# VLCDoC: Vision-Language Contrastive Pre-Training Model for Cross-Modal Document Classification

Souhail Bakkali[a,*], Zuheng Ming[a], Mickael Coustaty[a], Marçal Rusiñol[b,c], Oriol Ramos Terrades[c]

[a] *L3i, La Rochelle University, La Rochelle, France*
[b] *AllRead MLT, Barcelona, Spain*
[c] *CVC, Universitat Autonoma de Barcelona, Barcelona, Spain*

**Abstract**

Multimodal learning from document data has achieved great success lately as it allows to pre-train semantically meaningful features as a prior into a learnable downstream task. In this paper, we approach the document classification problem by learning cross-modal representations through language and vision cues, considering intra- and inter-modality relationships. Instead of merging features from different modalities into a joint representation space, the proposed method exploits high-level interactions and learns relevant semantic information from effective attention flows within and across modalities. The proposed learning objective is devised between intra- and inter-modality alignment tasks, where the similarity distribution per task is computed by contracting positive sample pairs while simultaneously contrast-

*Corresponding author at: L3i, La Rochelle University, La Rochelle, France
*Email addresses:* `souhail.bakkali@univ-lr.fr` (Souhail Bakkali), `zuheng.ming@univ-lr.fr` (Zuheng Ming), `mickael.coustaty@univ-lr.fr` (Mickael Coustaty), `marcal@allread.ai,marcal@cvc.uab.cat` (Marçal Rusiñol), `oriolrt@cvc.uab.cat` (Oriol Ramos Terrades)

ing negative ones in the joint representation space. Extensive experiments on public document classification datasets demonstrate the effectiveness and the generality of our model on low-scale and large-scale datasets.
*Keywords:*
Multimodal Document Representation Learning, Document Classification, Contrastive Learning, Self-Attention, Transformers

## 1. Introduction

Document data combines different sensory input modalities such as vision, language, and layout, which allow to extract useful structured information and to learn meaningful representation of its content. These types of inputs are approximated by combining visual-textual information as two coherent and complementary signals that can be further enhanced with layout information. Unlike general images from natural scenes, the extraction of accurate and structured information from the wide variety of document data is very challenging considering their different structural properties through vision, and their textual semantic information through language, as displayed in the Figure 1. Therefore, recent research has started to consider how to leverage and incorporate the relations within those different modalities in a unified network to capture latent information for exploring better yet effective multimodal representations. Such systems have shown their effectiveness in improving multimodal representation learning in a pretrain-then-finetune paradigm, where models are first pre-trained with large-scale data and then

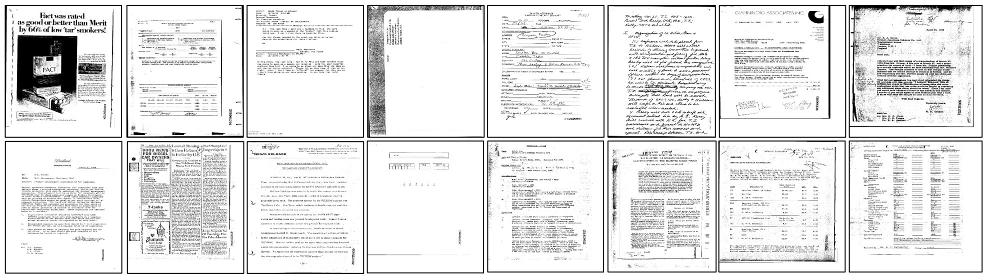


Figure 1: Document samples from the categories of the RVL-CDIP dataset which show the different structural properties of each document in each category. From left to right: *Advertisement, Budget, Email, File folder, Form, Handwritten, Invoice, Letter, Memo, News article, Presentation, Questionnaire, Resume, Scientific publication, Scientific report, Specification*

fine-tuned to each downstream task [1, 2].

Several studies that have been devoted to perform the document classification task, often used shallow cross-modal feature fusion modules to leverage visual-textual features such as naive concatenation, element-wise multiplication, and ensemble methods to extract cross-modal features [3, 4]. Despite being studied extensively, the shortcomings of the preceding cross-modal feature fusion approaches are twofold. First, during inference, the vision-language sample pairs need to be fed to the fusion modules to calculate the prediction scores in order to perform the document classification task, which remains computationally expensive. Second, the existing vision-language modality gap makes it difficult to capture high-level interactions between image regions and text sequences, as the feature representations of the vision-language modalities are usually inconsistent and their distributions span different feature space.

Therefore, the learning process of each modality is independent one from another and fails to capture the non-linearity of image and text sentences within document data based on simple linear operations. In contrast, to embody the idea that better features make better classifiers, a framework based on the pretrain-then-finetune paradigm which allows to learn more general and model-agnostic cross-modal representations is highly required, where the feature projections lead to more compact common representations, by incorporating intra-modality and inter-modality relations from vision and language modalities. The introduced common space is an intermediate that implicitly measures the cross-modal similarities between image and text sequence sample pairs. Intuitively, the multimodality of documents require multimodal reasoning over multimodal inputs. For instance, some types of documents such as handwriting categories are mainly not recognizable by OCR algorithms, which leads to losing textual information, and thus, semantic meaning. Thus, the visual information within the image regions of the document should be strongly emphasized. In the meantime, some type of documents such as file folder category do not contain any visual spatial information,in which case a stronger emphasis on the textual information within the language cues is highly required.

To address the heterogeneity gap and the lack of closer interactions between image regions and text sequences within and across vision-language modalities, we propose a cross-modal contrastive vision-language pre-training model by learning cross-modal representations as a prior in a unified pre-training network. To encourage cross-modal learning, we model intra-modality and inter-modality representations between the cues of the vision-language modalities in the pre-training stage. We design an Inter-Modality Cross-Attention module denoted as (InterMCA) to capture relevant features from image regions and semantic meaning from text sequences. We aim to ensure that features from vision and language modalities map to closer points in the joint embedding space. Nevertheless, existing cross-modal document understanding approaches lack an explicit measure which ensures that similar features from the same modality stay close in the joint embedding space. We assume that if similar features from the same category of each modality map to distant points in the joint embedding space, then the embeddings generated within vision and language modalities will lack semantically enriched information, and thus, will generalize badly for downstream tasks. As a remedy, we introduce intra-modality representation which is carried within an Intra-Modality Self-Attention module denoted as (IntraMSA). This module is devoted to constructing intra-modality relations within each modality according to the self-attention weights of image regions and text sequences.

