Vision-Language Pre-Training with Triple Contrastive Learning

Chevula Sujeeth Reddy
Department of Computer Science
Texas Tech University
Lubbock, United States
schevula@ttu.edu

Gorrepati Sushmitha
Department of Computer Science
Texas Tech University
Lubbock, United States
sugorrep@ttu.edu

Jayasri Jetti
Department of Computer Science
Texas Tech University
Lubbock, United States
jjetti@ttu.edu

Mucheli Sulakshana
Department of Computer Science
Texas Tech University
Lubbock, United States
smucheli@ttu.edu

Thokala Maheswari
Department of Computer Science
Texas Tech University
Lubbock, United States
mthokala@ttu.edu

***Abstract*—Vision-language representation learning has significantly benefited from image-text alignment techniques that employ contrastive losses, such as InfoNCE loss. This alignment strategy's success can be attributed to its ability to maximize mutual information (MI) between an image and its corresponding text.However, relying solely on cross-modal alignment (CMA) overlooks the inherent data potential within each modality, potentially leading to degraded representations. While CMA-based models can map image-text pairs closely together in the embedding space, they fail to ensure that similar inputs from the same modality remain in close proximity. This issue can be exacerbated by noisy pre-training data.To address these limitations, we propose triple contrastive learning (TCL), a vision-language pre-training method that leverages both cross-modal and intra-modal self-supervision. In addition to CMA, TCL introduces an intra-modal contrastive objective to enhance representation learning.To effectively utilize localized and structural information from image and text inputs, TCL maximizes the average MI between local regions of image/text and their global summary. TCL is the first method to incorporate local structure information into multi-modality representation learning.Experimental evaluations demonstrate that our approach achieves competitive results and establishes new state-of-the-art performance on various downstream vision-language tasks, including image-text retrieval and visual question answering.**

***Index Terms*—Triple Contrastive Learning, Cross Modal Alignment, Intra Modal Contrastive, Local MI maximization, Masked Language Modeling, Image Text Matching, Vision-Language Pre-training.**

I. Introduction

In both vision and language representation learning, self-supervision is an active research topic. In comparison to image's content with information that has been retrieved, the success can be determined. Several complicated tasks are there to achieve while extracting the relevant data from a picture such as wavy textual images, disoriented texts, noisy images and a variety of texts and fonts.These challenges can be minimized by various other techniques such as CMA (Cross-modal alignment), OCR (Optical Character Recognition) and so on. Even though there are number of ways, each of it has limitations which may result in degraded representations.The TCL incorporates intra-modal contrastive objective to significantly provide representation learning. We can increase the performance by pre-training TCL with huge data set and perhaps it will make much more progress.

II. Related Work

A. Vision-Language Pre-training (VLP)

The effectiveness of self-supervised learning in intra-modal tasks (e.g., language and vision) has created an increasing interest in developing pre-training goals for tasks involving multiple modalities. The approaches deliver impressive results, but they do not perform image-text alignment previous to fusion, which makes it difficult for the understand of the interplay between different modalities. Although the objectives of our method and those of ALBEF are similar, there are a few main differences. To ensure that the learned representations are semantically meaningful, we suggest using cross-modal and intra-modal self-supervision in addition to cross-modal alignment (CMA). We augment the cross-modal scenario with local alignment by raising the mutual information (MI) between local areas and global representations.

B. Mutual Information (MI) Maximization

Mutual Information(MI) refers to either measuring the link between random variables or determining the amount of shared knowledge. The MI is frequently employed in unsupervised feature learning, where the main aspect is to maximize MI between the input and output. However, for deep neural networks, the mutual information of huge random variables is challenging and unsolvable. InfoNCE, which is known as a categorical value, is used to find the positive sample among the group of negative samples.It also demonstrated that InfoNCE is a lower bound on MI, therefore decreasing InfoNCE loss can indirectly increase MI. For an accurate representation, maximizing the global mutual information is insufficient. With the local MI, most of the Augmented Multiscale Detailed InfoMax maximizes Mutual Information between the features that were recovered from different image augmentations, whereas Long InfoMax maximizes the difference in MI between neighboring areas. The limitations about intra-modal activities our approach presents the concept of local MI maximization for multi-modal scenarios.

