DocCT: Shift Document Image Classification Research from Format to Content

Anonymous ACL submission

Abstract

Document image understanding is challenging, given the complexity of the combination of illustrations and text that makes up a document image. Previous document image classification datasets and models focus more on the document format while ignoring the meaningful content. In this paper, we introduce DocCT, the first-of-its-kind document image classification dataset that covers various daily topics that require understanding fine-grained document content to perform correct classification. Further, since previous image models cannot sufficiently understand the semantic content of document images, we present DocMAE, a new self-supervised pre-trained document image model. Experiments show that DocMAE's ability to understand fine-grained content is far greater than previous models and even surpasses OCR-based models, which proves that it is possible to well understand the semantics of document images only with the help of pixels.[1]

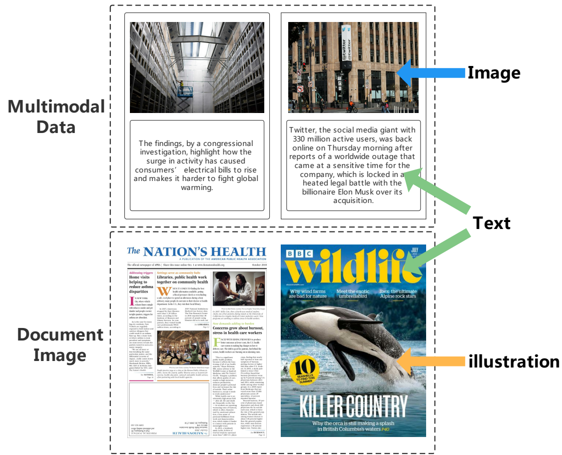

Figure 1: Comparison between multimodal data and document images. Multimodal data consist of a separate pair of images and text, while text and illustrations compose a whole document image.

1 Introduction

The task of visual document understanding (VDU) aims at automatically reading and understanding document images. Digital images of documents are an important source of information; for example, in digital libraries, documents are often stored as scanned images before further processing such as optical character recognition (OCR) (Harley et al., 2015). Figure 1 shows a document image example and its difference to common multimodal data (Wang et al., 2022b). A document image contains rich content elements, like text, images, and diagrams, organized in various styles. One important task toward visual document understanding is document image classification (DIC), which aims to classify a document image into a category, similar to vanilla image classification like ImageNet (Deng et al., 2009). DIC can be used in various applications, such as automatic book classification in the library, helping Internet search engines better integrate different information, or determining which domain-specific model should be used for OCR. It is also an essential step toward a more fine-grained understanding of document images, which can inspire some downstream tasks such as document visual question answering (Mathew et al., 2021).

RVL-CDIP (Harley et al., 2015) is the most widely used large-scale dataset for DIC research. It categorizes document images into 16 classes like "email", "invoice" and "magazine", based on their formats. However, it pays little attention to the document image's concrete content, while semantics conveyed by the content is also essential. For example, rather than knowing whether a document is an email, we want to know more about what topic the email talks about. Further, the data in RVL-CDIP are all under a similar topic, which makes it unable to be used for classification by distinguishing detailed content between different documents. The obstacle that is hindering the further development of DIC methods that can achieve content type classification is the lack of suitable datasets.

[1]The dataset and source code will be available at Github.

