DocumentCLIP: Linking Figures and Main Body Text in Reflowed Documents

Fuxiao Liu
University of Maryland, College Park
fl3es@umd.edu

Hao Tan
Adobe Research
hatan@adobe.com

Chris Tensmeyer
Adobe Research
tensmeye@adobe.com

Abstract

Vision-language pretraining models have achieved great success in supporting multimedia applications by understanding the alignments between images and text. While existing vision-language pretraining models primarily focus on understanding single image associated with a single piece of text, they often ignore the alignment at the intra-document level, consisting of multiple sentences with multiple images. In this work, we propose DocumentCLIP, a salience-aware contrastive learning framework to enforce vision-language pretraining models to comprehend the interaction between images and longer text within documents. Our model is beneficial for the real-world multimodal document understanding like news article, magazines, product descriptions, which contain linguistically and visually richer content. To the best of our knowledge, we are the first to explore multimodal intra-document links by contrastive learning. In addition, we collect a large Wikipedia dataset for pretraining, which provides various topics and structures. Experiments show DocumentCLIP not only outperforms the state-of-the-art baselines in the supervised setting, but also achieves the best zero-shot performance in the wild after human evaluation. Our code is available at https://github.com/FuxiaoLiu/DocumentCLIP.

1. Introduction

Images and text act as natural complements on the modern web. A number of works [28, 13, 16, 10, 1, 9, 23, 44] have been proposed to interpreting an image with the corresponding short text such as a caption or question. However, real-world media such as news articles, Wikipedia pages, magazines, product descriptions consist of multiple sentences with multiple images. Algorithms that identify document-internal connections between specific images and specific unit of text could have the long-term promise. For example, alt-text for vision-impaired users could be produced automatically via intra-document retrieval. Additionally, by designing a user interface hinting the readers that there is a figure associated with a part of the text upon users's action, it can not only help users to understand the whole document, but also make it easy and comfortable to read on smart phones, where the font and button is small.

This intra-document setting introduces several challenges compared to conventional cross-model retrieval task with a single sentence and caption [26, 17, 27, 46]. More specifically, since the images are to be incorporated in the same article, their contents should be consistent according to the theme of the article, thus making disambiguation more difficult than in the usual one-image/one-sentence case (Figure 1). Additionally, a sentence in the longer document may correspond to multiple images or no related images. Another challenge is that it requires considering longer texts with linguistically richer content and the relations between them, like the structure and layout information. The previous relevant works [28, 13, 38, 11, 14, 40] trained to align short literal text descriptions of images with the image failed to encode the structure features. Document structure analysis models [5, 4, 37] process pages one by one, but the figures may relate to text located on other pages.

To address the aforementioned challenges, we propose *DocumentCLIP*, a simple yet effective pretraining learning method of intra-document vision and language understanding tasks. Inspired by the BERT model [24], where input textual information is mainly represented by text embeddings and position embeddings, *DocumentCLIP* further adds two types of input embeddings: (1) an entity embedding that indicates which section sharing more common entities. This is because the images/caption pairs can be the literal visualization of entities mentioned in document sentences. (2) a section position embedding that denotes the relative position within a document to encode the layout and structure information. In addition, we introduce the salient sentence extraction strategy to extract the important information from the long section. To effectively transfer event knowledge across modalities, we employ early-fusion method to aggregate images and caption by [12] before the layout transformer in Figure 2. Unlike the state-of-the-art vision-language pretraining model CLIP [28], we optimize a salience-aware contrastive learning objective between intra-document components: images, captions, section text. In order to train robust

