DIT: self-supervised pretraining for document image transformer

Document Ai models usually start with OCR or doc-layout analysis which still relies heavily on the supervised computer vision models with human labeled training sample.

- **Domain shifts and template format mismatch** is usually the reason why this can't be done on large scale dataset.
- Domain of these datasets are usually from academic papers that share similar templated and formats. Therefore, when used in the real world dataset, it doesn't work.

Vital to pretrain the document imaged models with large scale of unlabeled data from general domains

DiT : A self-supervised pretrained Document Image Transformer for general docs inspired by BEiT model.

Difference: BEiT model utilizes discrete VAE in DALL-E. For DiT, dVAE is retrained with large-scale document images so that its more domain relevant.

Similarity:

Input text is resized (224x224) and split into sequence of 16x16 patches which are used as the input image transformer.

Pretraining objective is to recover visual tokens from dVAE based on the corrupted input document images using MIM (Masked Image Modelling)

Implementing the similarities, DiT doesn't rely on human-labeled document images. Only leverages unlabeled data to learn the global patch relationships within each document image.

The datasets: RVL-CDIP (Image classification) / PubLayNet (document layout analysis) / ICDAR 2019 CTDaR dataset for table detection and FUNSD dataset for OCR text detection.

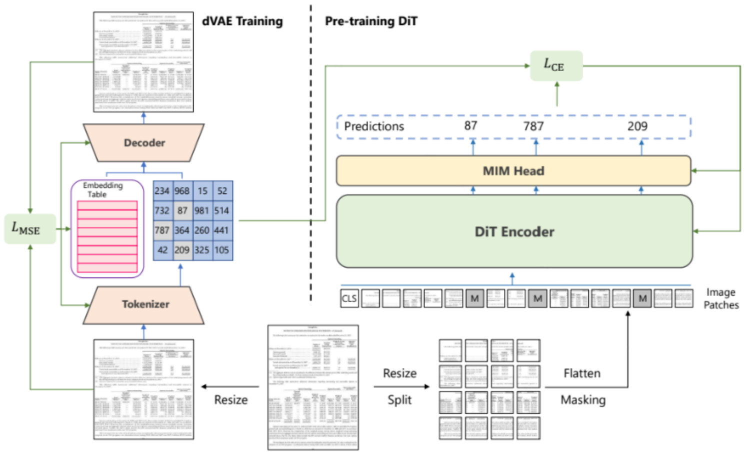
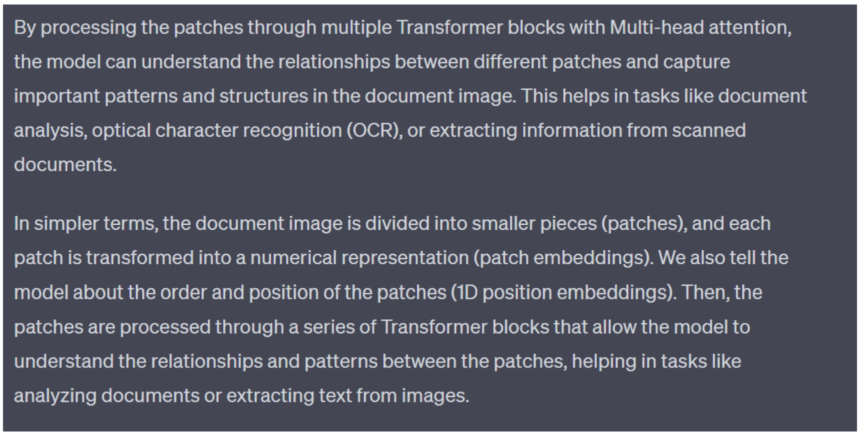


Figure 2: The model architecture of DiT with MIM pre-training.

Vanilla Transformer architecture as the backbone of DiT.

- **Divide document image into non overlapping patches and obtain sequence of patch embeddings.**
- After adding the **1d position embedding**, these patches pass through **stack of Transformer** blocks with multi-head attention.



By processing the patches through multiple Transformer blocks with Multi-head attention, the model can understand the relationships between different patches and capture important patterns and structures in the document image. This helps in tasks like document analysis, optical character recognition (OCR), or extracting information from scanned documents.

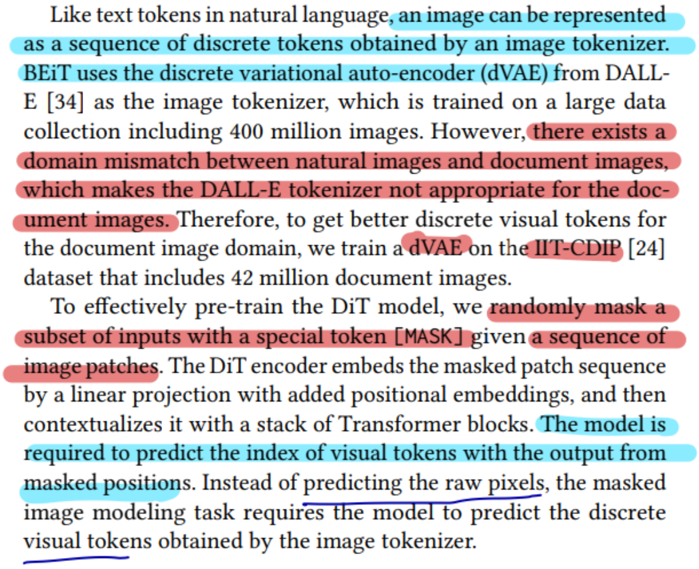
In simpler terms, the document image is divided into smaller pieces (patches), and each patch is transformed into a numerical representation (patch embeddings). We also tell the model about the order and position of the patches (1D position embeddings). Then, the patches are processed through a series of Transformer blocks that allow the model to understand the relationships and patterns between the patches, helping in tasks like analyzing documents or extracting text from images.

