

Contrastive Training Improves Zero-Shot Classification of Semi-structured Documents

Muhammad Khalifa^{1*}, Yogarshi Vyas^{2†}, Shuai Wang²,
Graham Horwood², Sunil Mallya, Miguel Ballesteros²

¹University of Michigan, ²AWS AI Labs

khalifam@umich.edu,

{yogarshi, wshui, ghorwood, ballemig}@amazon.com,

mallya16@gmail.com

Abstract

We investigate semi-structured document classification in a zero-shot setting. Classification of semi-structured documents is more challenging than that of standard unstructured documents, as positional, layout, and style information play a vital role in interpreting such documents. The standard classification setting where categories are fixed during both training and testing falls short in dynamic environments where new document categories could potentially emerge. We focus exclusively on the zero-shot setting where inference is done on new unseen classes. To address this task, we propose a matching-based approach that relies on a pairwise contrastive objective for both pretraining and fine-tuning. Our results show a significant boost in Macro F₁ from the proposed pretraining step in both supervised and unsupervised zero-shot settings.

1 Introduction

Textual information assumes many forms ranging from *unstructured* (e.g., text messages) to *semi-structured* (e.g., forms, invoices, letters), all the way to fully structured (e.g., databases or spreadsheets). Our focus in this work is classification of semi-structured documents. A semi-structured document consists of information that is organized using a regular visual layout, and includes tables, forms, multi-columns, (nested) bulleted lists, and that is either understandable only in the context of its visual layout or that requires substantial more work to understand without the visual layout. Automatic processing of semi-structured documents comes with a unique set of challenges including a non-linear text flow (Wang et al., 2021), layout inconsistencies, and low-accuracy optical character recognition. Prior work has shown that integrating the two-dimensional layout information

of such documents is critical in models for analyzing such documents (Xu et al., 2020, 2021; Huang et al., 2022; Appalaraju et al., 2021). Due to these challenges, methods for unstructured document classification, such as static word vectors (Socher et al., 2013) and standard pretrained language models (Devlin et al., 2019; Reimers and Gurevych, 2019; Liu et al., 2019) perform poorly with semi-structured inputs as they model text in a one-dimensional space and ignore information about document layout and style (Xu et al., 2020).

Past work on semi-structured document classification (Harley et al., 2015; Iwana et al., 2016; Tensmeyer and Martinez, 2017; Xu et al., 2020, 2021) has focused exclusively on the *full-shot* setting, where the target classes are fixed and identical across training and inference, neglecting the *zero-shot* setting (Xian et al., 2018), which requires generalization to unseen classes during inference.

Our work addresses zero-shot classification of semi-structured documents in English using the matching framework, which has been used for many tasks on unstructured text (Dauphin et al., 2014; Nam et al., 2016; Pappas and Henderson, 2019; Vyas and Ballesteros, 2021; Ma et al., 2022). Under this framework, a matching (similarity) metric between documents and their assigned classes is maximized in a joint embedding space. We extend this matching framework with two enhancements. First, we use a pairwise contrastive objective (Rethmeier and Augenstein, 2020; Radford et al., 2021; Gunel et al., 2021) that increases the similarity between documents and their ground-truth labels, and decreases it for incorrect pairs of documents and labels. We augment the textual representations of documents with layout features representing the positions of tokens on the page to capture the two-dimensional nature of the documents. Second, we propose an unsupervised contrastive pretraining procedure to warm up the representations of documents and classes. In summary, (i) we study

*Work done while at AWS AI Labs

†Corresponding author

zero-shot classification of semi-structured documents, which, to the best of our knowledge, has not been explored before. **(ii)** we use a pairwise contrastive objective to both pretrain and fine-tune a matching model for the task. This technique uses a layout-aware document encoder and a regular text encoder to maximize the similarity between documents and their ground-truth labels. **(iii)** Using this contrastive objective, we propose an unsupervised pretraining step with pseudo-labels (Rethmeier and Augenstein, 2020) to initialize document and label encoders. The proposed pretraining step improves F1 scores by 9 and 19 points in supervised and unsupervised zero-shot settings respectively, compared to a setup without this pretraining.