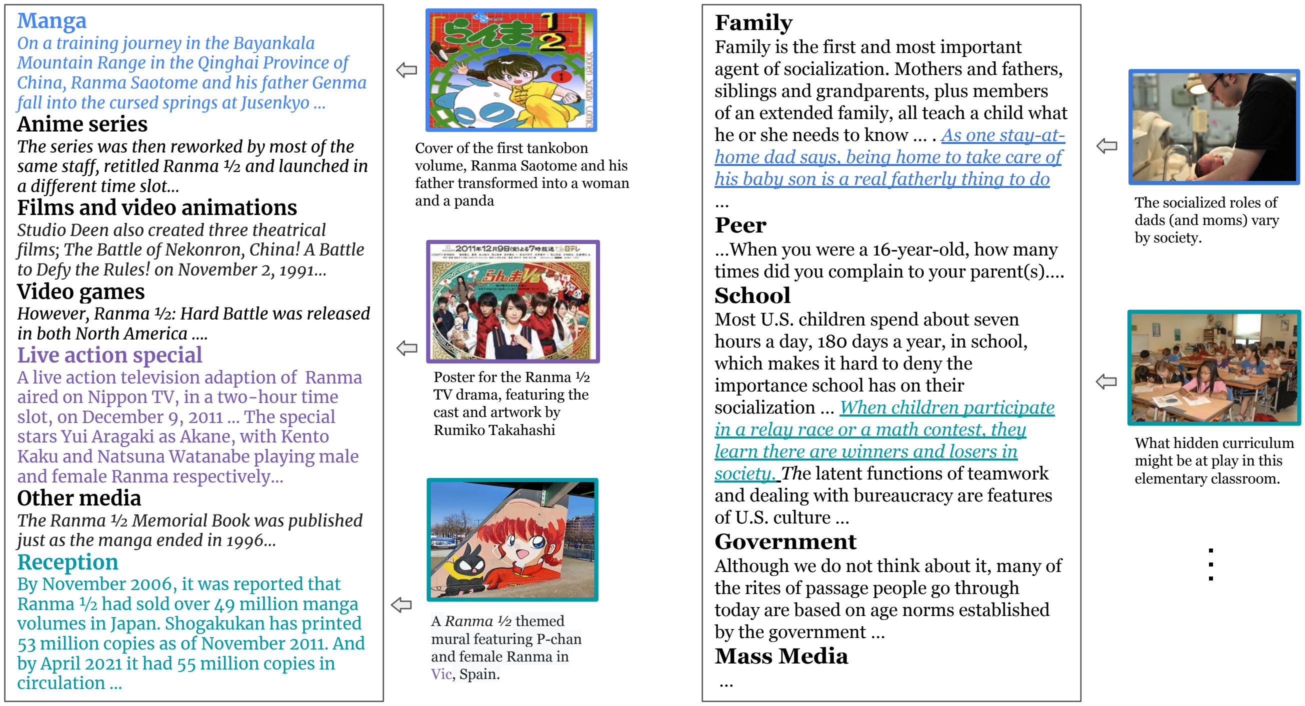

Figure 1: Illustration of our task. Given an article, our model extracts the most relevant text to the image/caption pair. (Left): The document is from our pretraining *Wikipedia dataset*, which has the section to image/caption pair links as the groundtruth label. We leverage this weak supervision to explore multimodal intra-document interaction by contrastive learning. (Right): An example is from *Open Textbook dataset* in the wild. The highlights are the most relevant sentences extracted by our model.

representations capable of discriminating subtle differences between different sentences in the same document, we not only have normal negatives by replacing groundtruth section text with others from the same document, but also propose to generate hard negatives by manipulating the image and caption pairs by switching one of them.

In this work, we pretrain and evaluate *DocumentCLIP* on the Wikipedia articles, which have natural correspondences between the text sections and image/caption pairs in its HTML sources and describe various topics. Also, our evaluations focus on zero-shot settings, since it is crucial to understand real-world documents with various topics and sources. Examples are shown in Figure 1. Experiment results show that *DocumentCLIP* significantly outperforms baselines and meet users' requirement in the wild[1]. The contributions of this paper are summarized as follows:

- We introduce a novel contrastive learning framework with layout information, multimodal interaction, novel section encoding strategy, salience-aware contrastive loss and hard negative samples.
- We collect a large visual article dataset from Wikipedia with various topics, containing 66k articles, 320k images/caption pairs.
- Our model outperforms baselines on the intra-document understanding task in both the supervised and zero-shot settings after human evaluation.

2. *DocumentCLIP*

2.1. Problem Statement

We assume that the structure of the document has been parsed, so that we can apriori identify figure images and their captions (if present), and segment the main body text into a sequence of sections, paragraphs, and sentences. The granularity of the associated text (section vs sentence) can be set based on the use case and/or training data available. In our prototype, we primarily used Wikipedia articles, whose source markup associates figure-captions with sections. Additionally, by leveraging this weak supervision with the section ground truth label, our model is often able to retrieve the most relevant sentence to the associated image and caption.

We define relevance based on the following criteria. (1) The images are the literal visualization of the entities mentioned in the document. (2) Text explicitly refers to the images. (3) The images provide evidences for the text claims.