Moreover, leveraging cross-modal relations through the InterMCA and IntraMSA attention modules require a cross-modal learning objective. In the pre-training stage, we propose to train the network with a combinatorial cross-modal contrastive learning loss, which aims to simultaneously learn visual-textual features that represent document data in a more efficient manner, than direct adoption of a uni-modal contrastive loss for vision or language only modalities. For the downstream application, we run uni-modal inference on top of the generated cross-modal embeddings to perform document classification. The superior performance on three document datasets demonstrates that the proposed cross-modal learning network, denoted as VLCDoC, can lead to learn meaningful cross-modal representations. The main contributions of this work are summarized as follows:

- We design a unified network for cross-modal representation learning. Our network consists of leveraging two flexible extra levels of cross-modal interactions through InterMCA and IntraMSA attention modules, to capture high-level interactions between visual-language cues in document images. The proposed VLCDoC approach shows its superiority over the uni-modal methods.

- We propose a cross-modal contrastive learning objective to further explore the relations between vision and language cues. Comparing to the classic uni-modal contrastive learning, the proposed cross-modal contrastive loss allows to learn and align the feature representations within and across vision-language modalities.

- Under a fair comparison setting, our VLCDoC demonstrates a good generality among vision-language based approaches on the benchmark document datasets, and enables to learn robust and domain-agnostic feature representations for document classification.

- We show that the vision transformer-based architecture used as a backbone of the vision modality in our VLCDoC network can achieve comparable performance when pre-trained on fewer data.

## 2. Related Work

### *2.1. Multimodal Document Understanding*

Deep learning methods have shown great performance in the field of CV and NLP. Specifically, they have been extensively applied to document understanding such as document classification [2, 5], table detection and recognition [6, 7], document visual question answering [8]. Since documents are multimodal, they require multimodal reasoning over multimodal inputs that are mapped into a joint embedding space. Earlier attempts have focused on shallow cross-modal fusion methods to leverage visual, textual, and/or layout information into a joint embedding space [9, 10, 11]. Yang *et al.* [9] proposed a multimodal, fully convolutional network to extract meaningful semantic structures from document images. Based on a graph convolution based model, Liu *et al.* [10] combined textual and visual information presented in visually rich documents to perform entity recognition on document data. Zhang *et al.* [11] proposed a multimodal framework for simultaneous text reading and information extraction in visually rich document for document understanding. By utilizing the graphical property of business documents, Raja *et al.* [6] employed deep neural networks for table structure recognition. Madhav *et al.* [7] utilized an end-to-end trainable cascade deep architecture for table detection in document images. Olivier *et al.* [8] proposed a document retriever model for answering questions on handwritten document image collections.

### 2.2. Multimodal Document Pre-training

Multimodal document pre-training has seen increased attention recently as it allows to train semantically meaningful embeddings as a prior to a learnable downstream task. The mechanisms used to leverage features from document modalities differ one from another. LayoutLMv1 [12] jointly models interactions between text and layout information across document images by adding 2D word position in the language representation to better align the layout information with the semantic representation. LayoutLMv2 [13] leverages vision, language, and layout modalities in a cross-modal pre-training scheme for a better cross-modality interaction. In LayoutLMv3 [14], the authors propose a joint multimodal approach to model the interaction between textual, visual, and layout information in a unified multimodal pre-training network, with different pre-text tasks for a better generality to image-centric and text-centric downstream document AI tasks. Besides, SelfDoc [2] exploits cross-modal learning in the pre-training stage to perform a task-agnostic framework to model information across textual, visual, and layout information modalities without requiring document data annotation. In DocFormer [5], the authors encourage multimodal interaction using a multimodal transformer architecture to perform visual document understanding.

Although these works employ visual, textual, and layout information for document pre-training, and often achieve superior performance, they have several limitations in real-world scenarios: (1) When performing cross-modal document classification during inference, the image-text sample pairs need to be fed to the fusion modules to calculate the prediction scores to classify documents, which remains computationally expensive. (2) To model high-level interactions between image regions and text sequences, in contrast to the previous related works that leverage different modalities into a joint embedding space, align them on the final embedding level, and thus, fail to model fine-grained interactions between the different modalities, our proposed VL-CDoC considers only visual-textual information, and exploits cross-modal representation learning by incorporating intra- and inter-modality relations. Beyond that, we introduce cross-modal contrastive learning as a pre-training objective for effective cross-modality representation learning.

### *2.3. Vision-Language Alignment*

A broad category of pre-training techniques are those that use contrastive losses, which have been used in a wide range of CV applications like image-text similarity, and cross-modal retrieval [15, 16]. Such methods aim at mapping text and images into a common space, where semantic similarity across different modalities can be learned by ranking-based contrastive losses [17, 18, 19]. While dealing with vision-language sample pairs, though individual samples may demonstrate inherent heterogeneity in their content, they are usually coupled with each other based on some higher-level concepts such as their categories. This shared information can be useful in measuring semantics of samples across modalities in a relative manner. Verma *et al.* [20] analyzed the degree of specificity in the semantic content of a sample in the vision modality with respect to semantically similar samples in the language modality. Krishnan *et al.* [21] measured the similarity score between the words distributions across two document images, by detecting patterns of text re-usages across documents written by different individuals irrespective of the minor variations in word forms, word ordering, layout or paraphrasing of the content. Different from the recent research, we propose intra-modality and inter-modality alignment objectives to ensure that samples with semantically similar content stay close in the common representation space, regardless of the modality, to emphasize the interaction and agreement between the visual regions and the semantic meaning of text sequences, as well as to intensify the inner-modality information, by simultaneously preserving the original features and establishing inner-interactions within each modality.