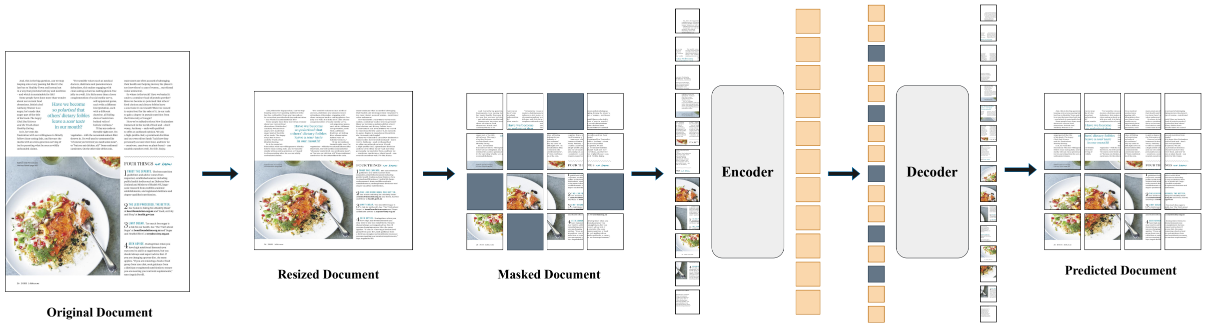

Figure 2: The overall pipeline of DocMAE consists of an encoder and a decoder, mainly following the architecture of MAE (He et al., 2022). The input document image is first resized to 640×640 and then split into numbers of patches. Some patches are masked by a certain ratio. Then the unmasked patches are concatenated to a sequence and fed into the transformer encoder. The masked patches and the output of the encoder are combined together and sent to the transformer decoder to predict the pixel of the masked patches.

Therefore, in this paper, to facilitate the further research in DIC, we present the first document image dataset including fine-grained topic annotations - **DocCT** (***D****ocument Image* ***C****lassification via* ***T****opic*). In DocCT, there are 10 categories, all of which are common topics in daily life. Each category contains documents in various formats. DocCT can prompt models' content understanding ability about document images since the model can classify them correctly only when their content is understood.

With DocCT, we then evaluate some state-of-the-art models developed for document images. Current DIC methods can be summarized into two categories. One is directly using image classification methods like CNN (Harley et al., 2015) or transformers (Li et al., 2022), which are usually used in document format or layout analysis. The other is a two-stream multimodal method that first extracts text by OCR and then performs classification with both OCR-text and image features (Appalaraju et al., 2021; Huang et al., 2022), while its model performance is heavily restricted by the quality of text extracted by OCR. Our experiments reveal a huge performance drop of those two kinds of DIC models from humans, which proves that document image understanding is still challenging, and DocCT is thus worth researching.

To develop an effective method for the content-based document image classification problem, we present a new self-supervised pre-trained model - **DocMAE** (**Doc**ument **M**asked **A**uto**E**ncoder), which is trained with large-scale unlabeled document images. In DocMAE, we enlarge the input image size to better understand the semantics of text composed of pixels. Experimental results on DocCT demonstrate that this adjustment dramatically improves the model's ability to recognize a fine-grained semantic topic in images, thus significantly surpassing previous models, even OCR-based methods, in classification, making it more suitable for some content-dependent image tasks.

Our contributions can be summarized as follows: (1) We present DocCT, the first DIC dataset with fine-grained content type annotations that can be used for document image topic classification tasks. (2) We present DocMAE, a self-supervised pre-trained model with a deeper understanding of content in images. (3) Our experimental results reveal some unique challenges from DocCT. Further, with DocMAE, we prove that the model can also understand the document image content by pixels without explicitly extracting its text by OCR.

2 Related Work

Document Image Classification With the development of deep image models, document image related research is attracting more attention. Compared to vanilla image research on ImageNet (Deng et al., 2009), document images are more complex given their much richer content. As an important task for the document images, DIC (Chen and Blostein, 2007) is one of the earliest and most researched directions. In DIC, a given document image should be classified into a correct category by specific requirements. The most widely used dataset is RVL-CDIP (Harley et al., 2015), a subdataset of IIT-CDIP (Lewis, 2006). The images in IIT-CDIP are scanned documents collected from the public records of lawsuits against American tobacco companies. RVL-CDIP contains 16 different document formats such as “letter” or “invoice”,

