DocumentNet: Bridging the Data Gap in Document Pre-Training

Lijun Yu[‡◇], Jin Miao[†], Xiaoyu Sun[†], Jiayi Chen[†], Alexander G. Hauptmann[‡],
Hanjun Dai[†], Wei Wei[†◇]
[‡]Carnegie Mellon University, [†]Google, [†]University of Virginia
[◇]lijun@cmu.edu, wewei@google.com

Abstract

Document understanding tasks, in particular, Visually-rich Document Entity Retrieval (VDER), have gained significant attention in recent years thanks to their broad applications in enterprise AI. However, publicly available data have been scarce for these tasks due to strict privacy constraints and high annotation costs. To make things worse, the non-overlapping entity spaces from different datasets hinder the knowledge transfer between document types. In this paper, we propose a method to collect massive-scale and weakly labeled data from the web to benefit the training of VDER models. The collected dataset, named DocumentNet, does not depend on specific document types or entity sets, making it universally applicable to all VDER tasks. The current DocumentNet consists of 30M documents spanning nearly 400 document types organized in a four-level ontology. Experiments on a set of broadly adopted VDER tasks show significant improvements when DocumentNet is incorporated into the pre-training for both classic and few-shot learning settings. With the recent emergence of large language models (LLMs), DocumentNet provides a large data source to extend their multi-modal capabilities for VDER.

1 Introduction

Document understanding is one of the most error-prone and tedious tasks many people have to handle every day. Advancements in machine learning techniques have made it possible to automate such tasks. In a typical Visually-rich Document Entity Retrieval (VDER) task, pieces of information are retrieved from the document based on a set of pre-defined entity types, known as the *schema*. For example, "amount", "date", and "item name" are major parts of an invoice schema.

The current setup of VDER tasks presents several unique challenges for acquiring sufficient training data. First, the availability of raw document images is greatly limited due to privacy constraints. Real-world documents, such as a driver's license or a bank statement, often contain personally identifiable information and are subject to access controls. Second, detailed annotation is costly and typically requires intensive training for experienced human annotators. *E.g.*, it takes deep domain knowledge to correctly label different fields in complex tax forms. Finally, knowledge sharing between various types of documents is constrained by inconsistent label spaces and contextual logic. For example, the entity sets (*i.e.*, schema) could be mutually exclusive, or the same entity type could take different semantic meanings in different contexts.

A number of models have been proposed for VDER tasks with various success (Huang et al., 2022; Lee et al., 2022; Appalaraju et al., 2021; Gu et al., 2021). To tackle the aforementioned challenges, most prior works initialize from a language model followed by BERT-style (Devlin et al., 2019) pre-training on document datasets with additional layout and visual features. However, even the largest dataset currently in use, *i.e.* IIT-CDIP (Lewis et al., 2006) dataset, has a limited size and only reflects a subset of document types.

In this paper, we introduce the method of building the DocumentNet dataset, which enables massive-scale pre-training for VDER modeling. DocumentNet is collected over the Internet using a pre-defined ontology, which spans hundreds of document types with a four-level hierarchy. Experiments demonstrated that DocumentNet is the key to advancing the performance on the commonly used FUNSD (Jaume et al., 2019), CORD (Park et al., 2019), and RVL-CDIP (Lewis et al., 2006) benchmarks in both classic and few-shot setups. More recently, LLMs (OpenAI, 2023; Anil et al., 2023) have shown great potential for VDER tasks given their reasoning capabilities. DocumentNet provides massive-scale multimodal data to boost the performance of LLMs for document understanding.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 707–722
December 6-10, 2023 ©2023 Association for Computational Linguistics

Dataset	#Samples↑	Ontology	Diverse Domains	High-quality OCR	Annotation
FUNSD (Jaume et al., 2019)	199				E=3
Kleister-NDA (Stanisławek et al., 2021)	540			✓	E=4
VRDU-Ad-buy (Wang et al., 2022b)	641			✓	E=14
SROIE (Huang et al., 2019)	973				E=4
CORD (Park et al., 2019)	1K				E=30
DeepForm (Borchmann et al., 2021)	1.1K			✓	E=5
VRDU-Registration (Wang et al., 2022b)	1.9K			✓	E=6
Kleister-Charity (Stanisławek et al., 2021)	2.7K			✓	E=8
DocVQA (Mathew et al., 2021)	12.8K			✓	Q
CC-PDF (Powalski et al., 2021)	350K		✓		
PubLayNet (Zhong et al., 2019)	358K		✓		B=5
RVL-CDIP (Lewis et al., 2006)	400K		✓		C=16
UCSF-IDL (Powalski et al., 2021)	480K		✓		
IIT-CDIP (Lewis et al., 2006)	11.4M		✓		
ImageNet (Deng et al., 2009)	1.3M images	✓	-	-	C=1K
ActivityNet (Caba Heilbron et al., 2015)	20K videos	✓	-	-	C=200
DocumentNet-v1 (ours)	9.9M	✓	✓	✓	**C=398, E=6**
DocumentNet-v2 (ours)	**30M**	✓	✓	✓	**C=398, E=6**

Table 1: Comparison between the proposed DocumentNet dataset and existing document understanding datasets. Datasets from other areas also built with ontology are listed in gray. Annotation includes class label (C), bounding box (B), entity (E), and question (Q), where the value refers to the number of classes.

2 Related Work

Tab. 1 provides an overview of relevant document datasets, with more details in App. B.1.

Single-domain document datasets. Many small document datasets with entity-span annotations have been used for tasks such as entity extraction. They contain less than 100k pages from a single domain. Newer datasets come with high-quality OCR annotation thanks to the advantage of relevant tools, while older ones, such as FUNSD (Jaume et al., 2019), often contain OCR errors. These datasets do not contain sufficient samples for the pre-training of a large model.

Large document datasets. A few larger datasets contain over 100k pages from different domains. However, they usually do not contain OCR annotations or entity-level labels. IIT-CDIP (Lewis et al., 2006) has been the largest dataset commonly used for pre-training of document understanding models. Although these datasets are large, their image quality and annotation completeness are often unsatisfactory. To complement them, we collect high-quality document images from the Internet to build the DocumentNet datasets with rich OCR and entity annotations, and demonstrate their effectiveness in document model pre-training.

Ontology-based datasets. Large labeled datasets are usually collected following an ontology. ImageNet (Deng et al., 2009) for image recognition is built upon the synsets of WordNet (Miller, 1998). ActivityNet (Caba Heilbron et al., 2015) for activity recognition adopts an activity taxonomy with four levels.To the best of our knowledge, DocumentNet is the first large-scale document dataset built upon a well-defined ontology.

Pretrained document models. A variety of pre-trained document models have emerged, including LayoutLM (Xu et al., 2020), UDoc (Gu et al., 2021), LayoutLMv2 (Xu et al., 2021), TILT (Powalski et al., 2021), BROS (Hong et al., 2022), DocFormer (Appalaraju et al., 2021), SelfDoc (Li et al., 2021), LayoutLMv3 (Huang et al., 2022), *etc.* App. B.2 provides detailed comparisons of their designs.