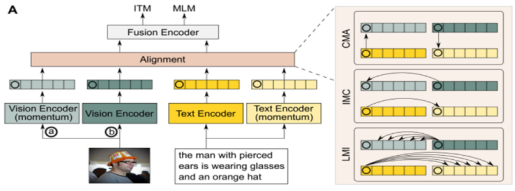
III. IMPLEMENTATION METHODS

We first describe the model architecture of our approach and then discuss uni-modal representation learning.The proposed triple contrastive learning modules are thoroughly detailed.

1. CMA (Cross-Modal Alignment)
2. IMC (Intra Modal Contrastive)
3. LMI (Local MI Maximization)

Finally, we summarize two pre-training objectives: image-text matching and masked language modeling.

IV. MODEL ARCHITECTURE



This architecture includes an vision encoder,text encoder and alignment module, fusion encoder.In this scenario, the input for the vision encoder is an image, whereas the input for the text encoder is text. The text and image content encoded in the text encoder and processed by the alignment module contains three learnings: LMI, CMA, and IMC. In this step, the similarities between the image and the text are matched and arranged in chronological order in the fusion encoder. This also predicts the truth labels of masked text tokens in masked language modeling (MLM) and determines whether or not an image text pair matches in image text matching (ITM).

A. Uni-modal Representation Learning

In order to get two corresponding perspectives of I1 and I2, we consider the image-text pair (I, T) and apply two independent augmentations to the picture. These two points of view are strengthened as a positive pair. Each view is divided into equal-sized patches that are constant in size and are represented with positional information. A specific categorization token [CLS] is used to describe the entire image. To learn visual representations vcls, $\{v_{\text{cls}}, v_1, \ldots, v_M\}$, where M is the total number of image patches, we save the data of I1 in g(.). Similarly, for image I2, we keep data in $\hat{g}(\cdot)$ to monitor representations $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_M\}$.

By using h(T) and h(T+), where N is the number of text tokens, we may obtain $\{t_{\text{cls}}, t_1, \ldots, t_N\}$ for the text input T as well as $\{\hat{t}_{\text{cls}}, \hat{t}_1, \ldots, \hat{t}_N\}$for the text input T+. By using

VLP, there is a disadvantage to learn interactions between image and text for the fusion encoder, so to overcome this disadvantage, alignment module is introduced. Alignment module consists of three phases which are CMA, IMC and LMI. We go over each phase in detail below.

B. Cross-Modal Alignment (CMA)

The main objective of CMA is to distinguish similar image-text pair and unsimilar image-text pair together. Basically, both are separately paired together. The other meaning of CMA tells that it maximizes Mutual Information (MI) between similar image-text pair.Because it is impractical to directly maximize MI for continuous and high-dimensional variables, we instead aim to minimize InfoNCE loss, which is the bottom bound of MI. The InfoNCE loss for image-to-text is defined as:

$$\mathcal{L}_{nce}\left(I_1, T_+, \tilde{T}\right) = -\mathbb{E}_{p(I,T)}\left[\log \frac{e^{(\text{sim}(I_1, T_+)/\tau)}}{\sum_{k=1}^{K} e^{(\text{sim}(I_1, \tilde{T}_k)/\tau)}}\right]$$

where τ is a temperature hyper-parameter,

$\tilde{T} = \{\tilde{T}_1, \ldots, \tilde{T}_K\}$ is a set of negative text examples, and we know that. $\text{sim}\left(T, I_2\right) = f_t\left(t_{\text{cls}}\right)^T f_v\left(\hat{v}_{\text{cls}}\right)$, $f_t(\cdot)$ and $f_v(\cdot)$ are two projection heads.