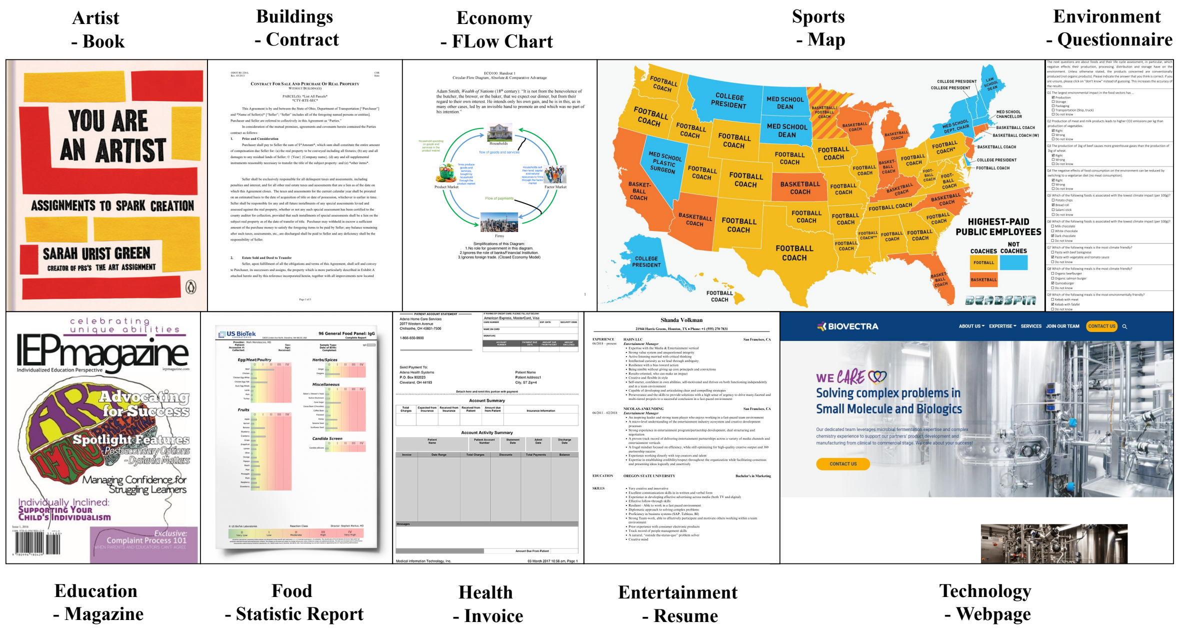

Figure 3: 10 categories and some of the formats in DocCT.

which can be used to evaluate models' classification ability.

However, compared to recognizing the format of an image, understanding its content is more critical and challenging, since it can facilitate lots of higher-level AI research such as visual question answering (Antol et al., 2015). Thus in this paper, we present DocCT, the first DIC dataset that focuses on document content understanding, hoping to prompt research in related fields.

Pre-training Document Models The goal of pre-training technologies is to use a large amount of unsupervised text to pre-train a model, so that the model can master prior knowledge, improving the performance of downstream tasks. After the success of ViT (Dosovitskiy et al., 2020), which first applies vanilla transformer (Vaswani et al., 2017) to vision tasks, researchers start to investigate how to better pre-train ViT in image-related tasks like BERT (Devlin et al., 2018) in natural language processing. Currently, there are also some pre-trained models for document image related research. DiT (Li et al., 2022) is a pre-trained document model based on BEiT (Bao et al., 2021). Some document models convert document image tasks into a multimodal task, such as LayoutLM (Huang et al., 2022), DocFormer (Appalaraju et al., 2021), and LiLT (Wang et al., 2022a). They use OCR to extract the text information from a document image and input both the original image and OCR text into the models. Compared to pure image models, they can obtain higher accuracy with the extra text input, while the training process is time-consuming and inefficient in making an inference.

However, most previous pre-trained document models aim at document layout analysis, making them unsuitable for solving fine-grained document content understanding when applied to datasets like DocCT. Thus in this paper, we present DocMAE, a large-scale self-supervised pre-trained model. It is a pure image model like DiT without OCR, while it is also helpful in understanding the semantic information in the image and can be further used in other document-related downstream tasks.

3 DocCT Dataset

In this section, we present the DocCT dataset. We first introduce the composition of the dataset, including the topics we adopted, and describe the procedure of how we collected, organized and annotated it. Then we analyze the dataset by comparing it with other datasets and in case studies.

3.1 Data Collection

We collected our dataset from web images with search engines. To cover as many topics as possible, we started from the root node of the wiki's category tree and selected 10 most commonly seen topics in our daily life, including "Artist", "Buildings", "Economy", "Education", "Food", "Entertainment", "Environment", "Sports", "Health", and

Category	#Count	Category	#Count
Artist	2531	Buildings	2089
Economy	2603	Education	2609
Food	3301	Entertainment	1984
Sports	2541	Environment	1544
Health	2032	Technology	2278

Table 1: Statistics of DocCT. The total number of document images is 23512.

