7: (a) The attention scores for MLM and MLM+GCL(Graph Contrastive Learning) models on CORD before fine-tuning. When pre-trained with the proposed GCL, the model can identify more meaningful token clusterings, leveraging multimodal input effectively; (b) Difficult cases where the model predictions do not match the human-annotated ground truth. In this visualization we highlight disagreements only.

multimodal input more effectively.

We also show sample model outputs that do not match the human-annotated ground truth in Figure 7(b). The model confuses between 'header' and 'other' on the top of the form and between 'question' and 'answer' for the multiple choice questions on the bottom half of the form. More visualization can be found in Figure 8 in Appendix.

5 Conclusion

FormNetV2 augments an existing strong FormNetV1 backbone with image features bounded by pairs of neighboring tokens and the graph contrastive objective that learns to differentiate between the multimodal token representations of two corrupted versions of an input graph. The centralized design sheds new light to the understanding of multimodal form understanding.

6 Limitations

Our work follows the general assumption that the training and test set contain the same list of predefined entities. Without additional or necessary modifications, the few-shot or zero-shot capability of the model is expected to be limited. Future work includes exploring prompt-based architectures to unify pre-training and fine-tuning into the same query-based procedure.

7 Ethics Consideration

We have read and compiled with the ACL Code of Ethics. The proposed FormNetV2 follows the prevailing large-scale pre-training then fine-tuning framework. Although we use the standard IIT-

CDIP dataset for pre-training in all experiments, the proposed method is not limited to using specific datasets for pre-training. Therefore, it shares the same potential concerns of existing large language models, such as biases from the pre-training data and privacy considerations. We suggest following a rigorous and careful protocol when preparing the pre-training data for public-facing applications.

References

Milan Aggarwal, Hiresh Gupta, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. Form2seq: A framework for higher-order form structure extraction. In *EMNLP*.

Joshua Ainslie, Santiago Ontañón, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured data in transformers. In *EMNLP*.

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

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *ICML*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *EMNLP*.

Adrian Cosma, Mihai Ghidoveanu, Michael Panaitescu-Liess, and Marius Popescu. 2020. Self-supervised representation learning on document images. In *International Workshop on Document Analysis Systems*, pages 103–117. Springer.

Timo I Denk and Christian Reisswig. 2019. Bert-grid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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*.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2020. Lambert: Layout-aware (language) modeling for information extraction. *arXiv preprint arXiv:2002.08087*.

Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, and Tong Sun. 2022a. Unified pretraining framework for document understanding. *arXiv preprint arXiv:2204.10939*.

Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022b. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4583–4592.

Jaekyu Ha, Robert M Haralick, and Ihsin T Phillips. 1995. Recursive xy cut using bounding boxes of connected components. In *ICDAR*.

Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*. PMLR.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*.

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 *ICDAR*.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial dependency parsing for semi-structured document information extraction. In *ACL-IJCNLP (Findings)*.

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

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. In *EMNLP*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Frank Lebourgeois, Zbigniew Bublinski, and Hubert Emptoz. 1992. A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. In *ICPR*.

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

Chen-Yu Lee, Chun-Liang Li, Chu Wang, Renshen Wang, Yasuhisa Fujii, Siyang Qin, Ashok Popat, and Tomas Pfister. 2021. Rope: Reading order equivariant positional encoding for graph-based document information extraction. In *ACL-IJCNLP*.

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*.

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

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. Dit: Self-supervised pre-training for document image transformer. *arXiv preprint arXiv:2203.02378*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and

Hongfu Liu. 2021b. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. 2019. Graph matching networks for learning the similarity of graph structured objects. In *International conference on machine learning*, pages 3835–3845. PMLR.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021c. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *CVPR*.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. 2021. Vibertgrid: a jointly trained multi-modal 2d document representation for key information extraction from documents. In *International Conference on Document Analysis and Recognition*, pages 548–563. Springer.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *ACL*.

Simone Marinai, Marco Gori, and Giovanni Soda. 2005. Artificial neural networks for document analysis and recognition. *IEEE Transactions on pattern analysis and machine intelligence*.

Lawrence O'Gorman. 1993. The document spectrum for page layout analysis. *IEEE Transactions on pattern analysis and machine intelligence*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rasmus Berg Palm, Ole Winther, and Florian Laws. 2017. Cloudscan-a configuration-free invoice analysis system using recurrent neural networks. In *ICDAR*.