Similarly, the loss of text-to-image is formulated by:

$$\mathcal{L}_{nce}\left(T, I_2, \tilde{I}\right) = -\mathbb{E}_{p(I,T)}\left[\log \frac{e^{(\text{sim}(T, I_2)/\tau)}}{\sum_{k=1}^{K} e^{(\text{sim}(T, \tilde{I}_k)/\tau)}}\right]$$

$\text{sim}\left(T, I_2\right) = f_t\left(t_{cls}\right)^T f_v\left(\hat{v}_{cls}\right), f_t(\cdot)$ and $f_v(\cdot)$are two projection heads. Combining, we have loss of CMA as:

$$\mathcal{L}_{cma} = \frac{1}{2}\left[\mathcal{L}_{nce}\left(I_1, T_+, \tilde{T}\right) + \mathcal{L}_{\text{nce}}\left(T, I_2, \tilde{I}\right)\right]$$

C. Intra-Modal Contrastive (IMC)

The semantic distinction between positive and negative samples within a modality is something that IMC

tries to understand. To ensure adequate learning of intra-modal representations, we minimize the following objective.

$\mathcal{L}_{imc} = \frac{1}{2}\left[\mathcal{L}_{nce}\left(T, T_+, \tilde{T}\right) + \mathcal{L}_{nce}\left(I_1, I_2, \tilde{I}\right)\right]$

The disadvantage of IMC is that the contrastive objective is only performed on [CLS] tokens from text and vision encoders, where [CLS] tokens are taken to represent the input's overall information. As a result, IMC maximizes overall MI between different types of augmented views.

D. Local MI Maximization (LMI)

The goal of local MI maximization is to create high MI between each local input region and the global representation. We increase our chances of success by reducing the following loss. MI between global and local regions:$\mathcal{L}_{lmi} = \frac{1}{2}\left[\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{nce}\left(I_1, I_2^i, \tilde{I}_l\right) + \frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_{nce}\left(T, T_+^j, \tilde{T}_l\right)\right]$

According to another viewpoint, the model is compelled to include fine-grained data, which is advantageous for joint representation learning. The model is also encouraged to predict local from the global representation by maximizing local MI.

E. Image-Text Matching (ITM)

To combine visual and linguistic representations, we use ITM, which has been extensively used in previous VLP studies. The ITM determines whether or not an image and text pair are matched, which is known as a binary classification problem. We generate negative data instances by batch sampling, assuming that each imagetext combination (I, T) sampled from the pre-training datasets is a positive example. ITM loss is defined as follows:

$\mathcal{L}_{itm}$ = $E_{p(1,T)}$ H(ϕ (I,T), $y^{(I,T)}$

where H (;) is the cross-entropy, $y^{(\mathrm{I,\ T})}$ denotes the label

F. Masked Language Modeling (MLM)

In order to forecast the ground truth labels of masked text tokens, we utilize MLM from BERT. We 15 percent randomly mask out text tokens and 80 percent of the time replace them with a specific [MASK] token, 10 percent with random words, and the other 10 percent keep it unaltered. The MLM loss is defined as:

$\mathcal{L}_{\mathrm{mlm}} = E_{p(1,T^{\mathrm{msk}})}H(\phi(I, T^{\mathrm{msk}}), y^{(I,T^{\mathrm{msk}})})$

The predicted probability of T_{msk} is denoted as $\Phi(I, T_{\mathrm{msk}})$, and $y^{(T_{\mathrm{msk}})}$ represents the ground truth. msk ground truth. The overall training objective of our model is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cma}} + \mathcal{L}_{\mathrm{imc}} + \mathcal{L}_{\mathrm{lmi}} + \mathcal{L}_{\mathrm{itm}} + \mathcal{L}_{\mathrm{mlm}}$$

V. EXPERIMENTS AND RESULTS

A. Pre-training Datasets

The pre-training datasets used in this experiment are SBU Captions, COCO, Visual Genome (VG), and Conceptual Captions (CC). To demonstrate that we have big-scale datasets that use CC12M, we can also apply our method to enormous data. With 14.97M distinct images and 16M image-text pairs, we can thus obtain large-scale pre-training data.

	COCO	VG	SBU	CC	CC12M
images	118K	110k	879K	2.95M	10.97M
text	577K	769K	859K	2.92M	10.95M

TABLE I: Statistics of pre-training datasets.