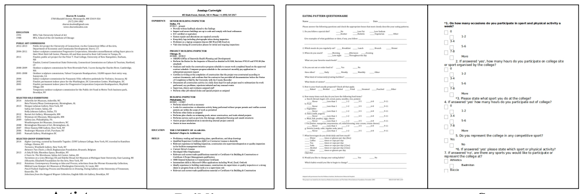

Figure 4: Comparison between two categories with the same format.

"Technology".

For each category, to ensure most of the search results from search engines are relevant documents, we constructed our search keywords with the category name alongside diverse document format names. As for the document format, we first adopted 16 types in RVL-DCIP and then added some novel formats to cover as many formats as possible. Finally, we settled on a total of 27 types of formats, including "book", "budget", "contract", "email", "exam", "flow chat", "form", "introduction", "invoice", "letter", "magazine", "map", "memo", "newspaper", "phone application", "poster", "presentation slides", "print advertisement", "questionnaire", "resume", "scientific publication", "specification", "statistical repost", "textbook", and "webpage". For each format, we collected up to 300 images. With those topics and formats as the search keywords, we roughly crawled nearly 80K images in the collection procedure.

3.2 Annotation and Quality Control

We then asked crowdworkers to annotate the crawled images. Given an image, the annotating procedure is as follows:

- **Step 1**: Determine whether the image is a document image. An image without any text information or with too vague text to recognize will be dropped.
- **Step 2**: Determine whether the document image conforms to the corresponding category. The irrelevant image will be removed. If an image can belong to more than one category, it will also be discarded.

Only images that pass the above judgments will be considered valid and be kept. After manual filtering, we obtained about 23K accurate document image samples. In Table 1, we provide the statistics for each category in DocCT.

3.3 Data Analysis

With lots of different formats, DocCT is able to reflect common knowledge content that documents in different formats can narrate under the same topic in our daily life, making the research on it more applicable. Compared with RVL-CDIP, the formats we chose contain more modern and diverse document formats with more vivid colors than a single white-black scanned file. In Figure 4, we present comparisons between two different categories with the same format. In DocCT, the layout of two different categories with the same format is very similar. This can ensure that models cannot cheat with layouts and must analyze the detailed content. Models can yield correct classification only through understanding the semantics conveyed by a document image.

4 DocMAE

In this section, we present the DocMAE model. We first describe the basic architecture of DocMAE and how we pre-train DocMAE, and then introduce the selection of input image size. Finally, we provide some image restoration examples to examine the performance of pre-trained DocMAE.

4.1 Architecture

Different from DiT and LayoutLM, which use BEiT (Bao et al., 2021) as the visual backbone, in this paper, we choose MAE (He et al., 2022) as the basic architecture of DocMAE. Compared with BEiT, using dVAE (Rolfe, 2016) to tokenize image patches first, MAE directly uses pixel reconstruction to calculate model prediction loss. This is a better choice for a document image since the pixels of a document image are more complex and contain more semantics. It is difficult to represent all cases with a limited number of tokens (8192 tokens used in BEiT).

Chen et al. (2022) proves that as an important part of MAE, the decoder can steal some abilities

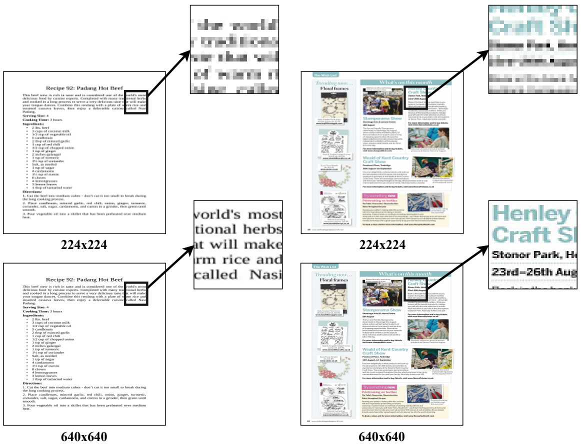

Figure 5: Comparison between 224×224 and 640×640. The top is a plain text image while the bottom is a rich text image. In either case, the image with the size of 224 loses most of the text information, while the image with 640 keeps the text legible.

from the encoder, which will significantly limit the encoder's ability when only using the encoder to do a downstream task. Thus in DocMAE, different from the original MAE, we keep both the encoder and decoder when fine-tuning to ensure a better performance.