3 DocumentNet Dataset

Blindly crawling the Web for images may seem easy, but it is not a practical solution since most images on the Web are not relevant to document types. We need a scalable pipeline to only select the concerned images. Broadly, this is achievable via a nearest-neighbor search of relevant keywords in a text-image joint embedding space. First, we

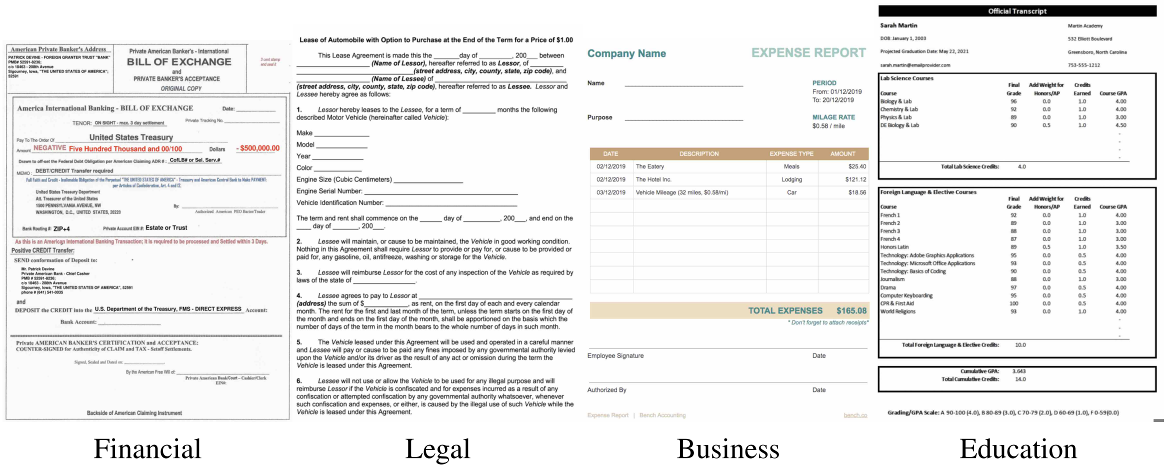

Figure 1: Exemplar documents of each of the four top-level hierarchies. Images are downloaded via keyword searching using a commercial search engine. All images are for demonstration purposes only and do not contain real transactions or personal information.

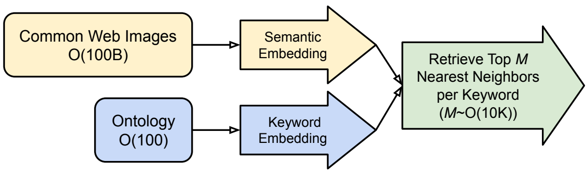

Figure 2: Data Collection Pipeline.

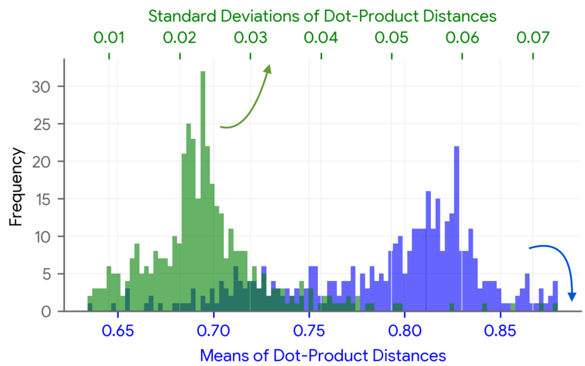

Figure 3: Mean and standard deviation of the dot-product distance between the retrieved 30M document images and each query keyword. A distance of 1.0 indicates the closest semantic relevance.

design a set of query keywords in English, *i.e.*, the document ontology, and encode them into the embedding space of general Web images. Further, a nearest-neighbor algorithm retrieves the top-K semantically closest images to each query keyword. Finally, a deduplication step consolidates all retrieved images across all query keywords. Fig. 1 illustrates several exemplar documents retrieved using our provided keywords.

Ontology creation. Each text string in the ontology list serves as a seed to retrieve the most relevant images from the general Web image pool. An ideal ontology list should therefore cover a broad spectrum of query keywords across and within the concerned downstream application domains. Although algorithmic or generative approaches may exist, in this paper, we manually curated about 400 document-related query keywords that cover domains of finance, business, personal affairs, legal affairs, tax, education, etc. The full ontology hierarchy and keyword list are provided in App. D.

Image retrieval from ontology. To retrieve only the most relevant document images out of the hundreds of billions of general Web images, we leverage a highly efficient nearest neighbor pipeline by defining the similarity metric as the dot product between the semantic feature vectors of the image and each of the target query keywords. Here we refer to Graph-RISE (Timofeev et al., 2020) for the semantic image embedding, and all query keywords are encoded into the same feature space as the images. Empirically, we pick the top 10k nearest neighbors in English for each query keyword. Note that the same image might be retrieved via multiple semantically similar keywords, so a de-duplication step is needed afterward. We summarize the main pipeline steps in Fig. 2. Fig. 3 shows statistical insights of the retrieved 30M document images with the mean and standard deviation histogram over each of the query keywords. The majority of the retrieved images are with mean distance values greater than 0.8 and standard deviations no more than 0.03, indicating high relevance to the document ontology.

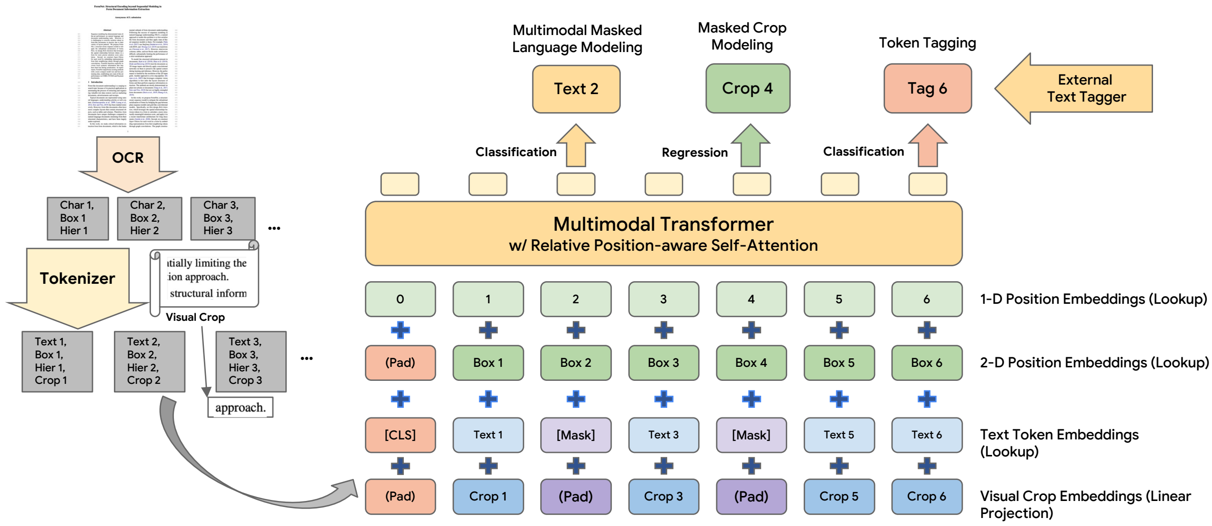

Figure 4: UniFormer pre-training pipeline. The multimodal tokenization process (left) outputs tokens with aligned image crops. The UniFormer model (right) learns a unified token representation with three objectives (top).