### *2.4. Attention Mechanism*

The attention mechanism has stimulated interest in the domain of deep learning, it was adopted to learn to attend to the most relevant regions of the input space in order to assign different weights to different regions. It was first proposed by Bahdanau *et al.* [22] for neural machine translation. The mechanism is firstly used for machine translation where the most relevant

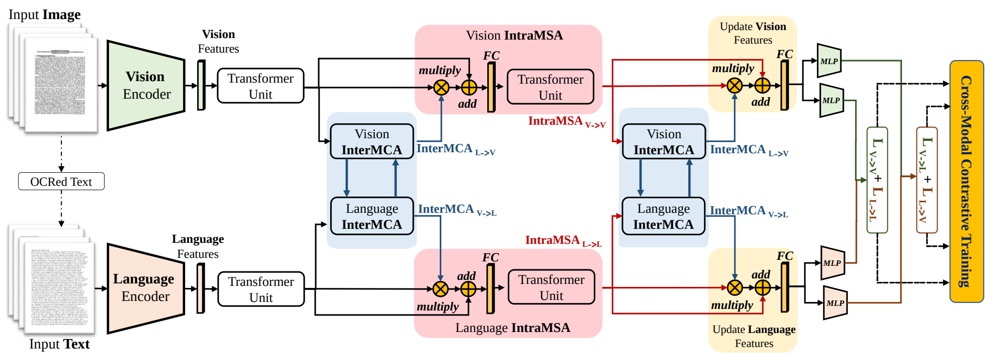



Figure 2: Overview of the proposed cross-modal contrastive learning method. The network is composed of InterMCA and IntraMSA modules with flexible attention mechanisms to learn cross-modal representations in a cross-modal contrastive learning fashion.

words for the output often occur at similar positions in the input sequence. Specifically, self-attention, and co-attention learning has seen increased interest in the field of multimodal vision-language learning such as, document understanding [2, 5], VQA [23, 24], and also image captioning [9, 25], aiming at learning the internal relations in a text sentence or in an image. To model the internal relationships among different modalities, we adopt the contextualized attention mechanism from NLP [26] to improve the location accuracy of a document image region in the vision modality for the desired text sequence in the language modality. Our proposal highlights both the cross-modal co-attention (InterMCA), and internal self-attention (IntraMSA) mechanisms which are integrated in the proposed model, which means that self-attention and co-attention are integrated in the proposed model.

## 3. Methodology

Figure 2 shows the overall architecture of the proposed cross-modal network. VLCDoC is an encoder-only transformer-based architecture trained in an end-to-end fashion. It has two main modalities to perform visual-textual feature extraction. VLCDoC enforces deep multimodal interaction in transformer layers using a cross-modal attention module. The VLCDoC architecture network consists of two main schemes: one contrastive learning branch for cross-modal representation learning, and one cross-entropy learning branch for classifier learning. This feature learning strategy aims to learn a feature space which has the property of intra-class compactness and inter-class separability, while the classifier learning branch is expected to learn a domain-agnostic classifier with less bias based on the discriminative features obtained from the encoder branch.

### *3.1. Model Architecture*

#### *3.1.1. Visual Features*

To extract visual embeddings, we follow the original pre-trained vision transformer architecture ViT-B/16 [27] as a backbone. Let $v_{visn} \in \mathbb{R}^{H \times W \times C}$ be the document image. We reshape it into a sequence of flattened $2D$ patches $v_{visn_p} \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where $(H, W)$ is the resolution of the document image, $C = 3$ is the number of channels, $(P, P)$ is the resolution of each document patch, and $N = HW/P^2$ is the resulting number of patches, which serve as the input sequence length for the transformer encoder. The patches obtained are

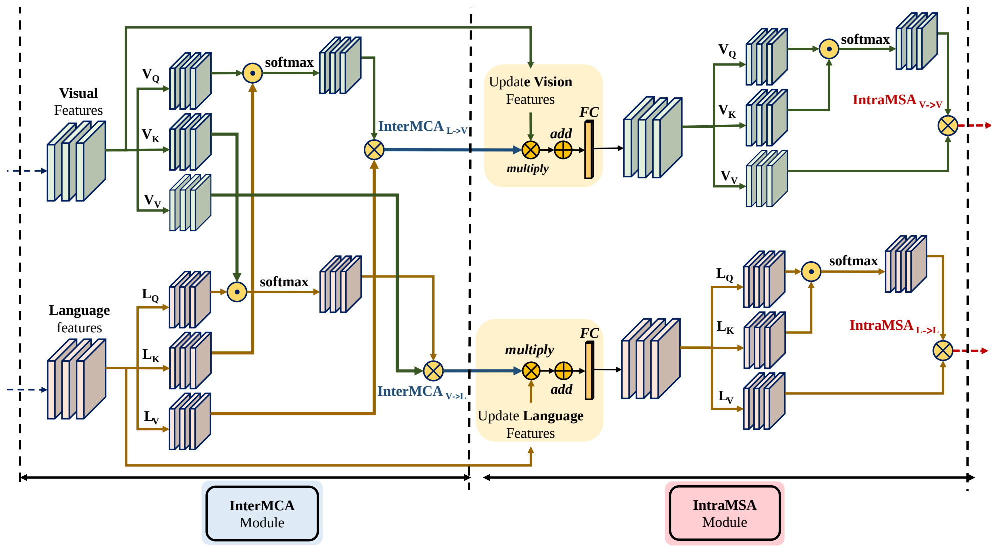



Figure 3: Illustration of the InterMCA and IntraMSA attention modules. The visual-textual features are transformed into query, key, and value vectors. They are jointly leveraged and are further fused to transfer attention flows between modalities to update the original features.

then flattened and mapped to $d$ dimensions as the hidden embedding size. The resulting visual embeddings are then represented as $V = v^i_{visn} \in \mathbb{R}^{d_{visn}}$, where $d_{visn}$ is a $2D$ vector.

### 3.1.2. Textual Features

To extract textual embeddings, we first extract the text $t_{lang}$ within document images via an off-the shelf optical character recognition (OCR) system, *e.g.* Tesseract OCR[1]. The input sequences extracted with the OCR are further fed into the pre-trained BERT$_{Base}$ uncased encoder [28]. The resulting textual embeddings are then represented as $T = t^i_{lang} \in \mathbb{R}^{d_{lang}}$, where $d_{lang}$ is

[1]https://github.com/tesseract-ocr/tesseract

a $2D$ vector of the same size as $d_{visn}$. This way, we ensure that the visual and the textual embeddings are of the same shape.