4.2 Pre-training Settings

We used MAE$_{base}$ as the basic architecture of DocMAE. The DocMAE encoder is a 12-layer transformer with 768 hidden size and 12 attention heads. The feed-forward network size is 3072. The DocMAE decoder is a 7-layer transformer with 512 hidden size and 16 attention heads. The feed-forward network size is 2048. The input image size is 640×640, and we employed 20×20 as the patch size. A special `[CLS]` token was concatenated to the start of the patch sequence. The mask ratio was set to 30%, which means that in pre-training, while the input sequence length to the encoder is 718 $(717 + 1)$, the input sequence length to the decoder is 1025 $(1024 + 1)$. To make DocMAE adapt to documents of different original resolutions and shapes, we randomly cropped the input images with 10% probability during pre-training.

We pre-trained DocMAE for 100 epochs with 512 batch size. The optimizer is Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and weight decay is 0.05. The start learning rate is 1e-4 with cosine annealing learning rate decay and without warmup. The dropout was disabled. The whole pre-training procedure lasted three weeks with four RTX 3090 GPUs.

4.3 Pre-training Corpus

To make DocMAE applicable to more diverse tasks, unlike DiT and LayoutLM, which directly use documents from IIT-CDIP, we used open-domain magazines as the pre-training corpus since magazines contain various document types, including both plain and rich text. We collected massive magazines and converted each magazine into a collection of document images. In total, we collected around 1.6 million open-domain document images. Since the collection method of the pre-training corpus is different from DocCT, we added an additional data filter to remove the data duplication between them.

4.4 Input Size Setting

Input size plays an essential role in a deep learning image model since too small image size will lead to loss of information, while too large image size will make it difficult to train the model. To balance training time and information retention, almost all previous image models chose 224×224 as the input image size. This image size has achieved excellent results on object image classification datasets such as ImageNet (Deng et al., 2009). Thus, some document image pre-trained models, such as DiT and LayoutLM, also chose this size for the input.

However, our investigation showed that 224 is not an appropriate size for images with text information such as document images. The small input size will lead to the loss of text information. This may have small effect on identifying whether an animal is a cat or dog, or figuring out the layout of a table in a document. However, if we want the model to identify more fine-grained text semantics in an image, the expansion of the input size is required since a word is much smaller in an image than a cat or table. We chose 640 as the input image size, ensuring that the text in most document images is recognizable while still applicable for training the model. The comparison of 224×224 and 640×640 is shown in Figure 5. The image with 640 size proves to contain richer and clearer text information either in plain or rich text images.

4.5 Evaluation

After pre-training, we used DocMAE to restore some document images randomly searched on the Internet. Some of the results are shown in Figure 6. We inputted a picture with 30% of the random area masked and observed the output. It can be found that the overall image can be restored relatively