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*.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *ICDAR*.

Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. 2020. Towards a multi-modal, multi-task learning based pre-training framework for document representation learning. *arXiv preprint arXiv:2009.14457*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Conference on Computational Natural Language Learning (CoNLL)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*.

Anikó Simon, J-C Pret, and A Peter Johnson. 1997. A fast algorithm for bottom-up document layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*.

Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep graph infomax. *ICLR*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *ACL: System Demonstrations*.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022a. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*.

Wenjin Wang, Zhengjie Huang, Bin Luo, Qianglong Chen, Qiming Peng, Yinxu Pan, Weichong Yin, Shikun Feng, Yu Sun, Dianhai Yu, et al. 2022b. Ernie-mmlayout: Multi-grained multimodal transformer for document understanding. *Proceedings of the 30th ACM International Conference on Multimedia*.

Mengxi Wei, Yifan He, and Qiong Zhang. 2020. Robust layout-aware ie for visually rich documents with pre-trained language models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2367–2376.

Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. 2021. Representing long-range context for graph neural networks with global attention. *Advances in*

Neural Information Processing Systems, 34:13266–13279.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL-IJCNLP*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*.

Xiaohui Zhao, Endi Niu, Zhuo Wu, and Xiaoguang Wang. 2019. Cutie: Learning to understand documents with convolutional universal text information extractor. In *ICDAR*.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *ICDAR*.

Yanqiao Zhu, Yichen Xu, Qiang Liu, and Shu Wu. 2021. An empirical study of graph contrastive learning. *arXiv preprint arXiv:2109.01116*.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*.

A Appendix

A.1 Image Embedder Architecture

Our image embedder is a 3-layer ConvNet with filter sizes {32, 64, 128} and kernel size 3 throughout. Stride 2 is used in the middle layer and stride 1 is used everywhere else. We resize the input document image to 512×512 with aspect ratio fixed and zero padding for the background region. After extracting the dense features of the whole input image, we perform feature RoI pooling (He et al., 2017) within the bounding box that joins a pair of tokens connected by a GCN edge. The height and width of the pooled region are set to 3 and 16, respectively. Finally, the pooled features go through another 3-layer ConvNet with filter size {64, 32, 16} and kernel size 3 throughout. Stride 2 is used in the first 2 layers horizontally and stride 1 is used everywhere else. To consume image modality in our backbone model, we simply concatenate the pooled image features with the existing layout features at edge level of GCN as shown in Figure 2.

A.2 More Implementation Details

We conduct additional experiments[5] on FUNSD and CORD using base and large versions of LayoutLMv3 (Huang et al., 2022). Instead of using entity segment indexes inferred from ground truth, we use word boxes provided by OCR. We observe considerable performance degradation when the model has access to word-level box information instead of segment-level. The results are shown in Table 3.

Method	Setting	FUNSD	CORD
LayoutLMv3-base	Reported	90.29	96.56
	Reproduced	90.59	95.85
	Word box	78.35	95.81
LayoutLMv3-large	Reported	92.08	97.46
	Reproduced	92.14	96.78
	Word box	82.53	95.92

Table 3: LayoutLMv3 results with entity segment indexes (reproduced) or word level indexes (word box). We observe considerable performance degradation when the model has access to word-level box information instead of segment-level.

A.3 Preliminaries

FormNetV1 (Lee et al., 2022) simplifies the task of document entity extraction by framing it as fundamentally text-centric, and then seeks to solve the

---

[5]github.com/Jyouhou/unilm-test

problems that immediately arise from this. Serialized forms can be very long, so FormNetV1 uses a transformer architecture with a local attention window (ETC) as the backbone to work around the quadratic memory cost of attention. This component of the system effectively captures the text modality.

OCR serialization also distorts strong cues of semantic relatedness – a word that is just above another word may be related to it, but if there are many tokens to the right of the upper word or to the left of the lower word, they will intervene between the two after serialization, and the model will be unable to take advantage of the heuristic that nearby tokens tend to be related. To address this, FormNetV1 adapts the attention mechanism to model spatial relationships between tokens using Rich Attention, a mathematically sound way of conditioning attention on low-level spatial features without resorting to quantizing the document into regions associated with distinct embeddings in a lookup table. This allows the system to build powerful representations from the layout modality for tokens that fall within the local attention window.