### *3.2. Cross-Modal Alignment*

In this subsection, we introduce the InterMCA and IntraMSA attention modules that capture intrinsic patterns by modeling the inter-modality and intra-modality relationships for image regions and texts. Specifically, our proposed attention modules are transformer-based architectures as in [26]. It consists of a multi-head self-attention sub-layer, and a position-wise feed-forward sub-layer $f_{FF}$. Meanwhile, residual connections followed by the layer normalization $f_{LN}$ are also applied around each of the two sub-layers. In the multi-head self-attention sub-layer, the attention is calculated $h$ times, making it to be multi-headed. This is achieved by projecting the queries $\mathcal{Q}$, keys $\mathcal{K}$, and values $\mathcal{V}$ $h$ times by using different learnable linear projections.

#### *3.2.1. Inter-Modality Alignment*

The inter-modality cross-attention module InterMCA aims to enhance the cross-modal features by embracing cross-modal interactions across image regions and texts. This module aims to transfer the salient information from one modality to another as illustrated in the Figure 3. Let $\mathbf{V}^l = \{v_1, v_2, ..., v_m\}$, $\mathbf{L}^l = \{l_1, l_2, ..., l_m\}$ be the sets of intermediate visual and textual features at the $l$-th layer of the vision and language modalities respectively, where $v_i \in \mathbb{R}^{1\times d_f}$, $l_i \in \mathbb{R}^{1\times d_f}$, and $\mathbf{V} \in \mathbb{R}^{m\times d_f}$, $\mathbf{L} \in \mathbb{R}^{m\times d_f}$. Note that the visual-textual features have the same dimensional feature vector $d_f$. To accomplish cross-modal interaction, we apply at first dot-product attention to combine the queries of each modality with the keys of the other. The weighted sum of the value of each modality is computed as:

$$\mathbf{InterMCA}_{\mathbf{L}\to\mathbf{V}}(\mathbf{V}^l) = \mathrm{softmax}\left(\frac{\mathcal{Q}_{\mathbf{V}^l}\mathcal{K}_{\mathbf{L}^l}^\top}{\sqrt{d_k}}\right)\mathcal{V}_{\mathbf{L}^l} \tag{1}$$

$$\mathbf{InterMCA}_{\mathbf{V}\to\mathbf{L}}(\mathbf{L}^l) = softmax\left(\frac{\mathcal{Q}_{\mathbf{L}^l}\mathcal{K}_{\mathbf{V}^l}^\top}{\sqrt{d_k}}\right)\mathcal{V}_{\mathbf{V}^l} \tag{2}$$

In this way, we emphasize the interaction and agreement between the visual regions and the semantic meaning of texts. The attention weights are then sent into the feed-forward sub-layer. Finally, we get the output features of the next layer of the vision modality $\mathbf{V}^{l+1}$ computed as:

$$\mathbf{V}^l_{Att} = f_{LN_{\mathbf{V}}}(\mathbf{InterMCA}_{\mathbf{L}\to\mathbf{V}}(\mathbf{V}^l) + \mathbf{V}^l) \tag{3}$$

$$\mathbf{V}^{l+1} = f_{LN_{\mathbf{V}}}(f_{FF}(\mathbf{V}^l_{Att}) + \mathbf{V}^l_{Att}) \tag{4}$$

Similarly, the output features $\mathbf{L}^{l+1}$ of the language modality are computed:

$$\mathbf{L}^l_{Att} = f_{LN_{\mathbf{L}}}(\mathbf{InterMCA}_{\mathbf{V}\to\mathbf{L}}(\mathbf{L}^l) + \mathbf{L}^l) \tag{5}$$

$$\mathbf{L}^{l+1} = f_{LN_{\mathbf{L}}}(f_{FF}(\mathbf{L}^l_{Att}) + \mathbf{L}^l_{Att}) \tag{6}$$

Further, the outputs of each vision and language InterMCA modules are subsequently fed into the vision and language IntraMSA modules.

*3.2.2. Intra-Modality Alignment*

The IntraMSA attention module illustrated in the Figure 3, aims to update the vision and language information and to capture inner-modality attention weights. For each modality, the information is updated according to a feature fusion scheme. At first, we perform element-wise product to the attention flow $\mathbf{V}^{l+1}$ with the the visual region features $\mathbf{V}^{l}$, then after a residual connection, features are fused by a linear additive function to yield the final updated visual information. To keep the dimension of the updated information consistent, a fully connected $f_{FC}$ layer is employed. The updated textual information is computed likewise, following the equations:

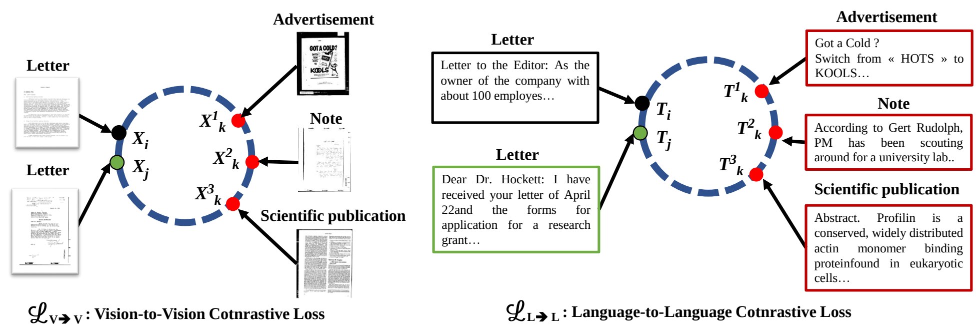
$$\hat{\mathbf{V}} = f_{FC}((\mathbf{V}^{l+1} \odot \mathbf{V}^{l}) + \mathbf{V}^{l}) \tag{7}$$

$$\hat{\mathbf{L}} = f_{FC}((\mathbf{L}^{l+1} \odot \mathbf{L}^{l}) + \mathbf{L}^{l}) \tag{8}$$