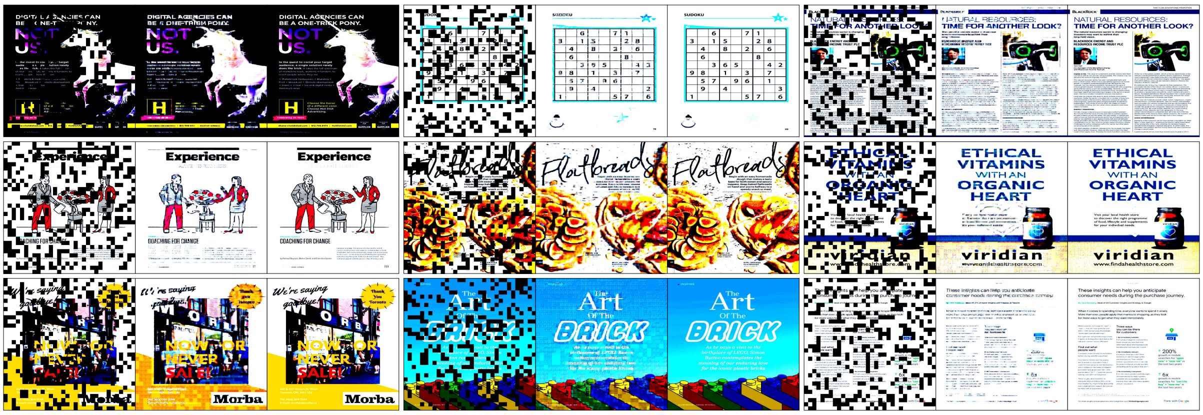

Figure 6: Image restoration for some document images. From left to right are the **masked image**, **restored image**, and **original image**. The mask ratio is 30%.

well, and the restoration for larger texts is excellent. However, for texts with small font size, the restoration is still kind of blurred. This shows that DocMAE still has some room for improvement.

5 Experiments

We conducted the experiments on different datasets, including RVL-CDIP and DocCT, with DocMAE and other document image related models. DocCT was split into training, validation and test sets with the ratio of 8:1:1. We used the training set to train the model, took the best model on the validation set, and then recorded its performance on the test set. We evaluated DocMAE in three ways. One is DocMAE$_{encoder}$, which uses only the encoder of DocMAE. The other is DocMAE$_{decoder}$, which fixes the parameters of the encoder and fine-tunes only the decoder. The last DocMAE$_{full}$ is to fine-tune all the parameters in the encoder and decoder. Compared models are mainly divided into two categories. One is image-only models which depends entirely on the processing of pixels, including BEiT (Bao et al., 2021), DiT (Li et al., 2022), and MAE (He et al., 2022). The other is OCR-enhanced multimodal models with text extracted by OCR as the additional input; here, we chose LayoutLMv3 (Huang et al., 2022).

5.1 Performance on RVL-CDIP

We first evaluated DocMAE on RVL-CDIP to see its performance in the document format classification task. The experimental results are shown in Table 2. DocMAE$_{decoder}$ achieves the state-of-the-art with 92.78 accuracy among the image-only models and surpasses the previous best model DiT$_{large}$ (92.69). This proves that enlarging the input image's resolution can help even in the document format classification task.

5.2 Performance on DocCT

5.2.1 Classification Accuracy

In the document content classification task on DocCT, DocMAE achieves the best performance among all the image-only models. DocMAE$_{full}$ obtains a comparable result to the OCR-based methods such as LayoutLMv3 (74.54 vs. 76.94 in F1), and DocMAE$_{decoder}$ greatly excels the OCR-based methods, which demonstrates that it is possible to capture the semantic information by using only pixel data in document images instead of directly using OCR.

It can also be found that MAE obtains a much higher F1 than DiT, proving that direct pixel prediction as a pre-training task is better in understanding document semantics than token prediction used in BEiT and DiT. This is mainly because text pixels are more complex, and it is difficult to summarize all the possible image patches by just using 8192 tokens.

Furthermore, we randomly selected 500 images for human annotators to classify, and the accuracy of human beings is 96.20%, which is much higher than the current deep learning models. It shows that the models still have a lot of room for improvement, and DocCT proves to be a challenging dataset that is worth researching.

5.2.2 Encoder vs. Decoder

We then performed ablation analysis for different parts of the DocMAE architecture to observe the effect of the different modules on the accuracy. We first fine-tuned DocMAE only with the encoder.