2 Approach

This section describes our proposed architecture (§ 2.1), pretrained model (§ 2.2), as well as the contrastive objective used for pretraining (§ 2.3) and fine-tuning (§ 2.4).

2.1 Model

Our goal is to learn a matching function between documents and labels such that similarity between a document and its gold label is maximized compared to other labels, which can be seen as an instance of metric learning (Xing et al., 2002; Kulis et al., 2012; Sohn, 2016). This requires encoding documents and class names¹ into a joint document-label space (Ba et al., 2015; Zhou et al., 2019; Chen et al., 2020; Hou et al., 2020). In this work, documents and class names are of different nature—documents are semi-structured (§ 1), while class names are one or two-word fragments of text.

We use two encoders to account for this difference: a document encoder Φ_{doc} suitable for semi-structured documents, and a label (class) encoder Φ_{label} suitable for the natural language representations of the class labels. Φ_{label} is simply a vanilla pretrained BERT_{BASE} model (Devlin et al., 2019). Φ_{doc} , as in prior work (Xu et al., 2020; Lockard et al., 2020), is a pretrained language model that encodes the text and the layout of the document using coordinates of each token. The next section explains this model, Layout_{BERT}, in detail. We choose this model for its simplicity, but our proposed approach can be combined with more sophisticated

¹We use class names as the natural language representation of a class, but more descriptive representations can be used if available (e.g. dictionary definitions) (Logeswaran et al., 2019)

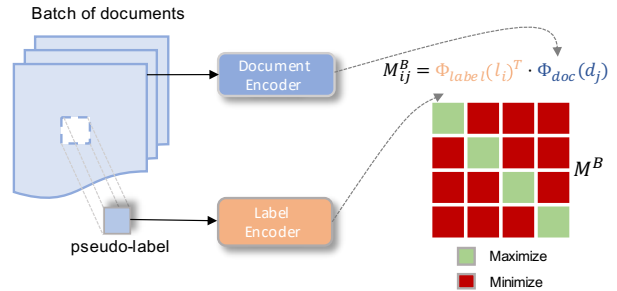


Figure 1: The unsupervised contrastive pretraining procedure. A random block of tokens from a document is used as the pseudo-label for that document. Dot products between documents and their labels are maximized and all other pairwise dot products are minimized.

document encoders that incorporate layout and visual information in different ways (Huang et al., 2022; Xu et al., 2021; Appalaraju et al., 2021).

2.2 Layout_{BERT}

Layout_{BERT} is a 6-layer Transformer based on BERT_{BASE} (Devlin et al., 2019) and is pretrained using masked language modeling on a large collection of semi-structured documents (§ 3). Unlike prior work, Layout_{BERT} has a simpler architecture that decreases model footprint while maintaining accuracy. Specifically, there are three main architectural differences between Layout_{BERT} and LayoutLM, which is the most comparable architecture in the literature (Xu et al., 2020): **(a)** LayoutLM uses 12 transformer layers while Layout_{BERT} uses only 6 layers **(b)** LayoutLM uses four positions per token, namely upper-left and bottom-right coordinates, while Layout_{BERT} use only two positions viz. the centroid of the token bounding box. **(c)** Unlike LayoutLM, Layout_{BERT} does not use an image encoder to obtain CNN-based visual features.²

2.3 Unsupervised Contrastive Pretraining

Φ_{label} and Φ_{doc} are models that have been pretrained independently. To encourage these models to produce similar representations for documents and their labels, we continue pretraining Φ_{label} and Φ_{doc} via an unsupervised procedure based on a pairwise contrastive objective. The unsupervised objective can learn from large amounts of unlabeled semi-structured documents. This also allows us to directly use the pretrained encoders in an unsupervised zero-shot setting (§ 3.3.1).

²The results in Xu et al. (2020) show that image features are not always useful. To keep things simple, we do not include the CNN component in our model.