	Task	Target Modality
MMLM	Multimodal Masked Language Modeling	OCR characters
MCM	Masked Crop Modeling	Image pixels
TT	Token Tagging	Segment tags

Table 2: UniFormer pre-training objectives and corresponding target modalities.

OCR and annotation. The retrieved images are fed into an OCR engine to generate a text sequence in reading order. We apply a text tagging model to weakly annotate the text segments of each sequence into 6 classes, including *email addresses, mail addresses, prices, dates, phone numbers,* and *person names*. Albeit noisy, these classification labels provide additional supervision for pre-training.

Post-processing and open-source tools. We adopt some heuristic-based filtering to improve sample quality. For example, we remove samples where the overall OCR result is poor due to blurry or noisy images. Some proprietary tools are used for scalable processing during the construction of DocumentNet, but open-source alternatives are readily available. *E.g.*, CLIP (Radford et al., 2021) for text-image embedding, Google ScaNN (Guo et al., 2020) for scalable nearest-neighbor search, Google Cloud OCR (`https://cloud.google.com/vision/docs/ocr`), and Google Cloud NLP (`https://cloud.google.com/natural-language/docs/reference/rest/v1/Entity#type`) for text tagging.

With all of the above steps, we have obtained a dataset of high-quality document images that are closely relevant to our query ontology. This dataset contains multiple modalities, including the image pixels, the OCR characters, the layout coordinates, and the segment tags.

4 UniFormer Model

To take advantage of all the modalities available in DocumentNet, we build a lightweight transformer model named UniFormer for document pre-training. Table 2 lists the pre-training objectives and corresponding target modalities.

UniFormer is built upon the BERT (Devlin et al., 2019) architecture similar to LayoutLM (Xu et al., 2020) and LayoutLMv2 (Xu et al., 2021). Figure 4 illustrates the pre-training pipeline. We highlight the new designs for multimodal pretraining here and defer more details into App. A.

Multimodal tokenization and embedding. With a pre-defined text tokenizer, *e.g.* WordPiece (Wu et al., 2016), we first tokenize the OCR characters into a sequence of text tokens $\mathbf{c}$. For each token c_i, we obtain its bounding box $\mathrm{b}_i = (x_0, y_0, x_1, y_1)_i$ by taking the union of the bounding boxes of its characters. We enlarge the bounding box by a context ratio r on each side and obtain the corresponding visual image crop v_i for each token from the raw image. To model visual information, we add a crop embedding by linearly projecting the flattened pixels in the image crop, following ViT (Dosovitskiy et al., 2020).

Model	Inputs	Pre-training Data	Pre-training Objectives	FUNSD Entity F1↑	CORD Entity F1↑	RVL-CDIP Accuracy↑
BERT	T	-	MLM	60.26	89.68	89.81
LayoutLM	T + L	IIT-CDIP	MVLM	78.66	94.72	91.78
UniFormer	T + L + C	IIT-CDIP	MMLM	80.63	95.17	93.47
UniFormer	T + L + C	IIT-CDIP + *DocumentNet-v1*	MMLM	82.61	95.91	94.86
			MMLM + MCM	83.45	96.08	95.15
			MMLM + MCM + TT	**84.18**	**96.45**	**95.34**

Table 3: Ablation studies on three document understanding benchmarks regarding pretraining datasets, pretraining objectives, and model architectures. Input modalities include text (T), layout (L), and crop (C).

Masked crop modeling. In addition to predicting the text token in the MMLM objective, A UniFormer parameterized by θ also predicts the visual modality by reconstructing the image crops for the masked tokens, in a way similar to MAE (He et al., 2022). It is formulated as a regression problem with a linear layer outputing flattened pixels and the objective is

$$\mathcal{L}_{MCM} = \mathop{\mathbb{E}}_{\text{data}} \Big[\sum_{\mathrm{v}_i \in \mathcal{M}} \| f_\theta(\overline{\mathbf{c}}, \overline{\mathbf{v}}, \rho)_i - \mathrm{v}_i \|_2^2 \Big] \quad (1)$$

where $\overline{\mathbf{c}}$ and $\overline{\mathbf{v}}$ denote the masked tokens and crops according to mask $\mathcal{M}$. ρ is the position and layout embeddings.

Token tagging. With fully unmasked sequences, UniFormer is pre-trained to predict the token tags $\mathbf{t}$ with a separate head. Since each token may have multiple tags, it is formulated as a multi-label classification problem with binary cross-entropy losses.

5 Experiments

We pre-train UniFormer on DocumentNet and evaluate on two settings: (1) the classic VDER setting with the full split of train and test; (2) the few-shot VDER setting where we have meta-train and meta-test task sets with each task containing a set of samples that satisfies the N-way K-shot setting.

5.1 Pre-Training

We initialize our UniFormer with BERT weights using the uncased vocabulary. The models are pre-trained using the Adam optimizer (Kingma and Ba, 2014). We adopt a cosine learning rate schedule with linear warmup during the first 2% steps and a peak learning rate of 10^{-4}. We use 20% of the samples for the token tagging pre-training task. The models are trained for 500K steps with a batch size of 2048 on 128 TPUv3 devices.

5.2 Classic VDER Setting

We evaluate the performance of pre-trained UniFormer models on three commonly used benchmarks: entity extraction on FUNSD and CORD, and document classification on RVL-CDIP. Detailed setups are provided in App. C.1.

Implementation details. For entity extraction on FUNSD and CORD, we add a *Simple* multi-class classification head on top of all text tokens to perform BIO tagging. We fine-tune with a peak learning rate of 5×10^{-5}, following a schedule of linear warm-up in the first 10% steps and then linear decay. Dropout with 0.1 probability is applied in the head layers. UniFormer is fine-tuned for 1000 steps with a batch size of 32 on FUNSD and 256 on CORD. For document classification on RVL-CDIP, we add a multi-class classification head on top of the [CLS] token. We fine-tune with a constant learning rate of 10^{-5} for 15000 steps with a batch size of 2048.