[1]https://open.umn.edu/opentextbooks/

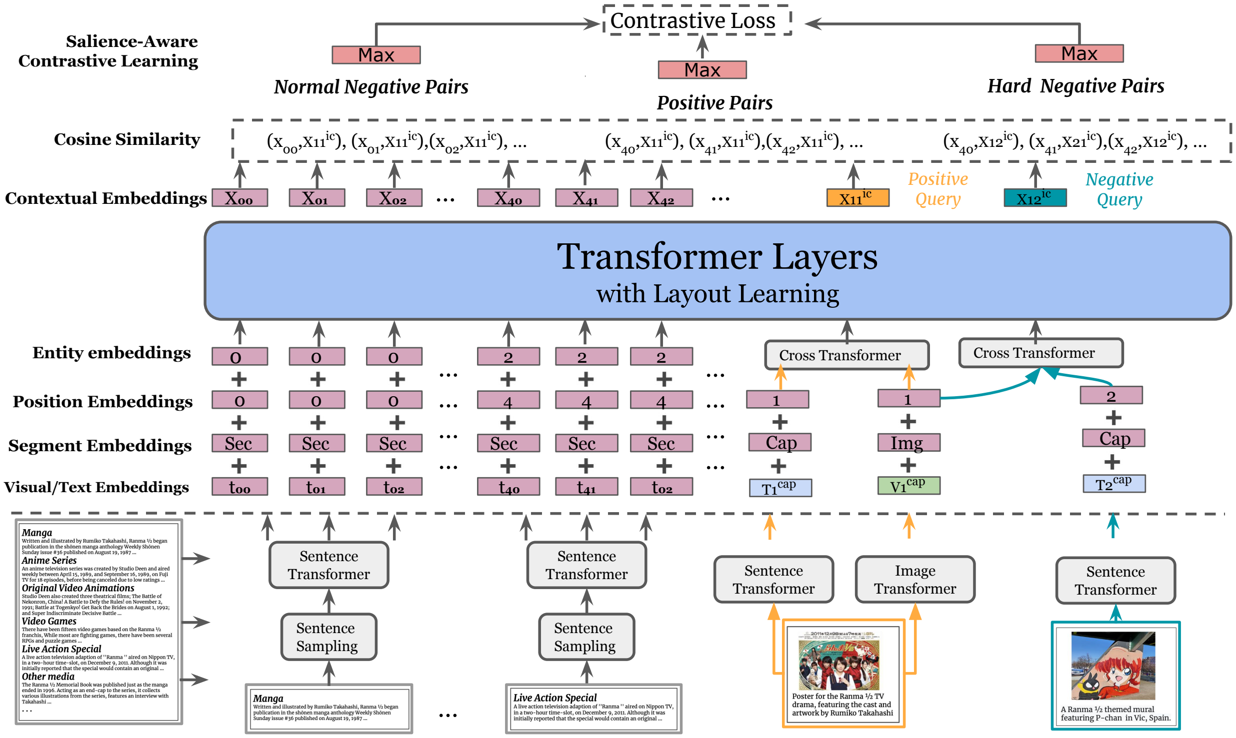

Figure 2: Overview of *DocumentCLIP* in the training phrase. In this example, t_{00}, t_{01}, t_{02} are the Top 3 sentences selected by the *Salient Sentence Extraction* from the image/caption pair in yellow. x_{12}^{ic} is the negative query from the image/caption pair in green to generated hard negative pairs. x_{11}^{ic} is the positive query to construct positive pairs and normal negative pairs.

2.2. Model Architecture

We build a multi-modal Transformer architecture as the backbone of *DocumentCLIP*, which takes text, images, and layout information as input to establish deep cross-modal interactions. We describe the strategies to encode sections of the document in 2.2.1. In 2.2.2, we present how *DocumentCLIP* leans layout information and feature extraction. Finally, we discuss the objective functions and model inference in 2.2.3. Detailed descriptions of the model are illustrated in Figure 2.

2.2.1 Section Encoding Strategy

Although BERT-like models become the state-of-the-art techniques on several challenging NLP tasks, they usually leverage only text information without image features. CLIP [28] models achieves great success in image-text retrieval task. However, it's only able to handle short text (less than 77), which is much shorter than the average length (195.5) of the sections within Wikipedia documents. Therefore, we experiment different sentence sampling strategies and propose the novel *Salient Sentences Extraction* method to learn the section representation.

First Sentence. It is evident that the first sentence plays an important role and always presents the central meaning of the sections. Therefore, we can embed the first sentence to represent the section.

Weighted Average. Apart from the first sentence, the associated image can also provides evidences for other sentences in a section. Thus, we train a one-layer transformer to aggregate all the important sentence embeddings in a section.

All Text Concat. Compared with the *Weighted Average* method, we regard the text in a section as one sentence and encode them together by a text transformer.

Salient Sentences. Motivated by the success of transferring image-text pre-training knowledge, we first adopt CLIP [28] to compute a scaled dot product similarity score Sim_{si} between each sentence in a section and images. Then we also do the operation Sim_{sc} for sentences and captions. After that, we utilize the average value of the Sim_{si} and Sim_{sc} to represent the correlation between the sentence and image/caption pairs. Finally, we pick the top K salient sentences in each section with the highest scores as the candidates. In Figure 2, K is 3 for simplicity and candidates are generated from image/caption in yellow.

2.2.2 Document Layout Learning

In this section, we extend to jointly model text, image and layout information in the *DocumentCLIP* framework.

Text Embedding. Following the common practice, we use lower-cased byte pair encoding (BPE) [32] to tokenize the sentences in documents. Then, we add `[CLS]` at the beginning of the sequence and `[SEP]` at the end. Extra `[PAD]` tokens are appended to the end so that the final sequence's length is exactly the maximum sequence length `L`. We directly adopt CLIP for initialization to extend its ability. The final token embedding is the sum of token embeddings and positional embeddings. Formally, the sentence embedding is computed by the text transformer [29] and output of `[CLS]` token which $T_j^{sec}$is utilized as sentence representations in documents. $T_j^{sec} \in [t_{j0}, ..., t_{jK}]$, where $j \in$ (0 $\leq$ j < N) means the section index in a document. $[t_{j0}, ..., t_{jK}]$ are the top K candidates selected by salient sentence extraction in the `j-th` section. Similarly, the caption representation is T_i^{cap}, where $i \in$ (0 $\leq$ i < M) and `M` means the caption number in a document.

Visual Embedding. In order to obtain image embedding, vision transformer is adopted to extract N non-overlapping image patches and perform linear projection to map every patch into 1D token. With injection of positional embedding and extra [CLS] token, the sequence of tokens are fed into image transformer layers, to model the correlation of each patch, where each layer ls comprises of Multi-Head Self-Attention, layer normalization, and Multi-layer Perception. Then, output `[CLS]` token embedding is used to represent the image feature V_k^{img}, where $k \in$ (0 $\leq$ k < M) and `M` means the image number in a document.