B. Downstream tasks

There are two tasks involved in image-text retrieval. First, text is the target and image is the query (TR), and second, text is the query and image is the target (IR). The pre-trained model is adequately modified on the training records and evaluated on the validation/test cases for the fine-tuning setting. On Flickr30K and COCO, the pre-trained model is evaluated using both zero-shot and fine-tuning settings.

Visual Question Answering (VQA): Guess the response to a question posed alongside a picture (in text format). To respond, one must possess linguistic, imaginative, and clairvoyant knowledge.

Visual Entailment (SNLI-VE): This three-class classification issue predicts whether a given image semantically entails a particular text.

Visual Reasoning (NLVR2): Ascertains if a caption in natural language is accurate for a pair of pictures.

C. Implementation Details

We conduct all of our experiments using the PyTorch framework on eight NVIDIA A100 GPUs. A 6-layer transformer implements our text encoder and the fusion encoder, while our picture encoder is implemented using ViT-B/16 with 12 layers and 85.8M parameters. The model undergoes 30 epochs of training during the pre-training phase, with a batch size of 512 and a weight decay of 0.02 for every mini-batch. The starting learning rate is set to -5. After 2000 training iterations, the learning rate is adjusted to -4. Using the cosine decay approach, we subsequently reduce the learning rate to -5.

D. Evaluation on Image-Text Retrieval

Standard procedure involves applying the trained model to downstream tasks in a zero-shot manner to assess how well the learned representations generalize. We evaluate our model through the standard evaluation

process on the COCO and Flickr30K datasets utilizing the zero-shot image-text retrieval benchmarking tasks. The table below illustrates how our method outperforms the state-of-the-art while still achieving the greatest results. We improve COCO by +9.5% on average and +12.2% on Flickr30K (on average), in contrast to ViLT, which directly represents the relationship between word and image using a transformer encoder, patch embeddings. This highlights the need for data fusion following crossmodal alignment. Strong ties exist between our approach and ALBEF, which aligns text and image embeddings first and then uses fusion. The ALBEF, which aligns text and image embeddings first before utilizing a fusion, and our work are closely related. yields +3.4% IR/R@1 and +2.7% TR/R@1 boosts on the MSCOCO (5K) dataset in comparison to ALBEF. For additional details on the intra-modal representation analysis, see the supplemental. Compared to ALIGN, our method yields a mean of 79.5 percent, which is superior to 70.9 percent on COCO and 94.0 percent compared to 92.2 percent on Flickr30K. This is particularly noteworthy. Almost 360 times as many image-text pairs as our model have been pre-trained by ALIGN, it should be noted.

All in all, our method's representations are more universal and transferable than those of the baselines that are currently in use.

Method	# Images	MSCOCO(5k)						Flickr30K(1K)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT	6M	45.0	72.3	81.4	32.4	60.1	72.2	71.2	91.2	95.4	53.5	78.5	89.4
UNITER	4.2M	63.0	85.6	93.4	48.7	75.8	86.2	81.2	95.6	98.1	65.6	87.8	93.8
ViLT	4.1M	57.6	83.5	88.5	41.2	71.2	82.2	87.0	98.8	95.07	56.0	83.5	90.1
CLIP	401M	59.7	88.7	87.6	36.8	63.4	73.4	88.0	98.1	99.4	68.5	90.8	96.3
ALBEF	4M	68.6	87.8	95.6	50.5	75.7	85.5	90.6	99.0	97.7	76.6	94.4	96.3
ALIGN	1.23B	57.4	89.5	90.4	45.5	68.6	78.6	88.7	98.4	98.7	79.2	93.7	95.6
OURS	4.2M	73.6	90.8	96.7	55.2	78.6	87.4	93.2	99.2	99.8	76.2	94.5	97.8

TABLE 2: The performance of zero-shot image-text retrieval on the COCO and Flickr30K datasets is compared. To be thorough, we also provide the results of ALIGN, which uses 1.8 billion image-text pairs (1.2 billion unique images) for pre-training. For image retrieval (IR) and text retrieval (TR), we provide the average of R@1, R@5, and R@10.