Ablation Studies. Table 3 lists the ablation results for pre-training data, pre-training objectives, and model design. Compared to LayoutLM, our unified embedding of the visual modality and MMLM pre-training results in a much stronger baseline. Adding our DocumentNet into the pre-training leads to a significant performance boost across all three tasks. Further incorporating MCM and TT pre-training objectives to fully leverage DocumentNet yields consistent improvements, where the entity extraction tasks benefit more from TT and the document classification task gains more from MCM.

Comparisons with existing methods. We compare the performance on the three benchmarks with existing approaches at the base model scale in Table 4. As shown, most prior methods use stronger language or image initialization compared to our lightweight UniFormer, but all of them are only

Model	Initialization	Total Parameters	Pretrain Data Source	FUNSD Entity F1↑	CORD Entity F1↑	RVL-CDIP Accuracy↑
LayoutLM	BERT	113M	IIT-CDIP	78.66	94.72	91.78
	BERT + ResNet-101	160M	IIT-CDIP	79.27	-	94.42
UDoc	BERT + ResNet-50	272M	IIT-CDIP	-	-	95.05
LayoutLMv2	UniLM + ResNeXt-101	200M	IIT-CDIP	82.76	94.95	95.25
TILT	T5 + U-Net	230M	RVL-CDIP + UCSF-IDL + CC-PDF	-	95.11	95.25
BROS	BERT	110M	IIT-CDIP	83.05	95.73	-
DocFormer	LayoutLM + ResNet-50	183M	IIT-CDIP	83.34	96.33	96.17
SelfDoc	BERT + ResNeXt-101	137M	RVL-CDIP	83.36	-	92.81
LayoutLMv3*	RoBERTa	126M	IIT-CDIP	-	96.11	95.00
UniFormer	BERT	115M	IIT-CDIP + *DocumentNet-v1*	**84.18**	**96.45**	95.34

Table 4: Comparison with existing document pretraining approaches on three document understanding benchmarks. Models at the base scale are listed for fair comparisons, while state-of-the-art results are obtained by models at larger scales. * denotes a variant that does not use its proprietary tokenizer in pre-training.

Datasets \ Setting / Prediction Head	4-way 2-shot Simple			4-way 2-shot Hierarchical			4-way 4-shot Hierarchical		
	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall
IIT-CDIP	0.099	0.253	0.062	0.108	0.103	0.114	0.115	0.110	0.123
IIT-CDIP + *DocumentNet-v1*	0.102	0.217	0.067	0.121	0.114	0.132	0.129	0.125	0.134
IIT-CDIP + *DocumentNet-v2*	**0.133**	**0.263**	**0.090**	**0.147**	**0.137**	**0.160**	**0.157**	**0.155**	**0.160**

Table 5: Performance comparisons on the few-shot VDER settings with the CORD dataset.

pre-trained on datasets no larger than IIT-CDIP. Although UniFormer is only using 115M parameters and BERT initialization, it outperforms all baseline approaches after pre-training on our DocumentNet dataset, with FUNSD entity F1 84.18, CORD entity F1 96.45, and RVL-CDIP accuracy 95.34.

5.3 Few-shot VDER Setting

We evaluate the performance of pre-trained UniFormer models on N-way K-shot meta-learning settings with the CORD dataset. Detailed task setups are introduced in App. C.2.

Implementation details. In addition to the *Simple* prediction head used in the classic setting, we also adopt a two-level *Hierarchical* prediction head. At the first level, it does a binary classification of the O-tag to identify background tokens. Non-background tokens are further classified by the second level. Hierarchical prediction helps reduce the label imbalance problem where the majority of the tokens are labeled as background. After eliminating a few entities that do not appear frequently enough, we use 18 entities for meta-train and 5 entities for meta-test, for a total of 23 entities. We fine-tune for 15 steps with a constant learning rate of 0.02.

Results. As shown in Tab 5, adding the DocumentNet data significantly boosts the performance of our models across all few-shot learning settings. In particular, the 30M DocumentNet-v2 variant yields a much larger improvement than the 9.9M DocumentNet-v1. The amount of data and the diversity in terms of the collected document type played a significant role in the performance improvements. Performance improvements are universal across each of the metrics, with recall improvements more significant than precision.

6 Conclusions

In this paper, we proposed a method to use massive and noisy web data to benefit the training of VDER models. Our approach has the benefits of providing a large amount of document data with little cost compared to usual data collection processes in the VDER domain. Our experiments demonstrated significantly boosted performance in both the classic and the few-shot learning settings.

7 Limitations

There are a number of areas that would warranty extensions or future work. First, a systematic study on the exact keywords and strategies of collecting such a data that would optimize the model outcome is yet to be studied. The methods proposed in this paper is merely a starting point for methods along this direction. Secondly, architecture changes that specifically targets the proposed methods of massive and noisy data collecting remains an open research question. One observation we had when examining the data is that many of them contains empty forms while others have filled in content. Models that can explicitly take advantage of both formats should further boost the performance of the model.

Acknowledgements

This research was supported in part by the Defence Science and Technology Agency (DSTA).

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv:2305.10403*.

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

Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. 2021. Due: End-to-end document understanding benchmark. In *NeurIPS*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*.

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. In *ICLR*.

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

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *ICML*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*.

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

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

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 *ACM MM*.

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*.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

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*.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *ACM SIGIR*.

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

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Selfdoc: Self-supervised document representation learning. In *CVPR*.

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:1907.11692*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774*.

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 *NeurIPS Workshops*.

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*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: key information extraction datasets involving long documents with complex layouts. In *ICDAR*.

Aleksei Timofeev, Andrew Tomkins, Chun-Ta Lu, Da-Cheng Juan, Futang Peng, Krishnamurthy Viswanathan, Lucy Gao, Sujith Ravi, Tom Duerig, Yi ting Chen, and Zhen Li. 2020. Graph-rise: Graph-regularized image semantic embedding. In *ACM WSDM*.

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

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2022b. A benchmark for structured extractions from complex documents. *arXiv:2211.15421*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. 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 *ACM SIGKDD*.

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

A Details on UniFormer Models

In this section, we detail our UniFormer model architecture and setups for pretraining and fine-tuning for VDER.

A.1 Multimodal Tokenization

Let $\mathbf{D} \in \mathbb{R}^{H \times W \times 3}$ be a visually-rich document image with height H and width W. We obtain a sequence of characters by applying OCR on the document image. The characters are accompanied by their bounding box coordinates. Then we perform a multimodal tokenization process as follows.

With a pre-defined text tokenizer, we first tokenize the character sequence into a sequence of text tokens $\mathbf{c}$. $\mathbf{p}$ represents the 1D position of the tokens ranging from 0 to $|\mathbf{c}| - 1$. For each token c_i, we obtain its bounding box $\mathrm{b}_i = (x_0, y_0, x_1, y_1)_i$ by taking the union of the bounding boxes of its characters. We enlarge the bounding box by a context ratio r on each side and obtain the corresponding visual image crop v_i for each token from $\mathbf{D}$.

