REGCLR: A Self-Supervised Framework for Tabular Representation Learning in the Wild

Weiyao Wang[*]
Johns Hopkins University
wwang121@jhu.edu

Byung-Hak Kim[*,†]
AKASA, Inc.
hak.kim@AKASA.com

Varun Ganapathi
AKASA, Inc.
varun@AKASA.com

Abstract

Recent advances in self-supervised learning (SSL) using large models to learn visual representations from natural images are rapidly closing the gap between the results produced by fully supervised learning and those produced by SSL on downstream vision tasks. Inspired by this advancement and primarily motivated by the emergence of tabular and structured document image applications, we investigate which self-supervised pretraining objectives, architectures, and fine-tuning strategies are most effective. To address these questions, we introduce REGCLR, a new self-supervised framework that combines contrastive and regularized methods and is compatible with the standard Vision Transformer [3] architecture. Then, REGCLR is instantiated by integrating masked autoencoders [6] as a representative example of a contrastive method and enhanced Barlow Twins as a representative example of a regularized method with configurable input image augmentations in both branches. Several real-world table recognition scenarios (e.g., extracting tables from document images), ranging from standard Word and Latex documents to even more challenging electronic health records (EHR) computer screen images, have been shown to benefit greatly from the representations learned from this new framework, with detection average-precision (AP) improving relatively by 4.8% for Table, 11.8% for Column, and 11.1% for GUI objects over a previous fully supervised baseline on real-world EHR screen images.

1 Introduction

Table objects are a compact representation that humans can easily understand. However, this is not true for machines since, unlike classic object detection classes, they might have widely disparate sizes, types, styles, and aspect ratios. In other words, table structure varies greatly between document domains (e.g., Word vs GUI screen), and a large variety of table styles are feasible even within the same document format (e.g., borderless vs bordered). In that regard, very few works have started exploring self-supervised learning (SSL) approaches to the problem of tabular rich document image domains. DiT [9], for example, directly employs BERT-style BEiT [1] to pretrain in a self-supervised manner on IIT-CDIP dataset [8] of 42 million document images and then fine-tune on a couple of classification and detection tasks.

However, to the best of our knowledge, there are several open research problems in developing SSL methods, particularly for tabular and structured document image domains, that require further study. Finding simple but effective self-supervised pretraining objectives, architectures, and fine-tuning strategies are examples of this. Keeping those questions in mind, our paper makes two main contributions:

[*] equal technical contribution
[†] project lead

Table Representation Learning Workshop at NeurIPS 2022.

- First, we introduce REGCLR, a new self-supervised framework that combines contrastive-based masked image modeling[3] and regularized methods and is Vision Transformer (ViT) compatible. It is then instantiated by coupling masked autoencoders (MAE) and enhanced Barlow Twins (eBT) with configurable image augmentations. MAE loss chooses to ignore irrelevant details, while eBT loss pushes relevant details to latent vectors (i.e., minimizing the ignored details), which is highly important for tabular representation learning.
- Second, we pretrain REGCLR on two distinct domains of document image datasets with rich tabular structures, the publicly accessible TableBank [10] and more challenging internal EHRBank, and then validate the higher sample efficiency of the learned tabular representations in finetuning with TableBank and the superior detection performance with EHRBank in various practical downstream settings.

2 REGCLR Description

To begin with, let X be the input image, (X_1, X_2, X_3) be the augmented or varied views of the image, and (Y_1, Y_2, Y_3) be the corresponding hidden/latent representation. Our goal is to learn the best representation Y that retains as much information about the input image X as possible given a reasonable representation constraint. Denote n be the number of samples, d be the feature dimension, and $Z_2, Z_3 \in \mathbb{R}^{n\times d}$ be the projected features.