We created new benchmark findings for the investigations that were refined, which are displayed in the table below. On the medium-sized COCO dataset, we outperform ALBEF by 2.5 percent absolute TR/R@1 and 2.2 percent absolute IR/R@1, indicating that our model can still benefit from fully supervised training. We also exceed previous baselines on the tiny Flickr30K dataset.

Method	# Images	MSCOCO(5k)						Flickr30K(1K)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT	6M	67	89.7	93.8	51.2	79.2	86.8	87.2	98.2	98.7	72.6	93.2	96.4
UNITER	4M	65.8	89.2	94.2	53.3	80.1	88.1	86.6	98.0	98.7	76.4	93.8	97.3
VILLA	4.3M	X	X	X	X	X	X	88.4	97.6	99.2	77.7	93.8	97.2
OSCAR	4.4M	69	70.3	90.6	95.6	54.3	88.7	X	X	X	X	X	X
ViLT	4.5M	62	85.7	93.4	43.4	73.3	82.9	84.2	97.5	99.7	63.8	89.4	94.5
UNIMO	4.1M	57.8	81.5	87.6	38.2	61.6	71.8	88.7	99.3	98.2	69.5	91.86	94.7
SOHO	220K	65.8	87.7	94.3	51.4	78.2	88.3	87.5	97.6	98.8	73.4	91.8	98.0
ALBEF	4M	72.8	90.7	96.2	56.2	87.2	81.7	93.7	98.5	99.6	83.7	97.6	97.3
Ours	5M	76.2	93.2	97.4	57.2	89.2	91.2	95.4	99.2	98.5	84.3	97.4	98.8
ALIGN	1.23B	77.0	93.6	97.2	57.3	82.7	90.2	95.8	100	99.4	85.8	98.5	99.2

TABLE 3: Fine-tuned image-text retrieval performance is compared using COCO and Flickr30K datasets. To be thorough, we additionally include the outcomes of ALIGN, which pre-trains using 1.8B image-text pairs (1.2B unique images).

E. VQA, VE, and NLVR^2

Table 4 displays the comparison of the picture text's performance in the necessary VQA, VE, and NLVR2. The model's capacity to learn joint multi-modal embeddings is a prerequisite for its performance on these tasks. For five of the six criteria, we present here state-of-the-art findings, showing that taking cross-modal alignment into explicit consideration.

Module	ZERO-SHOT				Fine-Tune			
	MSCOCO		Flickr30K		MSCOCO		Flickr30K	
	TR	IR	TR	IR	TR	IR	TR	IR
+IMC(w/o aug) (4M)	70.7	51.8	92.3	79.2	75.0	59.3	94.5	83.4
+IMC(W/o aug)(4M)	73.3	54.7	95.2	84.2	78.1	61.2	96.6	86.2

TABLE 4: Performance comaprision on vision language tasks.

F. Ablation Study

To find the efficacy modules (i.e., IMC and LMI) in improving multi-modal representation learning, we perform ablation studies on image-text retrieval tasks. Ablation study of each component for image-text retrieval tasks. Reports state that R@1. We apply CMA+ITM+MLM using the results of ALBEF.

Pooling	Intermediate	ZERO-SHOT				Fine-Tune			
		MSCOCO		Flickr30K		MSCOCO		Flickr30K	
		TR	IR	TR	IR	TR	IR	TR	IR
	X	71.6	53.3	91.6	78.7	76.3	59.4	95.4	84.2
	X	72.2	53.1	91.7	78.2	75.6	58.7	93.7	83.2
X		71.5	53.6	93.6	78.6	76.4	59.3	95.2	83.2
X		71.5	54.2	93.5	80.4	76.4	59.2	95.7	85

[h]

TABLE 5: Each component's ablation study for image-text retrieval tasks. There is a reported R@1. To calculate CMA+ITM+MLM, we utilize the ALBEF results.