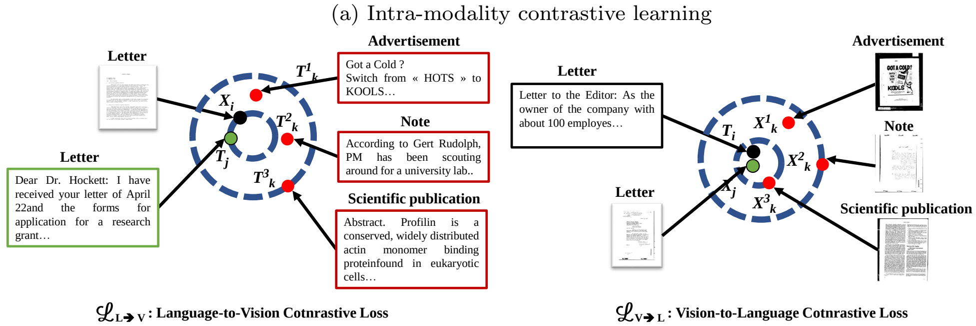
After updating original features based on cross-modal interactions, these features are fed into the transformer unit to intensify the inner-modality information, to preserve the original features and to establish inner-interactions simultaneously. Following the Equations 1, 2, we have:

$$\mathbf{IntraMSA}_{\mathbf{V}\rightarrow\mathbf{V}} = \mathrm{softmax}\left(\frac{\mathcal{Q}_{\hat{\mathbf{V}}^{l}}\mathcal{K}_{\hat{\mathbf{V}}^{l}}^{\top}}{\sqrt{d_k}}\right)\mathcal{V}_{\hat{\mathbf{V}}^{l}} \tag{9}$$

$$\mathbf{IntraMSA}_{\mathbf{L}\rightarrow\mathbf{L}} = \mathrm{softmax}\left(\frac{\mathcal{Q}_{\hat{\mathbf{L}}^{l}}\mathcal{K}_{\hat{\mathbf{L}}^{l}}^{\top}}{\sqrt{d_k}}\right)\mathcal{V}_{\hat{\mathbf{L}}^{l}} \tag{10}$$




(a) Intra-modality contrastive learning




(b) Inter-modality contrastive learning

Figure 4: The proposed cross-modal contrastive learning objective

These two modules can be stacked repeatedly, enabling to explore further latent intra- and inter-modality alignments between image regions and texts.

### *3.3. Cross-Modal Contrastive Learning*

We design a vision-language contrastive loss to force samples from language and vision that are semantically related to be closer. Besides, a projection head is implemented on top of the IntraMSA and InterMCA modules to map the image and text representations into a vector representation so that the two training schemes do not interfere with each other. The projection head is implemented as a nonlinear multiple-layer perceptron (MLP) with one hidden layer, as it is more suitable for contrastive learning [29]. Then,

$L_2$ normalization is applied to the visual-textual embeddings so that the inner product between features can be used as distance measurements. In the following parts, we denote cross-modal contrastive learning as CrossCL.

### 3.3.1. Intra-Modality and Inter-Modality Contrastive Learning

Let $\{\mathbf{x}_i^+\} = \{x_j | y_j = y_i, i \neq j\}$, $\{t_i^+\} = \{t_j | y_j = y_i, i \neq j\}$ be the sets of all positive samples from the same class of an anchor image $x_i$ and an anchor text $t_i$ respectively, and $\{\mathbf{x}_i^-\} = \{x_j | y_j \neq y_i, \}$, $\{t_i^-\} = \{t_j | y_j \neq y_i\}$ be the sets of the remaining negative samples from other classes within the minibatch N. Not only the pairs ($\mathbf{x}_i$, $\mathbf{x}_j$), ($\mathbf{t}_i$, $\mathbf{t}_j$) from the same modality should be mapped to a close location in the joint embedding space (intra-modality), but also similar samples $\mathbf{x}_i$ and $\mathbf{t}_j$ should be mapped in close proximity (inter-modality). Therefore, the vision modality loss shown on the left of the Figures 4a, 4b is computed as:

$$\mathcal{L}_V = \sum_{i=1}^{N} \mathcal{L}_{V \to V}(\mathbf{x}_i) + \sum_{i=1}^{N} \mathcal{L}_{L \to V}(\mathbf{x}_i) \tag{11}$$

$$\mathcal{L}_{V \to V}(\mathbf{x}_i) = \underbrace{\frac{-1}{|\{\mathbf{x}_i^+\}|} \sum_{\mathbf{x}_j \in \{\mathbf{x}_i^+\}} \log \frac{\exp(\mathbf{x}_i \cdot \mathbf{x}_j / \tau)}{\sum_{\mathbf{x}_k, k \neq i} \exp(\mathbf{x}_i \cdot \mathbf{x}_k / \tau)}}_{\text{Intra modality vision loss}} \tag{12}$$

$$\mathcal{L}_{L \to V}(\mathbf{x}_i) = \underbrace{\frac{-1}{|\{\mathbf{t}_i^+\}|} \sum_{\mathbf{t}_j \in \{\mathbf{t}_i^+\}} \log \frac{\exp(\mathbf{x}_i \cdot \mathbf{t}_j / \tau)}{\sum_{\mathbf{t}_k, k \neq i} \exp(\mathbf{x}_i \cdot \mathbf{t}_k / \tau)}}_{\text{Inter modality vision loss}} \tag{13}$$

where $\cdot$ computes similarity scores between example pairs and $\tau$ is a scalar temperature hyper-parameter. N is the minibatch size, $|\{\mathbf{x}_i^+\}|$ and $|\{\mathbf{t}_i^+\}|$ denote the number of positive samples of anchors $\mathbf{x}_i$ and $\mathbf{t}_i$ respectively. Similarly, the language modality loss shown on the right of the Figures 4a, 4b is computed as:

$$\mathcal{L}_L = \sum_{i=1}^{N} \mathcal{L}_{L\to L}(\mathbf{t}_i) + \sum_{i=1}^{N} \mathcal{L}_{V\to L}(\mathbf{t}_i) \tag{14}$$

Therefore, the learning objective is based on four contrastive components including ($V \to V$, $L \to V$, $L \to L$, $V \to L$) alignments, is computed as:

$$\mathcal{L}_{CrossCL} = \mathcal{L}_{V\to V} + \lambda\mathcal{L}_{L\to V} + \mathcal{L}_{L\to L} + \lambda\mathcal{L}_{V\to L} \tag{15}$$

where $\lambda$ is a hyper-parameter to control inter-modality alignment.