A.2 UniFormer Architecture

Fig. 4 illustrates the model architecture for our proposed UniFormer. UniFormer is built upon BERT (Devlin et al., 2019) and utilizes its tokenizer and pretrained weights. The input for each token consists of a text embedding and a 1D position embedding for p.

Following LayoutLM (Xu et al., 2020), we add 2D position embeddings x_0, y_0, x_1, y_1, w, h, where $w = x_1 - x_0$ and $h = y_1 - y_0$. These embeddings are used to represent the spatial location of each token. All the embeddings mentioned above are obtained from trainable lookup tables.

Following LayoutLMv2 (Xu et al., 2021), UniFormer adopts relative position-aware self-attention layers by adding biases to the attention scores according to relative 1D locations $\triangle$p and relative 2D locations $\triangle\frac{x_0+x_1}{2}$, $\triangle\frac{y_0+y_1}{2}$.

Image Crop Input To model visual information, we add a crop embedding by linearly projecting the flattened pixels in the image crop, following ViT (Dosovitskiy et al., 2020). Different from prior works using either uniform patches (Huang et al., 2022), regional features (Li et al., 2021; Gu et al., 2021), or global features (Appalaraju et al., 2021), our multimodal tokenization and linear embedding of image crops has the following advantages:

- It eliminates the separate preprocessing for the visual modality, such as feature extraction with a pretrained CNN (Xu et al., 2021) or manually defined patches (Huang et al., 2022).
- It obtains an aligned partition of the visual information with the text tokens, encouraging better cross-modal interaction.
- It eliminates the need for separate visual tokens as in (Xu et al., 2021; Huang et al., 2022), resulting in a shorter token sequence and better efficiency, as shown in Fig. 5.
- It provides a unified joint representation for text and visual modalities in document modeling with semantic-level granularity.

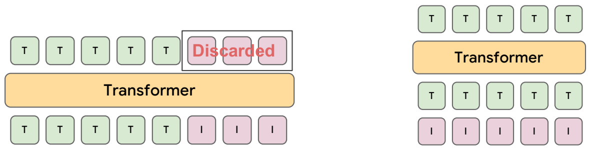

Figure 5: Unaligned (left) vs. Aligned (right) visual features. The unaligned visual features result in a longer sequence but are usually discarded in downstream tasks. T: Text, I: Image.

A.3 Pretraining

During pretraining, we adopt the following objectives on a UniFormer parameterized by θ. For each objective, we use a separate head upon the last attention layer. Let ρ denote the always available input embeddings, including the 1D and 2D positions.

Multimodal Masked Language Modeling (MMLM) We randomly select 15% (Devlin et al., 2019) of the tokens, denoted as $\mathcal{M}$, to mask and predict the language modality. In the masked language input $\overline{\mathbf{c}}$, 80% of the masked tokens are replaced with a special [MASK] token, while another 10% are replaced with a random token and the remaining 10% are kept as is. In the masked crop input $\overline{\mathbf{p}}$, crops for all masked tokens are replaced with an empty image. The language prediction is formulated as a multi-class classification problem with the cross-entropy loss as

$$\mathcal{L}_{MMLM} = \mathop{\mathbb{E}}_{\mathbf{D}} \Big[\sum_{\mathrm{c}_i \in \mathcal{M}} - \log p_\theta(\mathrm{c}_i \mid [\overline{\mathbf{c}}, \overline{\mathbf{v}}, \rho]) \Big] \quad (2)$$

Masked Crop Modeling (MCM) We also predict the visual modality by reconstructing the image crops for the masked tokens in MMLM, in a way similar to MAE (He et al., 2022). It is formulated

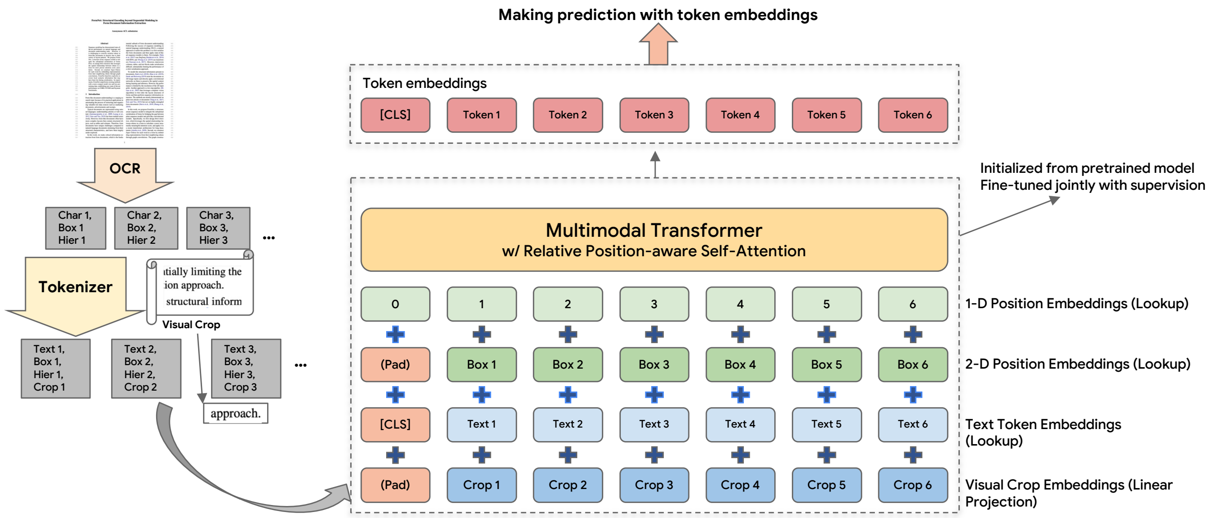

Figure 6: UniFormer finetuning model architecture.

as a regression problem with a linear layer over flattened pixels. The MCM loss is defined as

$$\mathcal{L}_{MCM} = \mathop{\mathbb{E}}_{\mathbf{D}} \Big[\sum_{\mathrm{c}_i \in \mathcal{M}} \|\hat{\mathrm{v}}_i - \mathrm{v}_i\|_2^2 \Big] \tag{3}$$

where $\hat{\mathbf{v}} = f_\theta(\overline{\mathbf{c}}, \overline{\mathbf{v}}, \rho])$.

Token Tagging (TT) We add an extra pretraining task by predicting the tags $\mathbf{t}$ for each token in an unmasked sequence. The tags are extracted from an external text tagger as described in Sec. 3. Since each token may have multiple tags, it is formulated as a multi-label classification problem with the binary cross-entropy loss as

$$\begin{aligned}\mathcal{L}_{TT} = \mathop{\mathbb{E}}_{\mathbf{D}} \Big[\sum_{i,k} &-\mathrm{t}_{i,k} \log p_\theta(\mathrm{t}_{i,k} \mid [\mathbf{c}, \mathbf{v}, \rho]) \\ &- (1 - \mathrm{t}_{i,k}) \log(1 - p_\theta(\mathrm{t}_{i,k} \mid [\mathbf{c}, \mathbf{v}, \rho]))\Big]\end{aligned} \tag{4}$$

where $k = 1, 2, \cdots, K$ refers to the K types of tags.