- Take the **output of the Transformer encoder as the representation of image patches**,

  1d position to tell the model the relative position or order of the patches within the document.

  Transformers to capture relationships and dependencies to indicate important patters and structures in the document image. This helps with document analysis, OCR or extracting information from scanned docs.

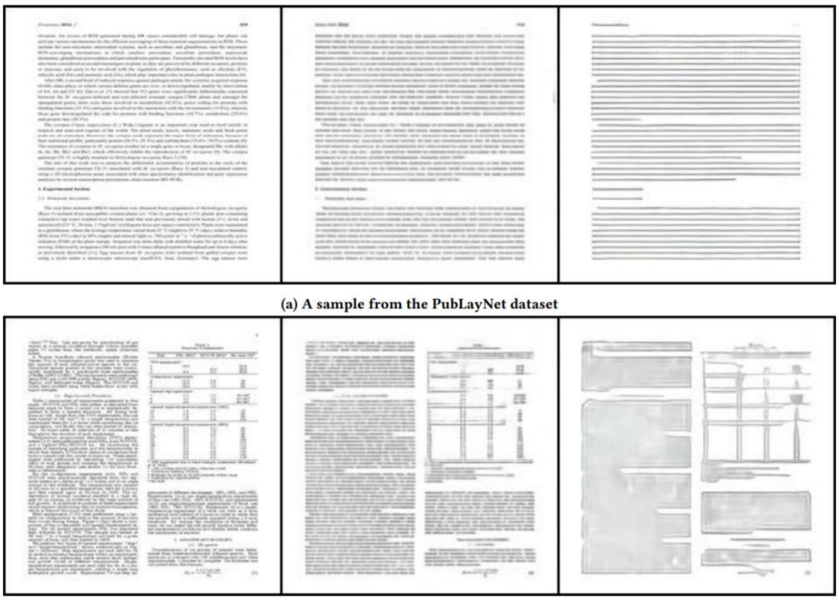
For domain relevance, dVAE is retrained with dataset that includes 42 million document images.



Like text tokens in natural language, an image can be represented as a sequence of discrete tokens obtained by an image tokenizer. BEiT uses the discrete variational auto-encoder (dVAE) from DALL-E [34] as the image tokenizer, which is trained on a large data collection including 400 million images. However, there exists a domain mismatch between natural images and document images, which makes the DALL-E tokenizer not appropriate for the document images. Therefore, to get better discrete visual tokens for the document image domain, we train a dVAE on the IIT-CDIP [24] dataset that includes 42 million document images.

To effectively pre-train the DiT model, we randomly mask a subset of inputs with a special token [MASK] given a sequence of image patches. The DiT encoder embeds the masked patch sequence by a linear projection with added positional embeddings, and then contextualizes it with a stack of Transformer blocks. The model is required to predict the index of visual tokens with the output from masked positions. Instead of predicting the raw pixels, the masked image modeling task requires the model to predict the discrete visual tokens obtained by the image tokenizer.

From left to right: the original document image, image reconstruction using the self-trained dVAE tokenizer, image reconstruction using the DALL-E tokenizer.



(a) A sample from the PubLayNet dataset

(b) A sample from the ICDAR 2019 cTDaR dataset

DallE is hard to distinguish the borders and lines.

Fine-Tuning to be adapted to specific tasks as document image classification, document layout analysis, table detection and text detection. **Model can learn task-specific patters**. **Each of these benchmark datasets represents different tasks such as variations in image quality etc.** Finetuning on these diverse datasets allows generalization.

For object detection, model used **Cascade RCNN** which ensures optimized extraction. This improves the model's ability to locate and identify images within the docs.

Also, Fine-tuning allows the model to refine the representation of image patches which improve the model's ability to classify image accurately based on their content.

The datasets: RVL-CDIP (Image classification) / PubLayNet (document layout analysis) / ICDAR 2019 CTDaR dataset for table detection and FUNSD dataset for OCR text detection.

- RVL-CDIP : {letter, form, email, handwritten, advertisement, report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo. Evaluation Metric is the overall classification accuracy.
- PubLayNet : title, list figure and table. Detect the regions for the assigned elements
- ICDAR 2019 CTDaR : different tables from archival documents and modern docs. E.g. stock-exchange lists, hand-drawn accounting books etc.
- FUNSD : noisy scanned document dataset for text detection.