## 4. Experiments

In this section, we evaluate the effectiveness of the proposed method on low-scale and large-scale document classification datasets.

### *4.1. Datasets*

**RVL-CDIP** The RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset is a subset of the IIT-CDIP Test Collection presented in [30]. It consists of gray-scale labeled documents split into 16 classes. The dataset is split into 320K training documents, 40K documents

documents for validation and test sets. For notation simplicity, we denote the dataset as RVL-CDIP.

**Tobacco-3482** The Tobacco-3482 dataset is a smaller sample containing 3482 gray-scale document images presented in [31]. This dataset is formed by documents belonging to 10 classes not uniformly distributed. For simplicity, we denote the dataset as Tobacco.

**NIST Special Database 6** The Nist-tax form [32] dataset is composed of structured forms of 5595 pages of binary, black-and-white images of synthesized documents containing hand-print and split into 20 different tax forms. For simplicity, we denote the dataset as NIST.

### *4.2. Experimental Settings*

The proposed VLCDoC method is implemented in Tensorflow with 4 NVIDIA GeForce 12Gb RTX 2080Ti GPU. For the vision modality, documents are resized into a fixed size of (H, W)=(224, 224). The image region feature vector extracted by the ViT-B/16 backbone is of $d_{visn}$=(197, 768). The final vision representation which is fed into the projection head is of dimension $d$=768. As for the textual data, we tokenize the plain text $t_{lang}$ using a word-peace tokenizer to get $t_{tok}$. Each input sequence is expected to start with a $[CLS]$ token, and should end with a $[SEP]$ token. The $t_{tok}$ is then represented as: $t_{tok} = [CLS], t_{tok_1}, t_{tok_2}, ..., t_{tok_n}, [SEP]$, where $n$=197 is the maximum sequence length. For each document, if $n$ >197, the input sequence is truncated so that it fits the desired length. For sequences that

Table 1: Ablation study on VLCDoC on cross-modality attention components, pre-trained on Tobacco dataset

| Pre-training setting | IntraMSA | InterMCA | #Parameters | Accuracy(%) |
|---|---|---|---|---|
| *-w/o language modality* | | | | |
| | | | 198M | 85.71 |
| | √ | | 201M | 86.66 |
| | | √ | 209M | 87.20 |
| | √ | √ | 217M | **90.94** |
| *-w/o vision modality* | | | | |
| | | | 198M | 86.01 |
| | √ | | 201M | 86.31 |
| | | √ | 209M | 87.50 |
| | √ | √ | 217M | **90.62** |

are shorter than $n$ <197, they are padded until they are $n$ =197 long. In the pre-training phase, the model is trained using AdamW optimizer with a learning rate of 2e-5, linear warmup ratio to 0.1 and a linear decay. We set the batch size to 64 and we use the pre-trained weights of both ViT-B/16 and BERT$_{Base}$ uncased backbones. We conduct pre-training for 100 epochs for the RVL-CDIP and Tobacco datasets. We fine-tune our network on 50 epochs for all datasets, we use Adam optimizer with learning rate of 5e-5. For Tobacco and NIST datasets, we split the original sets to 80% for training, and 10% for validation and test. The temperature parameter $\tau$ is set to 0.1, and $\lambda$ is set to 0.5. Note that we didn't use any type of data augmentation during pre-training, and we kept the OCRed text as is without any pre- or post-processing. Note that the InterMCA and IntraMSA modules in our method are flexibly stacked two times to enhance the modeling of

Table 2: Top-1 accuracy (%) comparison results of our proposed CrossCL loss against the SCL loss on Tobacco dataset

| Model | Modality | CrossCL(%) | SCL(%) |
|---|---|---|---|
| **VLCDoC** | Vision | **90.94** | 89.88 |
| | Language | **90.62** | 89.29 |

inter-modality and intra-modality relations during pre-training. We split the query, key, and value vectors of the visual features and textual features into four heads and concatenate the results in different sub-spaces.

### *4.3. Ablation Study*

We conduct ablation studies to characterize our VLCDoC network on the low-scale Tobacco dataset. We analyze the following contributions of: i) validating the effectiveness of the proposed InterMCA and IntraMSA attention modules in learning generic cross-modal representations, ii) investigating whether contrastive learning enhances the cross-modal representations, resulting in performance gain in terms of classification accuracy, iii) illustrating the generality and robustness of the proposed VLCDoC network.

#### *4.3.1. Effects of Attention Mechanisms*

To investigate the effectiveness of the attention mechanisms used in our VLCDoC model, we evaluate the performance of the learned cross-modal representations w/ and w/o the attention modules. Note that the evaluation protocol is uni-modal based. At first, we consider the scheme where the vision and language modalities are pre-trained independently. In Table 1, we ob-

Table 3: Cross-dataset test on datasets with different size and document types. Tob, RVL, and Nist denote Tobacco, RVL-CDIP, and Nist-tax form benchmark datasets. Tob $\rightarrow$ RVL denotes pre-train on Tobacco, and test on RVL-CDIP.

| Model | Accuracy (%) | | |
|---|---|---|---|
| | Tob $\rightarrow$ RVL | RVL $\rightarrow$ Tob | RVL $\rightarrow$ Nist |
| *w/o language modality* | | | |
| - EAML [33] | 78.89 | 84.82 | - |
| **- VLCDoC** | **79.04** | **89.73** | **99.99** |
| *w/o vision modality* | | | |
| - EAML [33] | 79.06 | 83.72 | - |
| **- VLCDoC** | **81.96** | **89.88** | **99.99** |

serve a significant drop to 85.71%, and 86.01% in classification performance when removing both attention mechanisms in the vision and language modalities respectively. When removing only the InterMCA module, we see that our model manages to improve slightly the performance of both modalities to 86.66% and 86.31% for the vision-language modalities. Further, removing the IntraMSA and keeping only the InterMCA module enables multimodal pre-training in an end-to-end fashion. The reported results in Table 1 show that our model gains in performance, and achieves the best performance with 90.94%, 90.62% top-1 accuracy for the vision and language modalities.