Pretraining Loss The overall pretraining objective is given as

$$\mathcal{L}_{pretrain} = \mathcal{L}_{MMLM} + \alpha\mathcal{L}_{MCM} + \beta\mathcal{L}_{TT} \tag{5}$$

where α, β are the corresponding loss weights.

A.4 Finetuning

Fig. 6 illustrates the pipeline for the finetuning of UniFormer. During finetuning, no tokens are masked. In this paper, we adopt the following two tasks in finetuning.

Entity Extraction Entity extraction is formulated as a sequence tagging problem. The ground-truth entity spans are converted into a sequence of BIO tags $\mathbf{e}$ over all tokens. The BIO tagging is formulated as follows: $\mathbf{e}$ is initialized with all $\mathcal{O}$ tags which indicates "Other" refering to background tokens. For each entity span with type $\mathcal{T}$, start position i and end position j (both inclusive), we assign

$$\mathrm{e}_i = \mathcal{T}_{\text{Begin}} \tag{6}$$

$$\mathrm{e}_{i+1} = \ldots = \mathrm{e}_j = \mathcal{T}_{\text{Intermediate}} \tag{7}$$

The prediction of BIO tags is modeled as a multi-class classification problem with the objective as

$$\mathcal{L}_{EE} = \mathop{\mathbb{E}}_{\mathbf{D}} \Big[\sum_{i} -\log p_\theta(\mathrm{e}_i \mid [\mathbf{c}, \mathbf{v}, \rho]) \Big] \tag{8}$$

Document Classification We use the embedding of the starting `[CLS]` token for document classification. The logits are predicted with an MLP head on top of the `[CLS]` embedding. Let l be the correct class, the objective is

$$\mathcal{L}_{DC} = \mathop{\mathbb{E}}_{\mathbf{D}} \Big[-\log p_\theta(l \mid [\mathbf{c}, \mathbf{v}, \rho]) \Big] \tag{9}$$

B Additional Related Works

B.1 Datasets

Smaller document datasets The Form Understanding in Noisy Scanned Documents (FUNSD dataset (Jaume et al., 2019), while being the most popular, only contains 199 document pages with three types of entities. The Consolidated Receipt Dataset for Post-OCR Parsing (CORD (Park et al., 2019) dataset comes at a larger scale with 1K document pages and 30 entity types. Other datasets, such as the Scanned Receipts OCR and key Information Extraction (SROIE (Huang et al., 2019), Kleister (Stanisławek et al., 2021) NDA and Charity, DeepForm (Borchmann et al., 2021), VRDU (Wang et al., 2022b) Ad-buy and Registration, have been introduced since then, at the scale of a few thousand documents. Among them, DocVQA (Mathew et al., 2021) contains 12.8K documents with question-answer annotations.

Larger document datasets IIT-CDIP (Lewis et al., 2006) consists of 11M unlabeled documents with more than 39M pages. PDF files from Common Crawl (CC-PDF) and UCSF Industry Documents Library (UCSF-IDL) have also been used for pretraining (Powalski et al., 2021), with a total of less than 1M documents. RVL-CDIP, a subset of IIT-CDIP, contains 400K documents categorized into 16 classes for the document classification task. PubLayNet (Zhong et al., 2019) is at a similar scale but for the layout detection task with bounding box and segmentation annotations.

B.2 Document Understanding Models

Document understanding models have emerged since LayoutLM (Xu et al., 2020), which extends BERT (Devlin et al., 2019) with spatial and visual information. Various models use different initialization weights, model scales, and pretraining data configurations. Table 4 provides a detailed comparison of existing models.

Text Modality. Document models are usually built upon a pretrained language model. As shown by LayoutLM (Xu et al., 2020), language initialization significantly impacts the final model performance. Many works have been built upon the standard BERT language model, such as LayoutLM (Xu et al., 2020), BROS (Hong et al., 2022), SelfDoc (Li et al., 2021), and UDoc (Gu et al., 2021). LayoutLMv2 (Xu et al., 2021) is initialized from the UniLM (Dong et al., 2019). TILT (Powalski et al., 2021) extends T5 (Raffel et al., 2020) for document analysis. DocFormer (Appalaraju et al., 2021) directly initializes from a pretrained LayoutLM. The recent LiLT (Wang et al., 2022a) and LayoutLMv3 (Huang et al., 2022) models are initialized from RoBERTa (Liu et al., 2019) to provide a stronger language prior. In our experiments, we adopt the vanilla BERT-base model for fair comparisons without the benefit of a stronger language model.

	Precision	Recall	F1-score	Support
Question	84.84	88.41	86.59	1070
Header	57.26	56.30	56.78	119
Answer	82.67	87.27	84.91	809
Average	82.41	86.04	**84.18**	1998

Table 6: Detailed metrics on the FUNSD entity extraction task.

Visual Modality. Existing document models rely on pretrained image models to utilize the document images. LayoutLM (Xu et al., 2020) adopts a pretrained ResNet-101 (He et al., 2016) as the visual feature encoder only during finetuning. LayoutLMv2 (Xu et al., 2021) further utilizes a ResNeXt-101 (Xie et al., 2017) at both pretraining and finetuning with encoded patch features as visual tokens. In addition, SelfDoc (Li et al., 2021), UDoc (Gu et al., 2021), TILT (Powalski et al., 2021), and DocFormer (Appalaraju et al., 2021) also adopt a pretrained ResNet (He et al., 2016) as the visual feature encoder. LayoutLMv3 (Huang et al., 2022) distills a pretrained document image dVAE (Ramesh et al., 2021) from DiT (Li et al., 2022) to learn the visual modality during pretraining. In contrast, we do not use pretrained image models but learn a joint vision-language representation by aligning both modalities at the token level.

C Detailed Experimental Setups and Analysis

C.1 Classic VDER Setting

Task setup. FUNSD contains 199 documents with 149 for training and 49 for evaluation. It is labeled with 3 entity types, i.e., header, question, and answer. CORD contains 1000 documents with 800 for training, 100 for validation, and 100 for testing. It is labeled with 30 entity types for receipts, such as menu name, price, *etc*. RVL-CDIP contains 400K documents in 16 classes, with 320K for training, 40K for validation, and 40K for testing.