We investigate these two options on image-text retrieval tasks, noting the importance of picture patch

Module	Zero-Shot				Fine-Tune			
	MISCOCO		Flickr30k		MISCOCO		Flickr30K	
	TR	IR	TR	IR	TR	IR	TR	IR
CMA+ITM+MLM	69.1	51.7	91.8	77.2	74.6	57.7	95.8	83.5
+LMC(w/o aug)	73.5	53.4	93.4	79.3	76.0	59.5	95.5	83.7
+LMC	73.5	54.8	94.2	79.8	76.0	59.8	96.2	84.4
+IMC+LMI(Ours)	73.5	54.8	93.4	79.3	76.7	60.0	95.7	85.0

TABLE 6: The impact of intermediate local features and picture patch pooling on image-text retrieval is investigated through ablation. R@1 has been documented.

Method	#Images	VQA		NLVR^2		SNLI-VE	
		Test-dev	test-std	dev	test-P	val	test
OSCAR[28]	4M	74.14	74.33	79.02	79.23	X	X
UNITERM[8]	4M	73.70	73.80	78.19	78.98	79.45	79.31
ViLT[24]	4M	72.28	X	76.08	77.19	X	X
UNIMO[27]	4M	73.29	75.08	X	X	81.5	80.04
VILLA[16]	4M	74.65	74.32	79.56	79.80	79.47	79.09
ALBEF[26]	4M	74.63	75.84	80.42	80.71	80.20	80.11
Ours	4M	75.40	75.80	80.36	82.51	80.76	80.45
Vin VL	4M	76.59	77.19	83.29	84.20	X	X

TABLE 7: Ablation study of the size of pre-training datasets.R@1 is reported.

pooling. Furthermore, the effectiveness of applying last-layer patches is on par with, if not slightly superior to, that of applying patches from intermediate layers.

To study the effects of training on larger-scale datasets, we perform ablation research on 14M datasets using +IMC (w/o aug). We find that training somewhat improves performance. We believe that.

VI. IMPROVEMENTS

A. Accuracy

Transform encoders may receive textual or visual input from VLP models, allowing them to fully use unimodal pre-trained models. Specifically, VLP models generate

use standard transformer encoders and random initialization to create textual or graphic representations. Furthermore, ViT-PF signals like ViT and DeiT may be encoded by VLP models using pre-trained visual transformers. Instead of being concatenated, text and images are transmitted independently to two distinct transformer blocks in a dual-stream architecture. There are no parameters shared by these two Transformer blocks. Cross-modal interactions are facilitated by cross-awareness, which enhances performance. Additionally, there should be no mutual attention between the Text Transformer and Visual Transformer blocks for efficiency's sake. One transformer block with a single stream architecture delivers both text and graphic capability. Single-stream structures integrate multimodal inputs using merged attention.The cross-attention approach is utilized for V-L interactions in the dual-stream architecture, as opposed to the self-attention operation of the single-stream design, where the key-value vector is acquired from one modality and the query vector is retrieved from the other modality. To mimic The mutual attention layer typically contains two unidirectional sublayers: visual-to-audio and audio-to-visual. Separating intramodal and cross-modal interactions is necessary for the dual stream architecture to produce a better multimodal representation. It uses two transformers to further represent the intra-modality interaction after the cross-modal module. To improve internal connections, LXMERT adds a self-attention sub-layer after a cross-attention sub-layer, rather than adding more transformers.n the cross-modal sub-layers, the two streams share the same attention module settings. ALBEF divides intra-modal and cross-modal interaction more well by using two distinct transformers for texts and pictures prior to cross-attention. This architectural design facilitates a more complete encoding of the input. However, the extra feature encoder makes it less efficient with respect to parameters.