The improvement of the classification accuracy is attributed to the flexible attention flows adopted in both the InterMCA and IntraMSA modules, which have shown their effectiveness and capability to enhance vision-language relations by capturing the relevant semantic information of images and sentences. The results demonstrate the effectiveness of cross-modal learning and the importance of both attention modules in learning more effective cross-modal representations during the pre-training stage.

#### *4.3.2. Effects of Cross-Contrastive Learning*

The Cross-modal Contrastive Loss (CrossCL) contains two components: intra- and inter-modality alignments. We show the effects of CrossCL on the proposed method against the standard supervised contrastive learning (SCL) loss. Table 2 shows that the CrossCL loss has a positive impact on the results. The VLCDoC with CrossCL loss yields the best performance gain compared to VLCDoC with the SCL loss. This indicates the importance of CrossCL by enforcing the compactness of intra-class representations (intra-modality), while separating inter-class features by contrasting positive and negative sample pairs within and across each modality. Note that, as described in Equation 15, the CrossCL can be vision cue-based or language cue-based, thus we have two different CrossCL presented in Table 2.

#### *4.3.3. Cross-Dataset Test*

To illustrate the generality and the robustness of the learned cross-modal features, we validate our VLCDoC model on document classification datasets with different size and document types. We refer as the cross-dataset test to the process of pre-training our VLCDoC on dataset $\boldsymbol{A}$, and fine-tune it and test it on dataset $\boldsymbol{B}$. The motivation behind is to confirm whether our model displays a good generality in terms of the document classification task. Since there is no publicly available cross-document datasets for this specific task, we evaluate the ability of our model to perform document classification on a new set of documents that had not been seen by our model during the pre-training phase. For example, as denoted in the Table 3, which refers to the cross-dataset test, RVL-CDIP→Tobacco denotes that the pre-training stage is firstly conducted on the RVL-CDIP dataset, then the fine-tuning stage of the previously pre-trained model is conducted on the Tobacco dataset. Finally, the test phase is conducted on the Tobacco dataset as well. Note that during the fine-tuning stage, we only train linear classifiers on the top of the final embeddings of the vision and language modalities of our pre-trained model, with the parameters of the rest of the layers freezed. Thus, even though the document categories are different between the dataset $\boldsymbol{A}$ used for pre-training and test dataset $\boldsymbol{B}$ used for fine-tuning and test, we can still evaluate our model on dataset $\boldsymbol{B}$. The results confirm that our approach leads to a model with a better generality compared to prior works.

As such, we compare our model with the related work EAML [33]. We first pre-train the model on Tobacco dataset, then we conduct fine-tuning and test on the RVL-CDIP dataset. The reported results in Table 3 show that we slightly outperform EAML on both vision and language modalities. Even-though EAML is an ensemble network trained with a different setting, based on vision, language, and fusion modalities, the results confirm that our model benefits from cross-modal pre-training with small amount of document data, achieving better performance with only vision and language modalities. Following similar protocol, we pre-train our encoder on RVL-

CDIP, and then conduct fine-tuning and test on Tobacco and NIST datasets with fewer document data. We clearly see that our model outperforms the work EAML with a significant margin of 4.91% and of 6.16% for vision and language modalities respectively. As for NIST dataset, the results achieve 99.99% classification accuracy for both modalities. These results demonstrate that our model displays a good generality which enables to learn a robust and domain-agnostic feature representation for classifying documents with different document types and document data size.

### *4.4. Results*

The comparison between the proposed VLCDoC network and existing methods on the large-scale RVL-CDIP document classification dataset is presented in Table 4. The compared methods cover various training strategies with different modalities used to perform document classification. These methods include vision-only, language-only, vision-language, and vision-language-layout methods. Although our VLCDoC network learns feature space with vision and language cues, it only uses uni-modality (either vision or language) to classify document during inference. In Table 4, we can see that our VLCDoC model achieves the best performance with 92.64% and 91.37% of top-1 accuracy for using the vision or language modality respectively even compared to the methods that use the fusion of visual and language modalities. Therefore, the results reported demonstrate that our proposed approach outperforms all the methods that do not require any supplementary information

Table 4: Top-1 accuracy (%) comparison results of different document classification methods evaluated on the of RVL-CDIP dataset. V+L denotes vision+language modalities

| Method | Pre-Train Data | Accuracy(%) | #Params |
|---|---|---|---|
| *vision methods* | | | |
| VGG-16 [34] | 320k | 90.31 | 138M |
| ResNet-50 [34] | 320k | 91.13 | - |
| Ensemble [35] | 320k | 92.21 | - |
| DiT$_{Base}$ [36] | 320k | 92.11 | 87M |
| *(language+layout) methods* | | | |
| BERT$_{Base}$ [28] | - | 89.81 | 110M |
| RoBERTa$_{Base}$ [37] | - | 90.06 | 125M |
| LayoutLM$_{Base}$ [12] | 11M | 91.78 | 113M |
| *(vision+language) methods* | | | |
| w/o language | | | |
| - Multimodal [38] | 320k | 89.1 | - |
| - Ensemble [39] | 320k | 91.45 | - |
| - EAML [33] | 320k | 90.81 | - |
| w/o vision | | | |
| - Multimodal [38] | 320k | 74.6 | - |
| - Ensemble [39] | 320k | 82.23 | - |
| - EAML [33] | 320k | 88.80 | - |
| **VLCDoC (V+L)** w/o language | 320k | **92.64** | 217M |
| **VLCDoC (V+L)** w/o vision | 320k | **91.37** | 217M |
| *(vision+language+layout) methods* | | | |
| SelfDoc [2] | 320k | 93.81 | - |
| LayoutLM$_{Base}$ [12] | 11M | 94.42 | 160M |
| TILT$_{Base}$ [40] | 1M | 95.25 | 230M |
| LayoutLMv2$_{Base}$ [13] | 11M | 95.25 | 200M |
| LayoutLMv3$_{Base}$ [14] | 11M | 95.44 | 133M |
| DocFormer$_{Base}$ [5] | 5M | 96.17 | 183M |

such as layout information as used in the prior related works [12, 13, 14, 2, 5]. Meanwhile, it achieves competitive results against the methods that include layout information in the pre-training setting.