Error analysis. Table 6 lists the detailed metrics on the FUNSD entity extraction task. Among the three labeled entity types, *header* has the poorest performance and the lowest number of examples. The other two types have much better performance with F1 86.59 for *question* and F1 84.91 for *answer*. Fig. 7 visualizes a few examples with annotations

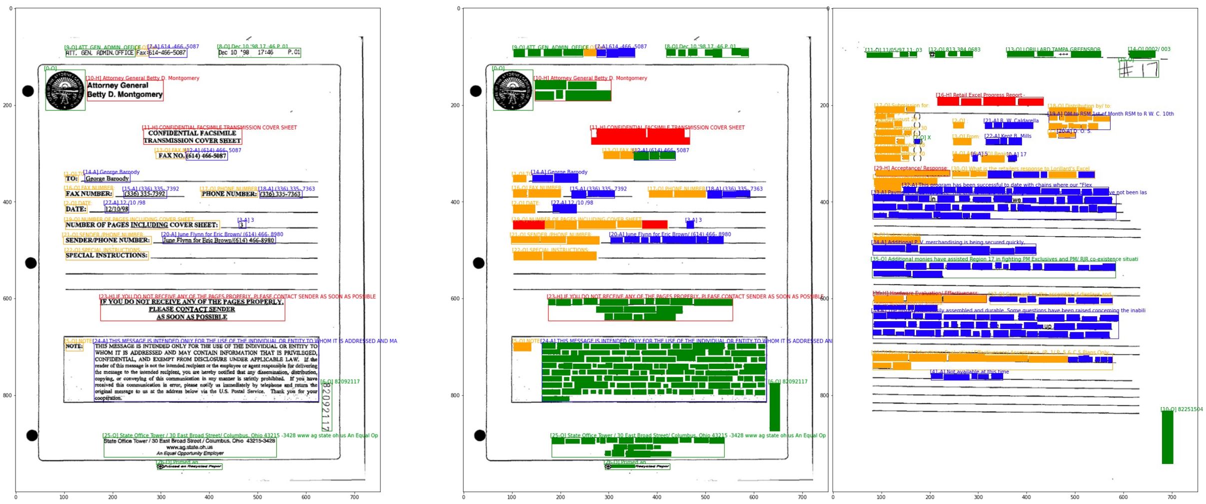

Figure 7: Visualization of annotation (left) and prediction examples (middle and right) from the FUNSD validation set. Zoom in for details.

and predictions from our UniFormer. As we can see in the annotation, the reading order is often weird and does not follow human conventions. However, the 2D positional embedding and spatial-aware attention can correctly handle them regardlessly. In the prediction samples, we observe that the predictions for *question* and *answer* fields are mostly correct, while a few errors are made for *header* due to ambiguity.

C.2 Few-shot VDER Setting

N-way K-shot meta-learning formulation. In our setting, we define a N-way K-shot problem to be one such that there are N novel classes that appear no more than K times in the training set. We then divide a dataset into several sub-groups with each of them satisfying the N-way K-shot definition. One unique characteristic on the VDER dataset is that documents usually contain multiple entities, with many of the entities occur more than once in a single document, we make the requirements on the number of occurrence K to be a soft one so that it would be realistic to generate such a dataset splitting. The few-shot learning problem will natually fit into a meta-learning scenario, meta-train and meta-test both contain a set of tasks satisfying N-way K-shot setting.

We sample datasets to achieve n-way, k-shot settings, which means that our training data contains n entities, each with at least k occurrences. The count of classes in testing is fixed at 5. For hyper-parameters, we follow most of the settings for classic VDER experiments. We fine tune with a learning rate of 0.02.

D DocumentNet Ontology

Fig. 8 illustrates the document ontology tree stub used for the construction of DocumentNet. Below we list all of the search keywords organized into four groups.

D.1 Financial Documents

- accounts receivable aging report
- bill of exchange pdf
- invoice
- receipt
- loan estimate
- loan application form
- credit report pdf
- employee insurance enrollment form
- property insurance declaration page
- renters insurance addendum
- auto insurance card
- dental insurance card
- dental insurance verification form
- vision insurance card
- medical insurance card
- liability insurance certificate
- insurance cancellation letter
- life insurance application form
- flood elevation certificate
- flood insurance application form
- hazard insurance application form
- tax return form
- form 1040 schedule C
- form 1040 schedule E
- form 1040 schedule D
- form 1040 schedule B
- form 1040 nr

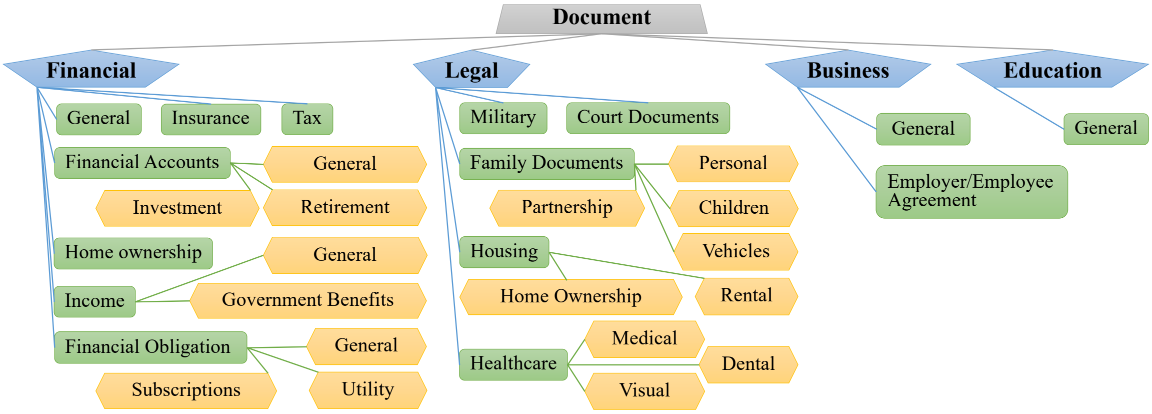

Figure 8: Document ontology tree stub, based on which the proposed DocumentNet datasets are collected. We create a document ontology with about 400 search keywords hierarchically connected by three intermediate layers.

- form 1040 sr
- form 4506T EZ
- form 4506T
- form 4506 C
- transfer of residence form 1076
- property tax bill
- W2
- W4
- 1099 B
- 1099-MISC
- 1099-NEC
- 1099 DIV
- 1099 G PDF
- 1099 R
- 1099 INT
- SSA 1099 form
- 1120 form
- 1120S Form
- form 1065
- W7 form
- W8BEN form
- W9 form
- SS4 form
- form 940 pdf
- form 5498
- ucc 1 form
- bank statement
- personal check
- check deposit slip pdf
- credit union statement pdf
- credit card authorization form
- credit card
- debit card
- credit card statement
- TSP election form
- 401k enrollment form
- IRA distribution request form
- stock certificate
- stock purchase agreement
- bond certificate
- bond purchase agreement
- mutual fund consolidated account statement
- HSA enrollment form
- FSA enrollment form
- verification of employment pdf
- wage paystub
- income verification letter
- music recording contract
- food stamp application form
- us treasury check
- child welfare services application form
- medicaid card
- medicaid application form
- club application
- membership renewal letter pdf
- mortgage statement
- rent invoice
- electric bill
- pg&e care fera application pdf
- gas bill
- water bill
- waste management invoice
- spectrum internet bill pdf
- phone bill pdf
- car payment agreement
- student loan payment agreement
- child support agreement
- child support receipt
- elder care facility agreement
- debt paymen tletter
- demand for payment letter
- magazine subscription form
- streaming service agreement
- gym waiver form