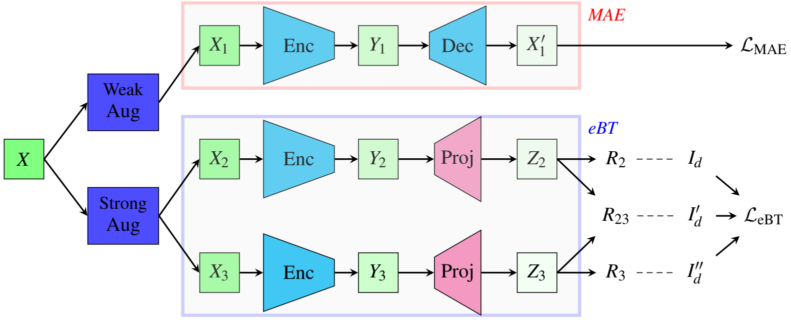

Figure 1: **REGCLR's self-supervised pretraining**. REGCLR is consisted of two branches: MAE and eBT. MAE branch uses weak augmentation and masking to obtain X_1 and then follows procedure introduced in MAE [6] to compute reconstructed loss for unmasked patches as $\mathcal{L}_{\text{MAE}}$ to obtain X_1'. In the eBT branch, strong augmentation is applied twice to get X_2 and X_3. Through the same ViT encoder used in the MAE branch, the encoded images are then projected into features Z_2 and Z_3. The proposed loss $\mathcal{L}_{\text{eBT}}$ computes three correlation matrices between Z_2 and Z_3, attempting to make each matrix near to an identity matrix. REGCLR's overall self-supervised pretraining is performed by jointly minimizing $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{eBT}}$ through optimizing both ViT encoder and decoder.

Training Loss Design[4] With the cross-correlation matrix R_{23} between Z_2 and Z_3 and auto-correlation matrices R_2 and R_3 respectively, we have the resulting objective denoted as $\mathcal{L}_{\text{eBT}}$ given by

$$\begin{aligned}\mathcal{L}_{\text{eBT}} = \nu_1 \sum_i \left(1 - R_{2,ii}\right)^2 + \nu_1 \sum_i \left(1 - R_{3,ii}\right)^2 + \nu_2 \sum_i \left(1 - R_{23,ii}\right)^2 \\ + \mu_1 \sum_i \sum_{j\neq i} \left(R_{2,ij}\right)^2 + \mu_1 \sum_i \sum_{j\neq i} \left(R_{3,ij}\right)^2 + \mu_2 \sum_i \sum_{j\neq i} \left(R_{23,ij}\right)^2,\end{aligned} \tag{1}$$

where ν_1, ν_2, μ_1 and μ_2 are hyper-parameters controlling diagonal and off-diagonal terms of each matrix. Combining the $\mathcal{L}_{\text{MAE}}$ defined in [6], the overall training loss is now straightforward to compose as:

$$\mathcal{L}_{\text{REGCLR}} = \mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{eBT}}. \tag{2}$$

Self-Supervised Pretraining We use the new objective, $\mathcal{L}_{\text{REGCLR}}$ in (2), for self-supervised pretraining, and build the model with two branches, MAE and eBT, as shown in Figure 1. For the MAE

[3] We follow Yan LeCun's broader classification of contrastive methods introduced in [7].

[4] For a detailed loss derivation, see [11].

branch, we choose to apply weak augmentation to the input image and then randomly mask out the selected patches. Only unmasked patches are fed into the ViT encoder, and subsequently the masked patches are reconstructed by the Vit decoder using the MSE calculated over the masked patches as the loss function $\mathcal{L}_{\text{MAE}}$ per the original MAE design. Secondly, the eBT branch operates on the cross embedding of the input image's two strongly augmented versions X_2 and X_3.

Detection via ViT Backbone For detection, we combine a ViT encoder and decoder pretrained in a self-supervised manner with MIMDet [4] to serve as the detection backbone and leverage Cascade Mask R-CNN [2], which is the common architecture in supervised state-of-the-art systems. Compared to previous representative approaches of adapting vanilla ViT for object detection, MIMDet replaces the pretrained patchify stem with a compact convolutional stem without further upsampling or redesigns, resulting in a ConvNet-ViT hybrid multi scale feature extractor that requires far fewer epochs in the fine-tuning procedure.

3 Experiments

We evaluate our methods in the TableBank dataset [10] and an internally collected EHRBank dataset which consists of screenshots collected by bots as they navigate the EHR systems of ten US health systems from various EHR providers. A total of 2,537 labeled images are collected for table detection task, 1,657 for table column detection task, and 28,121 unlabled images for pretraining.

TableBank Dataset The prediction results of the TableBank table detection are shown in Table 1 in AP (mAP @ IOU [0.50:0.95]) and AP_{75} (mAP @ IOU 0.75) with the results of two baselines. Our method REGCLR outperforms the other self-supervised and fully supervised baselines.