Since we do not assume access to ground truth labels for this step, our pretraining procedure relies solely on self-supervision via *pseudo-labels* (Rethmeier and Augenstein, 2020). These pseudo-labels are generated by sampling a continuous block of tokens from the document with a length drawn from a shifted geometric distribution. A pseudo-label extracted from a document is treated as the positive label for that document and is encoded using Φ_{label} .

We now describe our contrastive objective which is based on the multi-class n-pair loss (Sohn, 2016; Radford et al., 2021). Let B be a training batch that consists of training documents D and their pseudo-labels L , such that $D = (d_1, d_2, \dots, d_{|B|})$ and $L = (l_1, l_2, \dots, l_{|B|})$. Let Φ_{doc} and Φ_{label} be the document and label encoders, respectively. We start by encoding each document and pseudo-label in the batch and then computing a matching matrix $M^B \in \mathbb{R}^{|B| \times |B|}$ of pairwise dot products between every document-label pair, such that $M_{ij}^B = \Phi_{label}(l_i)^T \cdot \Phi_{doc}(d_j)$. Our objective is to increase the value of diagonal elements M_{ij} , where $i = j$, as compared to all other elements. More precisely, the loss function for a batch is a symmetric loss, \mathcal{L}^B , that can be expressed with the equation:

$$\mathcal{L}^B = \frac{1}{2}[\mathcal{L}_{row}^B + \mathcal{L}_{col}^B]. \quad (1)$$

Where \mathcal{L}_{row}^B and \mathcal{L}_{col}^B are the per-batch row-wise and column-wise losses, respectively, with

$$\mathcal{L}_{row}^B = \sum_{i=1}^{|B|} \left[-\log(\exp(M_{ii}^B)) + \log\left(\sum_{j=1}^{|B|} \exp(M_{ij}^B)\right) \right]. \quad (2)$$

The first term in Eq. 2 maximizes the diagonal elements, while the second term minimizes the off-diagonal elements. The column-wise loss is the same with i and j swapped. We directly optimize the raw dot products rather than cosine similarity as we observed dot-products to perform much better, which also agrees with Karpukhin et al. (2020).

2.4 Contrastive Fine-tuning

For the supervised zero-shot setting (§ 3.3.2), we fine-tune the model using the same objective as the pretraining step (Equation 1), except that the labels $L = (l_1, l_2, \dots, l_{|B|})$ for a batch B are ground-truth labels and not pseudo-labels.

3 Experiments and Results

3.1 Data

We evaluate our approach on the RVL-CDIP dataset (Harley et al., 2015), which consists of 400K documents balanced across 16 classes such as letter, advertisement, scientific report, form, etc. Since zero-shot performance can vary depending on which classes are used for train and test, we follow previous work (Ye et al., 2020) and create four zero-shot splits of the data with non-overlapping test classes. Thus, each split has 8 training classes (200K documents), 4 validation classes (100K documents), and 4 test classes (100K documents).³

Our document encoder is pretrained on documents from CommonCrawl (see Appendix B for more details).⁴ While this pretraining corpus is different from the one used for LayoutLM, our objective is not to compare directly with this model but to explore zero-shot classification. Our contrastive pretraining corpus consists of 800K documents sampled from this pretraining corpus. We first sample $l \sim \text{Geometric}(\frac{1}{20})$, and then sample a block of l tokens from each document to obtain a pseudo-label for that document. We run contrastive pretraining for 50K steps with batch size of 256.

3.2 Experimental Setup

Layout_{BERT} is a 6-layer model initialized using BERT_{BASE} weights and further pretrained using the MLM loss with layout information for 50K steps with a batch size of 2048 and a peak learning rate of 10^{-4} . Unlike LayoutLM, where the extra position embeddings are initialized from scratch, we initialize them from BERT positional embeddings, which we found to speed up convergence. We used dynamic subtoken masking (Liu et al., 2019) with $p_{mask} = 0.15$ and $p_{replace} = 0.80$.