Layout Embedding. The layout embedding layer is to understand the global context of the document. Thus, The final text embedding of captions l_i^{cap} is the sum of three embeddings. The sentence embedding T_i^{cap} represents the caption itself. `PosE(i)` is the positional embedding, where i represents the caption index in the document. The segment embedding SegE(T_i^{cap})=[C] is used to distinguish different segments. For instance in Figure 2, the position index of the image is `1` since it's the second image in the document. Similarly, the final image embedding l_k^{img} is aggregated by V_k^{img}, `PosE(k)` and SegE(V_k^{img})=[V].

$$l_i^{cap} = T_i^{cap} + \mathsf{PosE}(i) + \mathsf{SegE}(T_i^{cap}) \quad (1)$$
$$l_k^{img} = V_k^{img} + \mathsf{PosE}(k) + \mathsf{SegE}(V_k^{img}) \quad (2)$$
$$l_j^{sec} = T_j^{sec} + \mathsf{PosE}(j) + \mathsf{SegE}(T_j^{sec}) + \mathsf{EntE}(s_j) \quad (3)$$

Inspired by the fact that the complementary images/caption pairs can be the literal visualization of entities mentioned in document sentences, we calculate the common entity numbers between sections with captios and sort sections in the descending order. More formally, we introduce the entity embedding EntE(s_j) for l_j^{sec}, where $s_j \in$ (0 $\leq$ i < N). If a section share most entities with a images/caption pair, s_j=0, while s_j=N-1 if it shares the least entities. In Figure 2, given the positive query (image/caption in yellow), s_0=0 and s_4=2.

Image and Caption Fusion. The intuitive way to select the related section is to compute the similarity score between section text with the image and caption separately. Then train a learnable coefficient to combine them. However, it ignores the potential correlation between the image and caption effectively. Therefore, we apply the Cross-modal Transformer [12] to fuse the l_i^{cap} and l_k^{img} before the layout transformer. We also add the `[cls]` token in the first place of the input sequence. The outputs from Cross-modal Transformer is a sequence of contextualized embeddings. We use the output from the `[cls]` token as the unified representation of a image and caption pair.

$$l_{ki}^{ic} = \textit{Cross-Transformer}(l_k^{img}, l_i^{cap}) \quad (4)$$

Layout Transformer. We concatenates section sentence embeddings $\{l_0^{sec},\ l_1^{sec},\ \ldots,\ l_{N-1}^{sec}\}$ and the image/caption pair $\{l_{00}^{ic},\ l_{11}^{ic},\ \ldots,\ l_{(M-1)(M-1)}^{ic}\}$ to a unified sequence L. In order to make the features more compact, we employ an FC layer with a ReLU activation before the layout Transformer [19]. Formally, we obtain the contextual multimodal embeddings X after transformer layers with layout learning. More formally, $X = \{x_0^{sec},\ \ldots\ x_{N-1}^{sec},\ x_{00}^{ic},\ \ldots,\ x_{(M-1)(M-1)}^{ic}\}$. $x_j^{sec} \in [x_{j0}, ..., x_{jK}]$ are the contextual embeddings of the Top K sentences each section, where $j \in$ (0 $\leq$ j < N) means the section index.

$$X = \textit{Layout-Transformer}(ReLU(FC(L))) \quad (5)$$

2.2.3 Salience-Aware Contrastive Learning.

In this section, we introduce the definition of positive/negative pairs for similarity learning and our salient-aware contrastive loss..

Positive Pairs. We define that a section x_j^{sec} and image/caption pair $x_{k,i}^{ic}$ is positive if the section contains certain content that is relevant to both the image and caption in $x_{k,i}^{ic}$. For example in Figure 2, $\beta^p = \{(x_{40}, x_{11}^{ic}), (x_{41}, x_{11}^{ic}), (x_{42}, x_{11}^{ic}), \ldots\}$ are the positive pairs, where x_{11}^{ic} is the positive query. $[x_{40}, x_{41}, ..., x_{4K}]$ are the Top K sentences embeddings in the forth section.

Negative Pairs. Our negative pairs set β^n contains two parts: normal negative pairs and hard negative pairs. In normal negative pairs, the groundtruth section is replaced by other irrelevant sections in the same document. In Figure 2, $\{(x_{00},\ x_{11}^{ic}), (x_{01},\ x_{11}^{ic}), (x_{02}, x_{11}^{ic}), (x_{10}, x_{11}^{ic}), (x_{12}, x_{11}^{ic}), (x_{12}, x_{11}^{ic}) \ldots\}$ are the positive pairs, where x_{11}^{ic} is the positive query. $[x_{40}, x_{41}, ..., x_{4K}]$ are normal negative pairs. As for the hard negative pairs, either the caption items or the image items might be changed but not both of them. For example in Figure 2, x_{12}^{ic} is the hard negative query and $\{(x_{40},\ x_{12}^{ic}), (x_{40},\ x_{21}^{ic}), (x_{41}, x_{12}^{ic}), (x_{41}, x_{21}^{ic}), \ldots\}$ are hard negative pairs. This is inspired by the fact that both the caption and image play important role to determine the relevant text from the document.