Table 1: **Results of table detection on the TableBank test set** (in AP and AP_{75}). MAE denotes the representative SSL pretraining baseline, while ResNet [5] stands for the purely supervised baseline using the ResNet-152 backbone, with Cascade Mask R-CNN. The bold value represents the best (highest) value for each column metric. All baselines are outperformed by the proposed REGCLR.

	Word		**Latex**	
	AP	**AP_{75}**	**AP**	**AP_{75}**
ResNet (supervised baseline)	95.42	95.78	97.32	98.62
MAE (self-supervised baseline)	95.94	96.16	97.63	98.70
REGCLR (our method)	**96.03**	**96.22**	**97.68**	**98.75**

Table 2: **Results of GUI elements detection on the EHRBank Table and Column test sets** (in AP and AP_{75}). REGCLR performs best when pretraining with the EHRBank Screenshot dataset, increasing AP scores relatively by 4.8% for Table and 11.8% for Column over the supervised baseline, as seen by comparing the first and second rows. Interestingly, despite pretraining with approximately 10% volumes of TableBank, RegCLR fast approaches the best cross-domain transfer results from TableBank to EHRBank in the last row.

		Table		**Column**	
Pretrain on	Method	**AP**	**AP_{75}**	**AP**	**AP_{75}**
N/A	**ResNet**	40.53	44.46	61.43	67.07
EHRBank Screenshot	**REGCLR**	**42.46**	**45.32**	**68.68**	**75.17**
	MAE	36.78	39.05	64.67	71.09
TableBank (cross-domain)	**REGCLR**	40.96	43.47	67.77	75.06
	MAE	**43.99**	**48.77**	**69.83**	**77.29**

EHRBank Dataset We then evaluate REGCLR on internal EHRBank dataset which has higher values in our production scenario. As shown in Table 2, when pretrained on *unlabeled* EHRBank, our method outperforms the baselines in both Table and Column detection, increasing relative AP scores by 4.8% and 11.8% respectively over the supervised baseline, as seen by comparing the first and second rows. Furthermore, even though pretrained with only around 10% of the TableBbank volume, REGCLR quickly approaches the best cross-domain transfer performance from TableBank to EHRBank, as shown in the last row.

Additionally, it is worth noting that MAE performs worse than even ResNet on Table when pretrained on EHRBank (by comparing the first and third rows). It does, however, transfer better than REGCLR in scenarios involving cross-domain transfer from public TableBank to private EHRBank (by comparing the last two rows). In the future, we intend to investigate how quickly detection performance improves as the unlabeled data volume scales, as well as how effectively pretrained weights transfer across domains in the context of tabular rich images, so that they can be applicable to other document format datasets (e.g., Word to GUI and vice versa). More information on the experimental results can be found in [11].

4 Conclusion

In this paper, we present REGCLR, a brand-new framework combines contrastive and regularized self-supervised methods and has been pretrained on both public and private domain tabular rich images. We demonstrate that REGCLR outperforms previous self-supervised pretraining and fully supervised baselines by a large margin in various real-world contexts, with high sample efficiency for fine-tuning. We believe that this study is an important step towards semantic comprehension of real-world document images, and it will be interesting to see how this vision-based framework can be expanded to include textual content without manual data annotation.

References

[1] H. Bao, L. Dong, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Z. Cai and N. Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2019. ISSN 1939-3539. doi: 10.1109/tpami.2019.2956516. URL http://dx.doi.org/10.1109/tpami.2019.2956516.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR 2021)*, 2021.

[4] Y. Fang, S. Yang, S. Wang, Y. Ge, Y. Shan, and X. Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*, 2022.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2022)*. IEEE, 2022.

[7] Y. LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022.

[8] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 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*, pages 665–666, 2006.

[9] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei. Dit: Self-supervised pre-training for document image transformer. *arXiv preprint arXiv:2203.02378*, 2022.

[10] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li. Tablebank: A benchmark dataset for table detection and recognition. *arXiv preprint arXiv:1903.01949*, 2019.

[11] W. Wang*, B.-H. Kim*, and V. Ganapathi. RegCLR: A self-supervised framework for tabular representation learning in the wild. *arXiv preprint arXiv:2211.01165*, 2022.