The representation of the [CLS] token is used as the encoding of input documents and an affine layer with a dimension of 768 is applied to the output of both encoders. We fine-tune the matching model on the data from the train classes for 30 epochs with a batch size of 40 and a learning rate of 3×10^{-5} . The model with the best macro F_1 on the validation set is used for evaluation on the held out test set.

3.3 Results

We experiment with two settings — unsupervised zero-shot, and supervised zero-shot. In the former,

³The exact classes used for each split are in Appendix A.
⁴<https://commoncrawl.org/>

Method	I		II		III		IV		Avg.
	Valid	Test	Valid	Test	Valid	Test	Valid	Test	
BERT (doc and label)	12.05	10.64	13.77	14.08	10.89	13.28	13.94	12.25	12.61
Layout _{BERT} (doc), BERT (label)	12.05	30.64	16.77	22.04	31.11	17.32	21.75	12.04	20.47
CPT, Layout _{BERT} (doc), BERT (label)	50.5	21.25	24.60	61.36	21.65	24.58	61.50	51.57	39.63

Table 1: Unsupervised zero-shot performance (Macro F_1) on 4 splits of RVL-CDIP. **CPT**: Contrastive Pretraining.

Method	I		II		III		IV		Avg.
	Valid	Test	Valid	Test	Valid	Test	Valid	Test	
Cross-entropy FT	34.76	25.33	35.64	23.29	11.67	28.84	29.68	36.75	28.76
Contrastive FT	37.35	25.76	32.55	26.05	18.14	27.63	29.86	32.74	28.25
CPT + Standard FT	48.24	26.97	30.45	37.81	27.20	28.11	48.82	46.09	36.71
CPT + Contrastive FT	49.68	25.82	30.31	44.44	20.80	30.43	51.26	45.07	37.23

Table 2: Supervised zero-shot performance (Macro F_1) on 4 splits of RVL-CDIP and with two different finetuning objectives. **FT**: Finetuning using standard cross-entropy or contrastive losses. **CPT**: Contrastive Pretraining. Performance is averaged across 3 runs with different seeds.

no fine-tuning is involved and all models are directly used for inference. In the latter, all models are fine-tuned on data from classes different than those present in the test set. Thus, the latter is strictly more challenging.

3.3.1 Unsupervised Zero-shot

We start with the more challenging unsupervised setup and compare three models (Table 1). The first model uses a vanilla pretrained BERT_{BASE} as both the document and label encoders. The second model replaces the BERT_{BASE} document encoder with Layout_{BERT} model. For these two models, we remove the affine layer after both encoders (§ 3.2) since in the absence of pretraining/finetuning, they will not be trained. The third model uses the same components as the second model but is pretrained using the unsupervised contrastive loss (§ 2.3).

The results yield three key observations. First, the vanilla BERT model performs the worst with an F_1 score of 13. This is unsurprising as BERT does not capture any layout information. Second, the value of layout information can be verified by replacing the BERT_{BASE} document encoder with Layout_{BERT}. This improves the average F_1 by ~ 8 points. Finally, contrastive pretraining (CPT) is critical to produce better initialization for the encoders and it improves the average performance of the previous model by ~ 19 F_1 points.

3.3.2 Supervised zero-shot

Next, we turn to the supervised zero-shot setup, where models are finetuned on data from classes different than those in the test set. We only experiment with the Layout_{BERT} (doc), BERT (label) setup since it performed the best in unsupervised

settings. Table 2 shows the Macro F_1 with our in-batch contrastive training objective as well as a standard cross-entropy loss (Dauphin et al., 2014; Ye et al., 2020). We also show the fine-tuning performance with contrastive pretraining (§ 2.3).

We observe that the in-batch contrastive objective yields comparable F_1 to the cross-entropy loss on average (with and without pretraining). However, the in-batch loss also has higher variance across different runs compared to the cross-entropy loss,⁵ possibly due to the stochastic nature of in-batch contrastive training. Crucially, though, we observe a strong F_1 boost in almost all cases with contrastive pretraining, and in some cases as much as ~ 21 F_1 points. This reemphasizes the importance of pretraining in producing similar representations for related documents and labels.