	KVQA	TQA	Visual Genome	GoodNews	Visual News	Wikipedia (Ours)
Avg Unit Length	84.8	920.8	263.9	451.0	773.0	195.5
Avg Doc Length	11.4	106.0	5.3	18.0	18.8	3346.6
Avg Sent Length	11.4	12.2	5.3	18.0	18.8	22.3
Avg Sent Num per Unit	1.0	7.7	1.0	1.0	1.0	8.2
Avg Img Num per Doc	–	3.0	–	1.8	1.7	4.8

Table 1: Statistics in terms of the document(*Doc*), unit and images(*Img*). *Doc* donates document and *Sent* donates sentence. *Unit* means the section in our dataset.

Salience-Aware Contrastive Loss. Motivated by the fact that the section label is used as the weak supervision, we propose to equip the model with the ability to extract most salient sentence from each section given the query x_{ki}^{ic} and document context. Additionally, unrelated sentences in groundtruth section can introduce additional noise in the contrastive learning process. Formally, we measure the cosine similarity between query x_{ki}^{ic} and $[x_{j0}, ..., x_{jK}]$. Then we select the maximum score to represent the correlation between the section and image/caption pair:

$$q = x_{k,i}^{ic}, s = x_j^{sec} \in [x_{j0}, ..., x_{jK}] \tag{6}$$

$$S(q, s) = max\{cos(q, s)\}, (q, s) \in \beta^p \tag{7}$$

$$S^*(q, s) = max\{cos(q, s)\}, (q, s) \in \beta^n \tag{8}$$

$$L = -\sum_{(q,s)\in D} [log(\frac{S(q, s)}{S(q, s) + \sum_{S^* \in \beta^n} S^*(q, s)})] \tag{9}$$

where $S(q, s)$ collects the maximum similarity scores of positive pairs and $S^*(q, s)$ includes the negative pairs. D means the document. Finally, our model is trained by minimizing the infoNCE [28] loss.

Model Inference. Given a image/caption pair as the query $x_{k,i}^{ic}$, the cosine similarity between the query and all sentences in the document is computed. Then we sort the sentences in descending order according to the similarity and the first candidate will be picked as the most relevant sentence prediction. We also regard the section including this sentence as the most relevant section prediction.

3. Pretraining Dataset

In this section, we describe the data creation procedure of our pretraining dataset: Wikipedia dataset and present dataset statistic. First, we used the latest English Wikipedia dump[2] to extract Wikipedia articles [26]. In Wikipedia articles, images are nested in two HTML tags, which delineate the sections. We regard all the sentences within the same section as the ground truth since obtaining sentence-level associations would require additional manual labeling. [34]'s dataset has the groundtruth linking between image/caption with sentences in the news articles. However, we don't have the right to use them according to the official website[3].

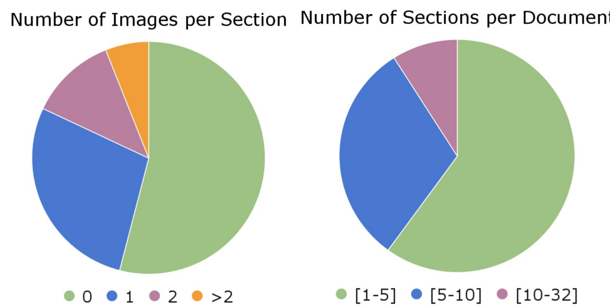

Figure 3: Distributions of the image number per section and section number per document in our Wikipedia dataset.

Consequently, we obtained 150,656 documents, not including ones failed to be parsed. Then we only kept documents with 2 to 30 images and no more than 32 sections. We also converted all the extracted images to an RGB format in order to guarantee quality. Eventually, this resulted in 66k articles and 320k images. We summarize the statistics of our dataset in Table 1 with previous datasets [31, 7, 19, 3]. Each section in our dataset has 8.14 sentences on average, the average sentence length is 22.3 words and the average section length is 195.5 words in our dataset, which is the largest among the compared datasets. In addition, over 50% documents in our dataset have more than 10 sections and many sections don't have corresponding images (Figure 3). Compared to other datasets used in Image-Text alignment methods (e.g. COCO captioning), the amount of text associated with an image is much larger. This confirms that understanding of longer texts is the required ability.

.

4. Experiments

In this section, we first introduce details of our implementation and present comprehensive experiment results.

4.1. Experimental Details

Evaluation Setting. Supervised settings: We first evaluate the most relevant section prediction on the Wikipedia

[2] https://github.com/muraoka7/tool4ipp

[3] https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-Service#2

Model	R@1	R@3	A-R@1
CLIP(*ViT32*) [28]	0.40	0.69	0.39
CLIP(*RN50*) [28]	0.39	0.67	0.38
SBert [30]	0.38	0.67	0.38
Pythia [26]	0.33	0.55	0.34
Newsroom [17]	0.31	0.51	0.31
SANDI [27]	0.34	0.57	0.33
LAIS [46]	0.45	0.74	0.45
DocumentCLIP*-Early Fusion*	0.57	0.87	0.56
DocumentCLIP*-Late Fusion*	0.56	0.85	0.54
DocumentCLIP*-Early Fusion*	0.57	0.87	0.56
-w/o salience-aware loss	0.54	0.84	0.53
-w/o entity check	0.46	0.76	0.46
-w/o layout information	0.39	0.67	0.39

Table 2: Comparative Experiments with baselines and ablation study on Wikipedia datasets.