Finally, while Rich Attention maximizes the potential of local attention, there remains the problem of what to do when there are so many interveners between two related tokens that they do not fall within the local attention window and cannot attend to each other at all. To this end FormNetV1 includes a graph convolutional network (GCN) contextualization step *before* serializing the text to send to the transformer component. The graph for the GCN locates up to K potentially related neighbors for each token before convolving to build up the token representations that will be fed to the transformer after OCR serialization. Unlike with Rich Attention, which directly learns concepts like "above", "below", and infinitely many degrees of "nearness", the graph at this stage does not consider spatial relationships beyond "is a neighbor" and "is not a neighbor" – see Figure 1. This allows the network to build a weaker but more complete picture of the layout modality than Rich Attention, which is constrained by local attention. A similar architecture is also found to be useful in graph learning tasks by Wu et al. (2021).

Thus the three main components of FormNetV1 cover each other's weaknesses, strategically trading off representational power and computational efficiency in order to allow the system to construct

useful representations while simplifying the problem to be fundamentally textual rather than visual. The final system was pretrained end-to-end on large scale unlabeled form documents with a standard masked language modeling (MLM) objective.

A.4 Output Visualization

Figure 8 shows additional FormNetV2 model outputs on FUNSD.

A.5 License or Terms

Please see the license or terms for IIT-CDIP[6], FUNSD[7], CORD[8], and SROIE[9] in the corresponding footnotes.

---

[6]ir.nist.gov/cdip/README.txt
[7]guillaumejaume.github.io/FUNSD/work/
[8]github.com/clovaai/cord/blob/master/LICENSE-CC-BY
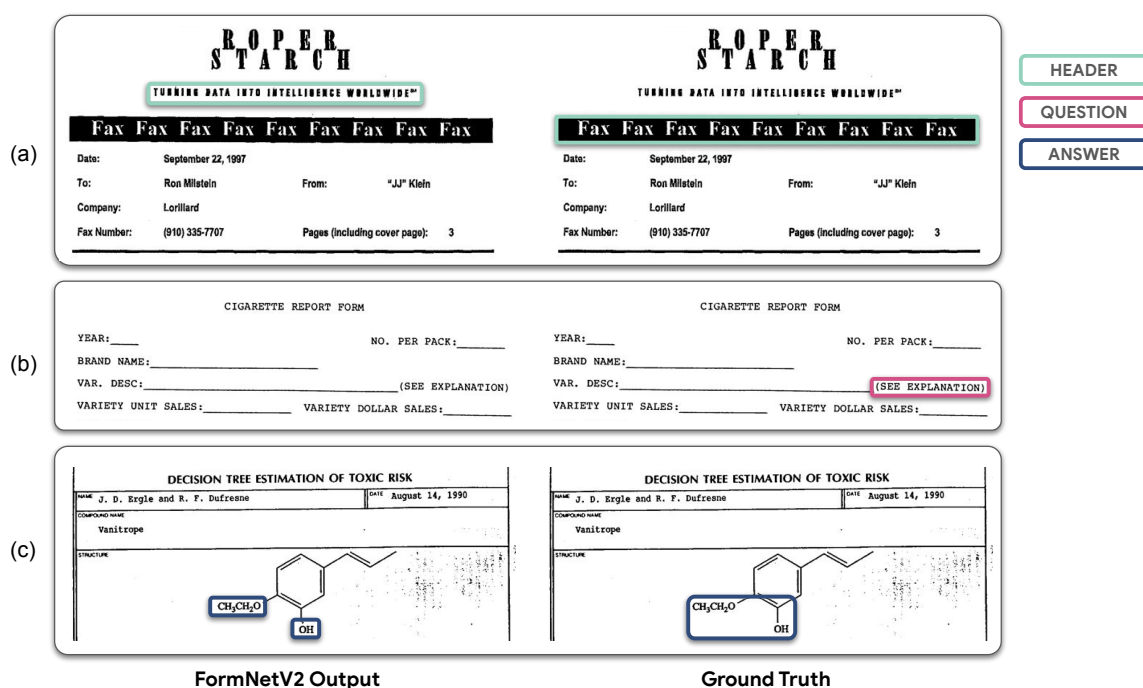[9]rrc.cvc.uab.es/?ch=13

Figure 8: The ambiguous cases where the model predictions do not match the human-annotated ground truth. In this visualization we only showcase mismatched entities.