Model	RVL-CDIP ACC	DocCT F1	DocCT ACC	DocCT Train/Epoch	DocCT Infer/Epoch	Image Size	#Param
Human	-	-	96.20	-	-	-	-
Image-Only Models							
BEiT$_{base}$	91.09	38.48	38.65	2m32s	30s	224	87M
DiT$_{base}$	92.11	39.89	39.92	2m30s	31s	224	87M
DiT$_{large}$	92.69	43.58	43.95	4m47s	40s	224	304M
MAE$_{base(encoder)}$	91.42	41.92	42.00	2m31s	31s	224	87M
MAE$_{base(decoder)}$	-	41.22	40.94	2m20s	30s	224	113M
MAE$_{base(full)}$	-	42.17	42.68	3m05s	35s	224	113M
DocMAE$_{224(encoder)}$	-	45.10	45.86	2m31s	30s	224	87M
DocMAE$_{224(full)}$	-	46.76	47.09	3m05s	35s	224	113M
DocMAE$_{224(decoder)}$	-	46.94	47.60	2m20s	30s	224	113M
DocMAE$_{encoder}$	-	36.30	37.46	17m40s	1m01s	640	87M
DocMAE$_{full}$	92.22	74.54	74.55	31m13s	1m19s	640	113M
DocMAE$_{decoder}$	92.78	**83.53**	**83.84**	14m22s	1m17s	640	113M
OCR-Enhanced Models							
LayoutLMv3$_{base}$	95.44	75.63	75.64	51m51s	7m05s	224	133M
LayoutLMv3$_{large}$	95.93	76.94	76.91	55m32s	7m11s	224	368M

Table 2: Experimental results on RVL-CDIP and DocCT with different models. DocMAE$_{encoder}$ means we utilized only the DocMAE encoder for the classification model. DocMAE$_{decoder}$ means the parameters in the encoder were fixed and only the decoder was fine-tuned. DocMAE$_{full}$ means both the encoder and decoder were used to be fine-tuned. Training and inference time was calculated on a single RTX3090 GPU within one epoch.

Compared to full DocMAE, the DocMAE encoder obtains only 36.30 in F1. This vast performance drop proves that in dealing with document images, the decoder is an essential part and cannot be removed as MAE does for ImageNet.

Another interesting finding is that, when the encoder module of DocMAE is fixed and only the decoder module is fine-tuned, the model obtains even higher accuracy (83.53 vs. 74.54 in F1). We think this phenomenon is because, when DocMAE is fully pre-trained, the encoder can already extract the features of a document image well. Any further fine-tuning of the encoder will affect the feature extraction ability, thus affecting the overall accuracy. Document images are more complex than images of simple objects, making this disturbance more obvious. Our experiments prove that in DocMAE, the encoder is suitable for acting as a feature extractor while the decoder can be used for migrating to downstream tasks.

5.2.3 Influence of Resolution

To confirm that the input image resolution does affect the model's understanding of the semantics in a document image, we additionally pre-trained a model named DocMAE$_{224}$ with the same settings as DocMAE. The only difference is that the input image size of DocMAE$_{224}$ is 224×224. The experimental comparison results are shown in Table 2. Although the performance of DocMAE$_{224}$ on DocCT is much better than the original MAE with the help of pre-training based on document image data, there is still a huge gap compared to DocMAE with 640 image resolution (46.94 vs. 83.53 in F1). This result effectively proves that larger resolution is crucial for the semantic understanding of document images.

5.2.4 Model Efficiency

Since the model structure of different methods varies, we also recorded the efficiency of the different models during training and inference. Compared with DiT$_{base}$, DocMAE$_{full}$ is much slower (31m13s vs. 2m30s), because, as the length of input image patches increases (1025 vs. 197), the training time also increases exponentially. However, when it comes to inference, DocMAE is not much slower than DiT$_{base}$ (1m19s vs. 31s) and DiT$_{large}$ (1m19s vs. 40s).

As for the OCR-based methods, they are the slowest among all methods, both during training and inference. DocMAE$_{full}$ takes half as long to train an epoch as LayoutLMv3 and reaches even a speed of nearly 6 times in inference. This is mainly because OCR is time-consuming no matter in training or inference.

DocMAE is proved to be a practical model that is well suited for solving document image related tasks by comparing all methods, including both the OCR-free and OCR-based methods. It has better accuracy than DiT while it also has higher efficiency than the OCR-based methods.