Training		Inference		R@1	R@3
Image	*Caption*	*Image*	*Caption*		
✓	*Original*	✓	*Original*	0.57	0.87
✓	*Generated*	✓	*Generated*	0.51	0.82
✗	*Original*	✗	*Original*	0.54	0.84
✗	*Generated*	✗	*Generated*	0.49	0.79
✓	✗	✓	✗	0.50	0.81
✓	*Original*	✓	✗	0.52	0.82
✓	*Original*	✗	*Original*	0.55	0.85

Table 3: Performance of *DocumentCLIP* with or without images and captions. *Generated* captions are from *BLIP* [10].

Section Encoding	Initialization	R@1	R@3
First Sentence	*CLIP(ViT32)*	0.54	0.84
Weighted Average	*CLIP(ViT32)*	0.55	0.85
All Sentences Concat	*CLIP(ViT32)*	0.52	0.82
Salient Sentence	*CLIP(ViT32)*	0.57	0.87
Salient Sentence	*CLIP(RN50)*	0.55	0.86
Salient Sentence	*ROBERTA**	0.56	0.86
Salient Sentence	*SCRATCH*	0.49	0.79

Table 4: Performance of *DocumentCLIP* to investigate different section encoding strategies and initialization methods. *ROBERTA** means using the vision transformer *CLIP(ViT32)* to encode images and *Roberta* [24] to encode text.

Section Encoding	Top$K$	R@1	R@3
Salient Sentence	$K = 1$	0.54	0.84
Salient Sentence	$K = 3$	0.56	0.86
Salient Sentence	$K = 5$	0.57	0.87
Salient Sentence	$K = 7$	0.57	0.86
Salient Sentence	$K = 9$	0.55	0.84

Table 5: Performance of different number of candidates (Top K) in the salient sentences extraction.

dataset in both the image level and document level. At the image level, we use *R@N(N=1,3)* to calculate the percentage of images where the groundtruth section is among the *TopN* predicted sections. At the document level, we measure the percentage of documents where all images in the document have a correct *Top1* predicted section: *A-R@1*. Zero-shot setting: We ask human annotators to evaluate the most relevant sentence prediction on the *Open Textbook* dataset. *DocumentCLIP* is trained on the Wikipedia dataset and directly tested in the zero-shot setting. More detailed more described in 4.3.

Implementation Details. As for the experiment setting, we randomly split our dataset into 80% for training, 10% for validation and 10% for testing. We trained our model with mini-batches of size 32 by using stochastic gradient descent. We used AdamW [39] as an optimization algorithm and a warm-up strategy with the initial learning rate of 0.001.

Baseline Models. We run the public release of *CLIP* [28] and *SBert* [30] and finetune them on the Wikipedia dataset in the supervised task. As for the zero-shot setting, they aren't finedtuned. *CLIP* takes both images and captions as the input while *SBert* only takes the captions. Additionally, only the first sentence in each section is fed into *CLIP* since its maximum length of text encoder is 77, which is much below the average section length. *Pythia* [26] encodes different modalities separately and computes an attention layer to attend specific parts in the image. *Newsroom* [26] adopts a hierarchical self-attention mechanism to attend to both key words within a piece of captions and informative components of a document. *LAIS* [46] introduce a two-stage architecture to select the insertion position for the images.

4.2. Supervised Experiment Results

4.2.1 Comparison with Baselines

We report the overall results of most relevant section prediction in the supervised setting on the Wikipedia dataset (Table 2). We make the following observations: our model *DocumentCLIP* achieves more competitive performance in all scores by a large margin up to 10%. This remarkable improvement indicates that *DocumentCLIP* successfully possesses the intra-document alignment between image and text. The major reason is that previous methods aren't able to distinguish between the sections in the same document since these sections have similar topics. Instead, we use salience-aware contrastive learning with layout features to learn the document context and compare common entities with captions to differentiate between sections.

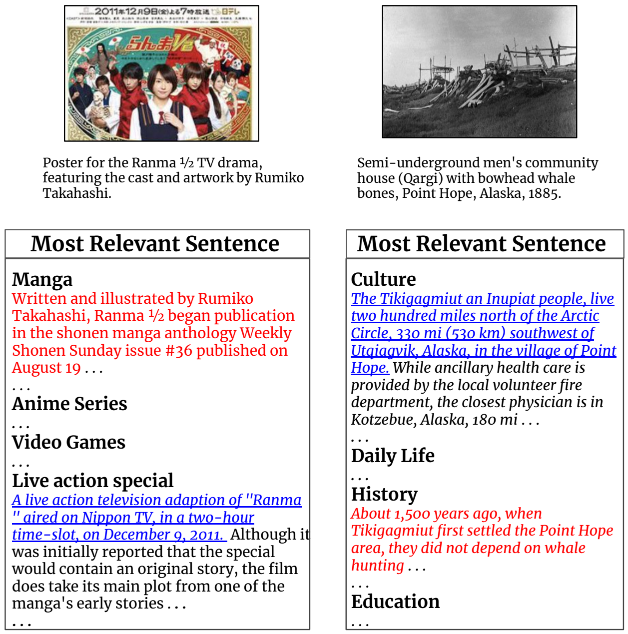

Figure 4: Two examples of Wikipedia articles. The red section is the prediction from baseline method while the blue section is the prediction by *DocumentCLIP*. The sentence in blue is the most relevant candidate.