Finally, comparing Tables 1 and 2 shows that the zero shot performance is better in the unsupervised case than the supervised case. This is likely due to the fact that in the latter, the model is fine-tuned towards a specific type of documents (i.e. those present in the training/validation) classes, which hinders generalization to unseen inference classes. More sophisticated approaches (Finn et al., 2017; Nichol et al., 2018) can potentially improved the supervised setup, but we leave this to future work.

4 Conclusion

This work explores zero-shot classification of semi-structured documents. We proposed two contrastive techniques for pretraining and fine-tuning

⁵Tables 4 and 5 in Appendix C show means and standard deviations with three random seeds. Experiments with more random seeds did not yield any meaningful differences.

of a matching model. Our fine-tuning objective showed comparable results to the standard cross-entropy loss used widely in the literature and our contrastive pretraining significantly boosted zero-shot F_1 in supervised and unsupervised scenarios.

5 Limitations

The current work is an initial attempt at studying the problem of zero-shot classification of semi-structured documents. There are two key aspects that this work does not cover and we encourage future work to explore.

First, as pointed out in §2.1, we choose Layout_{BERT} as our document encoder, Φ_{doc} . This work does not experiment with the variety of encoding strategies in the literature that combines textual, visual, and layout information (Appalaraju et al., 2021; Xu et al., 2021; Huang et al., 2022). It is likely that richer document representations derived from these diverse encoders will further push the limits of zero-shot classification when combined with our proposed unsupervised contrastive pre-training procedure.

Second, results in this paper are on a single dataset, i.e. the RVL-CDIP dataset. While we mitigate this to a large extent by creating four non-overlapping test splits (see §3.1 and Appendix A), results on more datasets might yield more useful insights. In practice, the lack of datasets for this task (of semi-structured document classification) is what makes this exploration difficult and might require creation of new resources

Acknowledgements

We thank our colleagues at AWS AI Labs and the ACL reviewers who have helped improve the paper through their feedback.

References

- Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 993–1003.
- Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions.
- Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, and Benjamin Van Durme. 2020. Reading the manual: Event extraction as definition comprehension.
- Yann N. Dauphin, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2014. Zero-shot learning and clustering for semantic utterance classification. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.
- Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.
- Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.
- Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393, Online. Association for Computational Linguistics.
- Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking.
- Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2016. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

- Brian Kulis et al. 2012. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364.
- 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](#). *CoRR*, abs/1907.11692.
- Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi. 2020. [ZeroShotCeres: Zero-shot relation extraction from semi-structured web-pages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8105–8117, Online. Association for Computational Linguistics.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. [Zero-shot entity linking by reading entity descriptions](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.
- Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. [Label semantics for few shot named entity recognition](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971, Dublin, Ireland. Association for Computational Linguistics.
- Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. All-in text: Learning document, label, and word representations jointly. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Alex Nichol, Joshua Achiam, and John Schulman. 2018. [On first-order meta-learning algorithms](#).
- Nikolaos Pappas and James Henderson. 2019. Gile: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.
- 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. *arXiv preprint arXiv:2103.00020*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Nils Rethmeier and Isabelle Augenstein. 2020. Data-efficient pretraining via contrastive self-supervision. *arXiv preprint arXiv:2010.01061*.
- Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. [Zero-shot learning through cross-modal transfer](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 935–943.
- Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865.
- Chris Tensmeyer and Tony Martinez. 2017. Analysis of convolutional neural networks for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 388–393. IEEE.
- Yogarshi Vyas and Miguel Ballesteros. 2021. [Linking entities to unseen knowledge bases with arbitrary schemas](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 834–844, Online. Association for Computational Linguistics.
- Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. [LayoutReader: Pre-training of text and layout for reading order detection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.
- Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. 2002. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15:521–528.
- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. [LayoutLMv2: Multi-modal pre-training for visually-rich document understanding](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.
- 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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.
- Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang,

and Huajun Chen. 2020. Zero-shot text classification via reinforced self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3014–3024.

Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2019. *Zero-shot open entity typing as type-compatible grounding*.

A Data Splits

As stated in section 3, we split the RVL-CDIP dataset into four splits with non-overlapping test classes. Table 3 shows the classes used in each split.

B Pre-training data from Common Crawl

We build our pre-training corpus by first extracting all documents from CommonCrawl with a `.pdf` extension. We then remove duplicate documents based on the MD5 hash using `fdupes`.⁶ The resulting documents are then passed through `PDF-PLUMBER`⁷ to extract both the text as well as the co-ordinates of the tokens in the documents, and any documents that cannot be processed by `PDF-PLUMBER` are discarded. We analyzed a sample of the crawled documents and found a large amount of structured information in the documents, so we use all documents at this stage without additional filtering. This leaves us with 2.3 million documents with approximately 850 million tokens.

C Supervised Zero-shot Results

Tables 4 and 5 shows the full results of the supervised zero-shot finetuning with macro F_1 means and standard deviations across three different runs. While in-batch contrastive fine-tuning outperforms the standard loss in many cases, we can see that, in general, the contrastive loss exhibits higher F_1 variance. For example, in Table 4, the standard deviation when evaluating on the test set of the split II is 10.28, which is very high.

⁶<https://github.com/adrianlopezroche/fdupes>

⁷<https://github.com/jsvine/pdfplumber>

Split	Train Classes	Val Classes	Test Classes
I	letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification	file folder, news article, budget, invoice	presentation, questionnaire, resume, memo
II	file folder, news article, budget, invoice, presentation, questionnaire, resume, memo	letter, form, email, handwritten	advertisement, scientific report, scientific publication, specification
III	advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice	presentation, questionnaire, resume, memo	letter, form, email, handwritten
IV	presentation, questionnaire, resume, memo, letter, form, email, handwritten	advertisement, scientific report, scientific publication, specification	file folder, news article, budget, invoice

Table 3: The four splits of the RVL-CDIP dataset. Each split contains 8 training classes, 4 validation classes and 4 test classes. Validation and test classes do not overlap across splits.

	I		II	
	Valid	Test	Valid	Test
Standard FT	34.76 \pm 6.75	25.33 \pm 2.40	35.64 \pm 2.25	23.29 \pm 2.92
Contrastive FT	37.35 \pm 2.34	25.76 \pm 1.70	32.55 \pm 1.03	26.05 \pm 2.78
CPT + Standard FT	48.24 \pm 3.08	26.97 \pm 3.10	30.45 \pm 1.05	37.81 \pm 5.36
CPT + Contrastive FT	49.68 \pm 0.95	25.82 \pm 1.96	30.31 \pm 0.99	44.44 \pm 10.28

Table 4: Supervised zero-shot performance (Marco F_1) on splits I and II of the RVL-CDIP dataset. We show the mean and standard deviations across 3 runs with different seeds.

	III		IV	
	Valid	Test	Valid	Test
Standard FT	11.67 \pm 0.98	28.84 \pm 1.84	29.68 \pm 7.03	36.75 \pm 3.32
Contrastive FT	18.14 \pm 1.37	27.63 \pm 3.91	29.86 \pm 4.55	32.74 \pm 2.33
CPT + Standard FT	27.20 \pm 4.70	28.11 \pm 1.55	48.82 \pm 1.88	46.09 \pm 2.10
CPT + Contrastive FT	20.80 \pm 0.40	30.43 \pm 0.71	51.26 \pm 2.19	45.07 \pm 5.27

Table 5: Supervised zero-shot performance (Marco F_1) on splits III and IV of the RVL-CDIP dataset. We show the mean and standard deviations across 3 runs with different seeds.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
5
- A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.
- A3. Do the abstract and introduction summarize the paper’s main claims?
1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

3.1

- B1. Did you cite the creators of artifacts you used?
3.1
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3.1
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Data used as in prior work
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Data used as in prior work
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C Did you run computational experiments?

Section 4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

3.2

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

3.3

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Not applicable. did not use existing packages

D **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.