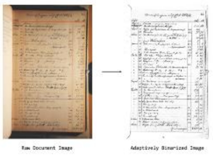
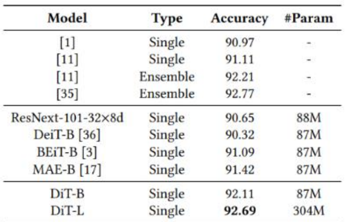


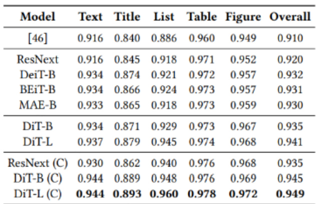
Figure 5: An example of pre-processing with adaptive image binarization on the ICDAR 2019 cTDaR archival subset.

결론: 결국 DiT-L 이 많은 부분에서 다른 classification model들을 이겼다.



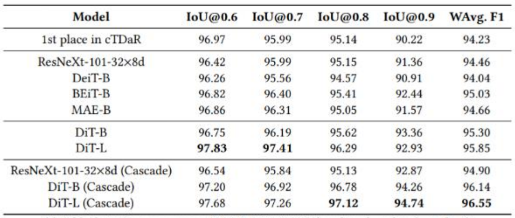
Model	Type	Accuracy	#Param
[1]	Single	90.97	-
[11]	Single	91.11	-
[11]	Ensemble	92.21	-
[35]	Ensemble	92.77	-
ResNext-101-32×8d	Single	90.65	88M
DeiT-B [36]	Single	90.32	87M
BEiT-B [3]	Single	91.09	87M
MAE-B [17]	Single	91.42	87M
DiT-B	Single	92.11	87M
DiT-L	Single	**92.69**	304M

Table 1: Document Image Classification accuracy (%) on RVL-CDIP, where all the models use the pure image information (w/o text information) with the 224×224 resolution.



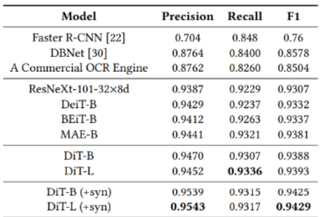
Model	Text	Title	List	Table	Figure	Overall
[46]	0.916	0.840	0.886	0.960	0.949	0.910
ResNext	0.916	0.845	0.918	0.971	0.952	0.920
DeiT-B	0.934	0.874	0.921	0.972	0.957	0.932
BEiT-B	0.934	0.866	0.924	0.973	0.957	0.931
MAE-B	0.933	0.865	0.918	0.973	0.959	0.930
DiT-B	0.934	0.871	0.929	0.973	0.967	0.935
DiT-L	0.937	0.879	0.945	0.974	0.968	0.941
ResNext (C)	0.930	0.862	0.940	0.976	0.968	0.935
DiT-B (C)	0.944	0.889	0.948	0.976	0.969	0.945
DiT-L (C)	**0.944**	**0.893**	**0.960**	**0.978**	**0.972**	**0.949**

Table 2: Document Layout Analysis mAP @ IOU [0.50:0.95] on PubLayNet validation set. ResNext-101-32×8d is shortened as ResNext and Cascade as C.



Model	IoU@0.6	IoU@0.7	IoU@0.8	IoU@0.9	WAvg. F1
1st place in cTDaR	96.97	95.99	95.14	90.22	94.23
ResNeXt-101-32×8d	96.42	95.99	95.15	91.36	94.46
DeiT-B	96.26	95.56	94.57	90.91	94.04
BEiT-B	96.82	96.40	95.41	92.44	95.03
MAE-B	96.86	96.31	95.05	91.57	94.66
DiT-B	96.75	96.19	95.62	93.36	95.30
DiT-L	**97.83**	**97.41**	96.29	92.93	95.85
ResNeXt-101-32×8d (Cascade)	96.54	95.84	95.13	92.87	94.90
DiT-B (Cascade)	97.20	96.92	96.78	94.26	96.14
DiT-L (Cascade)	97.68	97.26	**97.12**	**94.74**	**96.55**

(a) Table detection accuracy on ICDAR 2019 cTDaR (combined: archival+modern)



Model	Precision	Recall	F1
Faster R-CNN [22]	0.704	0.848	0.76
DBNet [30]	0.8764	0.8400	0.8578
A Commercial OCR Engine	0.8762	0.8260	0.8504
ResNeXt-101-32×8d	0.9387	0.9229	0.9307
DeiT-B	0.9429	0.9237	0.9332
BEiT-B	0.9412	0.9263	0.9337
MAE-B	0.9441	0.9321	0.9381
DiT-B	0.9470	0.9307	0.9388
DiT-L	0.9452	**0.9336**	0.9393
DiT-B (+syn)	0.9539	0.9315	0.9425
DiT-L (+syn)	**0.9543**	0.9317	**0.9429**

Table 4: Text detection accuracy (IoU@0.5) on FUNSD Task #1, where Mask R-CNN is used with different backbones (ResNeXt, DeiT, BEiT, MAE and DiT). "+syn" denotes that DiT is trained with a synthetic dataset including 1M document images, then fine-tuned with the FUNSD training data.