## 5. Conclusion and Future Work

In this paper, we have proposed a novel cross-modal representation learning model for document classification, which models the intra- and inter-modality relations between vision-language cues via cross-modal contrastive learning. We have introduced InterMCA and IntraMSA attention mechanisms which incorporate visual-textual features to further improve the cross-modal representations. We have performed a detailed analysis and evaluation on each module, demonstrating the suitability of the proposed approach. We have demonstrated a good generality of our multimodal transformer-based model to the document classification task, enabling to classify documents in different domains. We will push forward two research lines for the future. On the one hand, we will carry on further research on the integration of a third layout modality in our transformer-based multimodal model. As prior related works rely mainly on the three vision, language, and layout modalities to extract better cross-modality relations, we would like to propose a better solution for layout integration in our vision-language model. On the other hand, we would like to explore new pre-text task strategies to improve document understanding in a pretrain-to-finetune paradigm. Thus, we will further tune our model for different downstream applications related to document AI with more challenging heterogeneous data.

## 6. Acknowledgements

This work has been co-funded by the French National Research Agency (ANR), and partially supported by the Spanish projects RTI2018-095645-B-C21, the Catalan project 2017-SGR-1783 and the CERCA Program / Generalitat de Catalunya.

## References

[1] H. Bao, L. Dong, F. Wei, Beit: Bert pre-training of image transformers, ArXiv abs/2106.08254 (2021).

[2] P. Li, J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, V. Manjunatha, H. Liu, Selfdoc: Self-supervised document representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5652–5660.

[3] S. Bakkali, Z. Ming, M. Coustaty, M. Rusiñol, Visual and textual deep feature fusion for document image classification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 2394–2403.

[4] S. Sierra, F. A. González, Combining textual and visual representations for multimodal author profiling, Working Notes Papers of the CLEF 2125 (2018) 219–228.

[5] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, R. Manmatha, Docformer: End-to-end transformer for document understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 993–1003.

[6] S. Raja, A. Mondal, C. Jawahar, Visual understanding of complex table structures from document images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2299–2308.

[7] M. Agarwal, A. Mondal, C. Jawahar, Cdec-net: Composite deformable cascade network for table detection in document images, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 9491–9498.

[8] O. Tüselmann, F. Müller, F. Wolf, G. A. Fink, Recognition-free question answering on handwritten document collections, arXiv preprint arXiv:2202.06080 (2022).

[9] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, C. Lee Giles, Learning to extract semantic structure from documents using multimodal fully convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5315–5324.

[10] X. Liu, F. Gao, Q. Zhang, H. Zhao, Graph convolution for multimodal information extraction from visually rich documents, in: NAACL, 2019.

[11] P. Zhang, Y. Xu, Z. Cheng, S. Pu, J. Lu, L. Qiao, Y. Niu, F. Wu, Trie: end-to-end text reading and information extraction for document understanding, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1413–1422.

[12] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, LayoutLM: Pre-training of text and layout for document image understanding, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Jul 2020).

[13] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. A. F. Florêncio, C. Zhang, W. Che, M. Zhang, L. Zhou, Layoutlmv2: Multi-modal pre-training for visually-rich document understanding, ArXiv abs/2012.14740 (2021).

[14] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, Layoutlmv3: Pre-training for document ai with unified text and image masking, ArXiv abs/2204.08387 (2022).

[15] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, Y. Yang, Cross-modal contrastive learning for text-to-image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 833–842.

[16] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, B. Faieta, Multimodal contrastive training for visual representation learning,

in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6995–7004.

[17] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32 (2019).

[18] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, ArXiv abs/1908.03557 (2019).

[19] J. Gu, J. Kuen, S. Joty, J. Cai, V. Morariu, H. Zhao, T. Sun, Self-supervised relationship probing, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 1841–1853.

[20] Y. Verma, A. Jha, C. Jawahar, Cross-specificity: modelling data semantics for cross-modal matching and retrieval, International journal of multimedia information retrieval 7 (2) (2018) 139–146.

[21] P. Krishnan, C. Jawahar, Matching handwritten document images, in: European Conference on Computer Vision, Springer, 2016, pp. 766–782.

[22] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, CoRR abs/1409.0473 (2015).

[23] J. Wu, J. Du, F. Wang, C. Yang, X. Jiang, J. Hu, B. Yin, J. Zhang,

L. Dai, A multimodal attention fusion network with a dynamic vocabulary for textvqa, Pattern Recognition 122 (2022) 108214.

[24] Y. Liu, X. Zhang, Q. Zhang, C. Li, F. Huang, X. Tang, Z. Li, Dual self-attention with co-attention networks for visual question answering, Pattern Recognition 117 (2021) 107956.

[25] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

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

[28] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[29] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for

contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[30] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 991–995.

[31] J. Kumar, P. Ye, D. Doermann, Structural similarity for document image classification and retrieval, Pattern Recognition Letters 43 (2014) 119–126.

[32] D. Dimmick, M. Garris, C. Wilson, Nist structured forms reference set of binary images (sfrs), NIST Special Database 2 (1991).

[33] S. Bakkali, Z. Ming, M. Coustaty, M. Rusiñol, EAML: ensemble self-attention-based mutual learning network for document image classification, International Journal on Document Analysis and Recognition (IJDAR) 24 (09 2021).

[34] M. Z. Afzal, A. Kölsch, S. Ahmed, M. Liwicki, Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, IEEE, 2017, pp. 883–888.

[35] A. Das, S. Roy, U. Bhattacharya, S. K. Parui, Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks, in: 2018 24th international conference on pattern recognition (ICPR), IEEE, 2018, pp. 3180–3185.

[36] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, F. Wei, Dit: Self-supervised pretraining for document image transformer, ArXiv abs/2203.02378 (2022).

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach (2019).

[38] N. Audebert, C. Herold, K. Slimani, C. Vidal, Multimodal deep networks for text and image-based document classification, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 427–443.

[39] T. Dauphinee, N. Patel, M. M. Rashidi, Modular multimodal architecture for document classification, ArXiv abs/1912.04376 (2019).

[40] R. Tito, M. Mathew, C. V. Jawahar, E. Valveny, D. Karatzas, Icdar 2021 competition on document visual question answering, in: J. Lladós, D. Lopresti, S. Uchida (Eds.), Document Analysis and Recognition – ICDAR 2021, Springer International Publishing, 2021, pp. 635–649.