4.2.2 Ablation Studies

Contributions of Different Components. The below sections of Table 2 summarizes the ablative performance of our model with different components. Without layout information, the performance shows 7% degradation in terms of *R@1* and 9% degradation in *R@3*, which demonstrates that the structural information like the position features and segment features are essential for our model to learn the correlation and difference between each sections and how to connect to the semantic of image/caption pairs. We notice that *DocumentCLIP* benefits from entity check and salient-aware loss since they can not only filter out noisy sentences in sections but also check keyword alignment between captions and sections. Additionally, the improvement from early fusion strategy indicates the its efficiency to aggregate multi-modal features between the image and caption.

Importance of Different Modalities. Besides, we investigate the contribution of each modality (image or caption) in Table 3. Given only one modality in training and inference, the performance of captions is superior to that of images (third and fifth row). This can be reasonable because captions sometimes share similar or same words with textual sections. To further analyze the effect of captions, we replace the original captions with the ones generated by *BLIP* [10], an state-of-art and pretrained vision-language generation model (second and forth row). The performance degradation

Model	Avg Rank	Rank #1	Rank #2	Rank #3
SBert	1.29	22%	31%	47%
CLIP	1.13	29%	36%	35%
DocumentCLIP	**0.75**	**49%**	33%	18%

Table 6: Human evaluation of different models on *Open Textbook* dataset in the zero-shot setting. Annotators are asked to put the sentences in order of relevance to the Image/Caption pair. "*1.29*" means the average rank of *SBert* is 1.29. "*22%*" means only 22% of the predictions from *SBert* are put in the 1st place.

is expected since generated captions always miss the key entities in the original ones, which play the evidence role.

Section Encoding Strategies. Table 4 studies the effect of different section encoding strategies. Compared to other strategies, the improvement is evident when using salient sentence extraction. The main reason for this might because other strategies fail to remove the influence of the noisy sentences, which will mislead the model to pick the wrong the section. To further investigate its effect, we gradually increase the Top$K$ value and implement the salient sentence extraction in Table 5. We notice that a larger K will yield better accuracy when K is smaller than 5. The phenomenon indicates that single sentence isn't informational enough to represent the section while too many ones will introduce additional noise.

Initialization. We also analyze different initialization methods in Table 4. As the *ROBERTA** (six row) method suggests, we use the vision transformer *CLIP(ViT32)* to encode images and *Roberta* [24] to encode text. We found it achieves best results if we use *CLIP* to encode both images and captions. This is because *CLIP* is trained on a large vision and language dataset, alleviating the semantic gap between images and text. Besides, *SCRATCH* performs worst due to missing the prior knowledge.

4.2.3 Qualitative Analysis.

In order to evaluate the quality of the complementary text retrieved by our model in the sentence level, examples are shown in Figure 4. This left document talks about a Japanese manga series called Ranma ½. The image/caption pair is in the 2nd place and talks about the TV drama version of Ranma ½. Previous methods regard the first section as the prediction since it share more common entities with the caption. However, they fail to encode image/pair position information and understand commonsense knowledge. For example, *"live action special"* from the fifth section actually refers to *"TV drama"* in the caption. As for the right document, current methods only consider the keywords within the captions without the image position information.

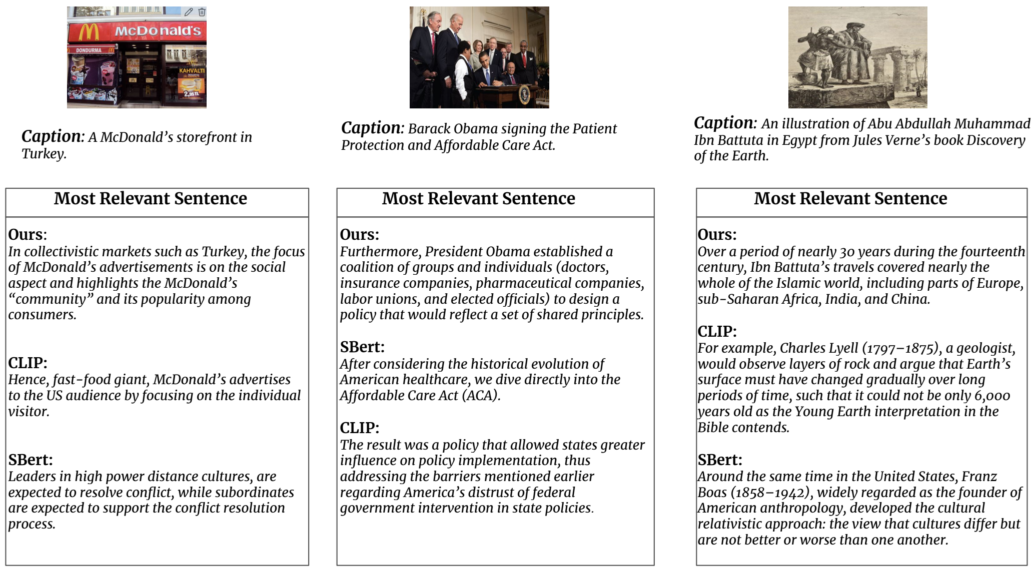

Figure 5: Selected image/caption samples and most relevant sentence predictions from Open Textbook dataset. We compare our predictions with the ones from *CLIP*, *SBert* and demonstrate the advance of *DocumentCLIP*.