5.2.5 Error Analysis

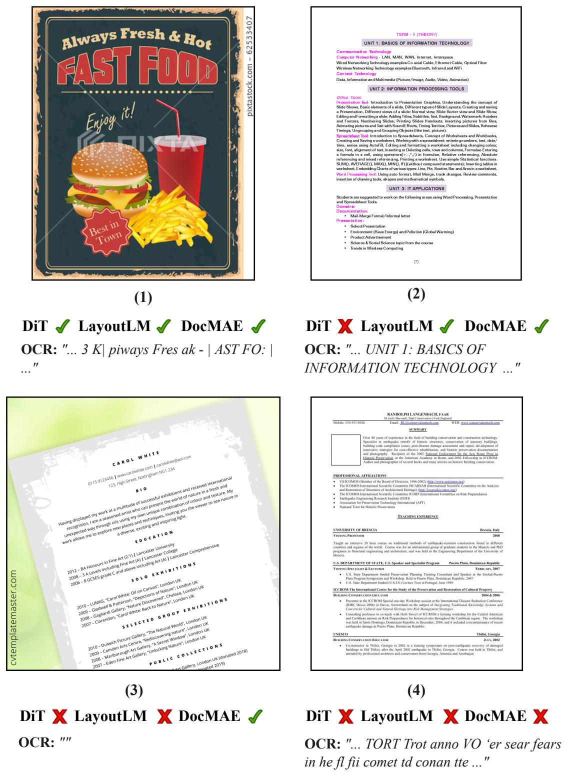

Figure 7: Classification results on the test set with DiT$_{large}$, LayoutLMv3$_{large}$, and DocMAE$_{decoder}$. ✓ indicates correct classification and × indicates incorrect classification. The OCR results come from LayoutLMv3.

To gain an intuitive perception of the features of cases where the model works or where it does not, we performed error analysis for several cases. We chose DiT$_{large}$, LayoutLMv3$_{large}$, and DocMAE$_{decoder}$ to compare, and their results are shown in Figure 7.

In the first case, all three models can classify it correctly. There are apparent objects and keywords in the image. Since the compression of the input image resolution will not lose important information, even the OCR does successfully extract the correct text. In the second case when there is no significant object and full of fine-grained text, due to the small image size, DiT is not able to recognize deep semantic information and just fails. However, in spite of the same image size, since LayoutLMv3 has OCR as a complement input, it can obtain enough meaningful information directly from the OCR text and thus can still classify it correctly. In the third case, because the text is relatively small and skewed, OCR cannot precisely extract the text, making the final classification result of LayoutLMv3 wrong. Those cases prove that DocMAE has a deeper understanding of pixel-based text semantics and is more robust to different text forms, enabling it to classify all three cases correctly. In the fourth case, all three models perform wrong classification. The words in the last image are minimal and blurry, and although humans can still distinguish some of the keywords, it is too difficult for the models.

From the above cases, we can find that OCR is not always so reliable and especially often fails for more complex document images. Our experimental results show that solving directly from pixels is a more direct and practical approach to understanding document content. Meanwhile, for more complex and fuzzy text, DocMAE still has room for improvement compared to human performance.

6 Conclusion

This paper investigated how to better understand the rich semantic content in document images. Given that the previous document image classification datasets mainly focused on document format while ignoring document's text content, we presented a new dataset called DocCT. DocCT is the first dataset to concentrate on the topic classification of document images. The models must analyze fine-grained document content to classify each image under a correct topic. DocCT can facilitate the research related to document image understanding.

Furthermore, we analyzed the shortcomings of previous document image classification models and presented a new self-supervised pre-trained model called DocMAE. The basic structure of DocMAE was borrowed from MAE with an enlarged input image size. Our experimental results showed that a larger image size is essential for understanding semantics by pixels. Meanwhile, compared to models that rely on OCR to obtain semantic text, DocMAE, as a purely pixel-based model, has better robustness, faster training and inference efficiency, and higher classification accuracy than previous methods on DocCT, proving it is possible to process document image semantics without OCR. For future research, we believe that it is necessary to introduce more fine-grained pre-training tasks because at present, DocMAE still has a particular gap compared with humans in understanding small and fuzzy text.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding.

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Nawei Chen and Dorothea Blostein. 2007. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1):1–16.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. 2022. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

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.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking.

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

Ddlc D. Lewis. 2006. Building a test collection for complex document information processing. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Jason Tyler Rolfe. 2016. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

J. Wang, L. Jin, and K. Ding. 2022a. Lilt: A simple yet effective language-independent layout transformer for structured document understanding.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022b. N24news: A new dataset for multimodal news classification. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France. European Language Resources Association.