- gym membership cancellation letter
- gym membership card
- massage therapy waiver
- HOA agreement
- HOA dues letter
- urla form 1003
- home appraisal report
- security instrument
- ucdp summary report
- audit findings report pdf
- sales contract
- purchase agreement
- title commitment pdf
- earnest money deposit pdf
- patriot act disclosure
- owner occupancy affidavit form
- compliance agreement
- name affadavit
- notice of right to reclaim abandoned property
- VBA 26-0551 debt questionnaire pdf
- VBA 26-8923 form pdf
- USDA-AD 3030
- loan application pdf
- homeowner insurance declaration page
- 1040
- wage and tax statement
- employee's withholding certificate
- miscellaneous income form
- nonemployee compensation form
- dividends and distributions form
- certain government payments
- distributions from pensions
- social security benefits form
- form 1005
- stimulus check pdf
- waste management bill
- comcast internet bill pdf
- car loan payment agreement
- gym release form
- one and the same person affadavit
- xfinity internet bill pdf
- car payment contract

D.2 Legal Documents

- birth certificate
- social security card
- social security form
- ssa 89 form
- social security change in information form
- passport book
- passport card
- new passport application
- passport renewal application
- green card
- green card application form
- naturalization certificate
- N-400 form pdf
- living will sample
- living will form
- living will declaration
- voter identification card
- disability card
- death certificate
- death certificate application
- name change form
- state issued identification card
- prenup form
- postnuptial agreement
- marriage license
- marriage certificate
- application for marriage license
- family court cover sheet
- complaint for divorce no children
- complaint for divorce with children
- divorce summons
- divorce certificate
- domestic partnership application form
- domestic partnership certificate
- domestic partnership termination form
- separation agreement
- pet custody agreement form
- pet ownership transfer form
- child adoption certificate
- child power of attorney
- child visitation form
- daycare contract
- child custody agreement
- child support modification form
- free minor travel consent form
- child identity card
- DNA paternity test order form
- petition for declaration of emancipation of minor
- vehicle registration card
- vehicle registration form
- vehicle registration renewal notice
- vehicle certificate of title
- motor vehicle transfer form
- driver's license
- application for driver's license
- truck driver application
- learner's permit card

- pilot's license card
- vehicle leasing agreement
- motor vehicle power of attorney
- mortgage interest credit form
- mortgage application form
- mortgage verification form
- mortgage loan modification form
- real estate deed of trust
- mortgage deed
- warranty deed
- quitclaim deed
- usps mail forwarding form PDF
- property power of attorney
- notice of intent to foreclose
- closing disclosure
- HUD 92541 form
- HUD 54114 form
- HUD 92561 form
- FHA loan underwriting and transmittal summary
- form HUD92051
- form HUD 92900-A
- form HUD 92544
- form HUD 92900 B important notice to housebuyers
- form HUD 92900 WS mortgage credit analysis worksheet
- Form HUD 92800 Conditional Commitment
- SFHDF
- lease agreement
- lease application
- notice to enter
- notice of intent to vacate premises
- notice of lease violation
- pay rent or quit
- lease offer letter
- roommate agreement
- eviction notice form
- lease termination letter
- lease renewal agreement
- pet addendum
- notice of rent increase
- sublease agreement
- record of immunization
- allergy record sheet
- allergy immunotherapy record
- medication log
- prescription sheet
- disability documentation
- advance directive form
- DNA test request form
- medical power of attorney
- health care proxy form
- revocation of power of attorney
- dnr form
- hipaa release form
- hipaa complaint form
- health history form
- birth plan form
- new patient form
- child medical consent
- grandparent medical consent for minor
- medical treatment authorization form
- dental policy and procedure document
- endodontic treatment consent form
- denture treatment consent form
- dental patient referral form
- patient dismissal letter
- dental record release form
- oral surgery postop instructions
- refusal of dental treatment form
- tooth extraction consent form
- corrective lens prescription pdf
- military id card
- dd214
- honorable discharge certificate
- supreme court distribution schedule pdf
- case docket pdf
- jury summons pdf
- jury duty excuse letter
- supplemental juror information pdf
- attorney termination letter
- certificate of good standing
- attorney oath of admission pdf
- substitution of attorney
- notice of appearance of counsel
- bankruptcy declaration form
- notice of lawsuit letter
- court summons pdf
- arrest warrant pdf
- promissory note
- tolling agreement pdf
- notary acknowledgement form
- cease and desist letter
- condominium rider pdf
- adjustable rate rider pdf
- family rider form 1-4 pdf
- balloon rider form pdf
- second home rider pdf
- revocable trust rider form pdf
- pud rider pdf
- birth certificate form

- ssn card
- application for a social security card
- application for naturalization pdf
- legal name change form
- state issued ID
- prenup sample
- postnuptial agreement sample
- declaration of domestic partnership
- marriage separation agreement
- minor power of attorney
- request for child custody form
- child care contract
- motion to adjust child support
- child travel consent form
- vehicle registration application
- vehicle certificate of title
- driver's license application
- mortgage loan application form
- real estate power of attorney
- foreclosure letter notice
- rental agreement
- rental application
- landlord notice to enter
- intent to vacate rental
- notice to pay rent or quit
- roommate contract
- eviction notice pdf
- early lease termination letter
- pet addendum to lease agreement
- rent increase letter
- vaccine record form
- prescription sample
- healthcare directive form
- do not resuscitate form
- medical records release form
- medical history form
- new patient registration form
- child medical release form
- root canal consent form
- eyeglasses prescription pdf
- military discharge form
- notice of intent to sue
- ssn application
- name change form example
- prenuptial agreement sample
- marital separation form
- motion to modify child support
- tenant application
- notice to enter premises
- notice to vacate
- notice to quit
- notice of lease termination
- doctor prescription
- patient history form
- patient intake form
- consent to treat minor
- form petition for name change
- medical intake form

D.3 Business Documents

- articles of incorporation
- corporate bylaws
- operating agreement
- shareholder agreement
- memorandum of understanding
- expense report
- purchase of business agreement
- purchase order
- invoice pdf
- late payment reminder letter
- arbitration agreement pdf
- business contract
- payment agreement document
- end user license agreement
- licensing agreement pdf
- job application form
- employment offer letter
- employment rejection letter
- employment agreement
- resume
- employment resignation letter
- notice of contract termination
- notice of employmen termination
- nda pdf
- non compete agreement
- leave of absence request
- employment evaluation form
- overdue payment reminder letter
- job application pdf
- job offer letter
- job rejection letter
- employment contract
- contract termination letter
- non disclosure agreement

D.4 Education Documents

- research papers pdf
- certificate of enrollment
- high school transcript
- high school diploma
- college diploma
- college transcript