4.3. Zero-Shot Experiment Results

In this section, we gain further insights into the most relevant sentence prediction on the *Open Textbook* dataset via human evaluation. With the human evaluation, we access whether our model could pick more relevant sentences than *CLIP* [28] and *SBert* [30] in the zero-shot setting.

Here, we randomly select a set of 40 image/captions pairs from 10 articles in *Open Textbook* dataset. We conduct our evaluations on 75 workers from Amazon Mechanical Turk[4]. The 40 questions in the questionnaire are shuffled for each worker and it takes 40 minutes to complete on average. We also add 3 attention questions to guarantee the quality of the response. As for each question, they are given a caption and image pair and asked to put the three sentences in order of relevance to the Image/Caption pair and the most relevant should be in the 1st place. Summary of the evaluation are in Tab 6. The average rank of the predictions from *DocumentCLIP* is 0.75, which is much lower than *SBert*: 1.29 and *CLIP*: 1.13. In addition, 49% of the samples have the *DocumentCLIP* in the 1st place and only 18% of the samples have the *DocumentCLIP* in the final place. Figure 5 demonstrate the advance of our model to understand the intra-document multimodal interaction.

5. Related Work

Cross-modal research [10, 1, 9, 23, 44, 38, 36, 23, 20, 19, 21, 22, 18, 15] has recently received increasing attention as a result of the advances in both computer vision and natural language processing. [6, 8, 45, 35] encode multimodal features into a common embedding space in which instance similarity is measured by conventional cosine or Euclidean distance. Recent state-of-the-art works [28, 13, 25, 33] employ pretraining methods, cross transformers and contrastive learning algorithms to learn the correlation between different modalities. Various pre-training objectives have also been proposed over the years, and have progressively converged to a few time-tested ones [38, 11, 14, 40, 42, 41, 43]. However, the limitation is that they are trained to align short literal text descriptions with images but fail to handle real-world media such as news articles [19], Wikipedia pages [26], magazines, product descriptions consisting of multiple sentences with multiple images and layout information. Document structure analysis models [5, 4, 37, 2] process pages one by one, but the figures may relate to text located on other pages.

Different from existing methods, *DocumentCLIP* is not only able to effectively process long text, but also understand text-image interaction at the sentence-level in the document.

6. Conclusions

In this paper, we are the first to explore the sentence-level multimodal intra-document links in vision-language pretraining. We also introduce a novel contrastive learning framework with layout information, multimodal interaction, novel section encoding strategy, salience-aware contrastive loss and hard negative samples. We also collect a large visual document dataset from Wikipedia with 66k articles, 320k images/caption pairs and various topics. Furthermore, our proposed model *DocumentCLIP* achieves state-of-the-art performance in both the supervised and zero-shot settings. We hope this work paves the way for future studies in document understanding and multi-modal alignment.

[4] https://www.mturk.com

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021.

[3] Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. *arXiv e-prints*, pages arXiv–1904, 2019.

[4] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50, 2021.

[5] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022.

[6] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.

[7] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.

[8] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[11] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[12] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[14] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[15] Zongxi Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *ArXiv*, abs/2307.05052, 2023.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[17] Fangyu Liu, Rémi Lebret, Didier Orel, Philippe Sordet, and Karl Aberer. Upgrading the newsroom: an automated image selection system for news articles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(3):1–28, 2020.

[18] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *ArXiv*, abs/2306.14565, 2023.

[19] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*, 2020.

[20] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2020.

[21] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visualnews : A large multi-source news image dataset. *arXiv: Computer Vision and Pattern Recognition*, 2020.

[22] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visualnews : Benchmark and challenges in entity-aware image captioning. *ArXiv*, abs/2010.03743, 2020.

[23] Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. Covid-vts: Fact extraction and verification on short video platforms. *arXiv preprint arXiv:2302.07919*, 2023.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[26] Masayasu Muraoka, Ryosuke Kohita, and Etsuko Ishii. Image position prediction in multimodal documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4265–4274, 2020.

[27] Sreyasi Nag Chowdhury, Simon Razniewski, and Gerhard Weikum. Sandi: Story-and-images alignment. In *16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 989–999. ACL, 2021.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[31] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019.

[32] Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. 1999.

[33] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[34] Alasdair Tran, Alexander Mathews, and Lexing Xie. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13035–13045, 2020.

[35] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*, pages 12–20, 2019.

[36] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[37] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020.

[38] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[39] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[40] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[41] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[42] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.

[43] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[44] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 1(6):8, 2021.

[45] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701, 2018.

[46] Zhengkun Zhang, Jun Wang, Adam Jatowt, Zhe Sun, Shao-Ping Lu, and Zhenglu Yang. News content completion with location-aware image selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14498–14505, 2021.