B. X-VLM

Most existing approaches for vision-language pretraining rely on object-centric features extracted through object detection and allow for fine-grained alignment between the derived features and texts. For these systems, learning relationships among many different things is a challenge. To achieve this, we propose a new method for multi-grained vision language pre-training called X-VLM 1. Learning multi-grained alignments involves finding visual concepts in an image given the corresponding texts and simultaneously aligning the texts with the visual concepts (where the alignments are multi-granular). Experiment results show that X-VLM consistentlysurpasses cutting-edge methods and effectively uses the multi-grained alignments it has learned for several downstream vision language tasks. We compare the 4M and 16M settings, respectively, between X-VLM and the existing approaches. To prepare our pretraining data (CC), we use two datasets from within the domain, Visual Genome (VG) and COCO, as well as two datasets from outside the domain, SBU Captions and Conceptual Captions.In the approximately 4M configuration, we only use image annotations from COCO and VG, which have approximately 2.5M object annotations and 3.7M region annotations. Remember that the same set of object annotations was used to train BUTD, the most widely used object detector. VG region annotations are also utilized by the existing methods of just learning image text alignments, with the understanding that area

descriptions can adequately represent the entire image. Rather than using the item labels as text descriptions of the objects, we re-formulate the picture annotations, producing an image with multiple boxes, each of which is linked to a text. The text describes the visual concept in the box, which could be an area, an object, or the image itself.We make use of the Conceptual 12M dataset, which is much noisier.

Following ALBEF, use (CC-12M) in the 16M setting. We also use Open Images and Objects365 after VinVL.We

As most downstream V+L tasks are built on top of COCO and VG to prevent information leakage, remove any images that are also present in the downstream tasks' validation and test sets. Since there are some overlaps and COCO and VG are sourced from Flickr, we additionally employ URL matching to remove any co-occurring Flickr30K images

VII. IMPLEMENTATIONS

The image encoder for X-VLM is Vision Transformer, which is initialized with Swin of Transformerbase. Six transformer layers are present in both the text encoder and the cross-modal encoder. The text encoder is initialized using the first six layers of the BERTbase, and the cross-modal encoder is initialized using the last six layers of the BERTbase. X-VLM has 215.6M total parameters for pre-training. X-VLM receives images with a resolution of 224 x 224. We restrict text input to a maximum of 30 tokens.

While fine-tuning, we increase the image resolution to 384X384. There is a mixed level of precision in pretraining. We sample the data by generating bounding box annotations for half of the batch photos. utilizing the optimizer's weight decay setting of 0.02.The learning rate follows a linear plan, warming up to 1e-4 from 1e-5 in the first 2500 steps before falling back to 1e-5. There will be a text encoder and a cross model encoder in the X-VLM we looked at. Each of them has six transformers in total. Here, we use the first six layers of the BERT base to initialize the text encoder. The final six layers are used to initialize the cross-model encoder, which is the other encoder. For pretraining purposes, our X-VLM comprises 215.6M parameters in total. Thus, the input image resolution is 224*224. The maximum number of tokens that can be entered for the text input is thirty. As we continue to fine-tune, we inherently

raise the 384*384 input resolution. Furthermore, we interpolate the image's positional embeddings in this instance.patches to improve eyesight.

Afterwards, we mix precisions to perform the pretraining. We have 4M and 16M settings in X-VLM, as was previously observed. With 8 NVIDIA 100 GPUs and a batch size of 1024, we attempt to train the model for 200k steps in order to accommodate 4M. This process can take up to three and a half days. On the other hand, we will train the 16M model using just 24 GPUs and a batch size of up to 3072. By dividing the data into half-half batches with boundary annotations, we test the samples. This is how the X-VLM for the vision pretraining program is being implemented.

A. Image processing and caption generation

.

VIII. LIMITATIONS

Apart from Contrastive Multimodal Attention (CMA), TCL (Text-Image Contrastive Learning) introduces an intra-modal contrastive objective to offer additional advantages in representation learning. To harness localized and structural information from both image and text inputs, TCL goes further by maximizing the mutual information (MI) between local regions of the image/text

and their global summaries. However, it is important to note that, conversely, the learned representation might exhibit a bias towards features prevalent in the existing data. In cases where certain groups are underrepresented, this bias may lead to suboptimal performance on those particular groups.

IX. FUTURE WORK

To learn the effectiveness of the newly proposed modules (i.e., IMC and LMI) in improving multimodal representation learning, we performed ablation studies on the impact of training on larger-scale datasets, by using IMC (w/o aug) on image-text retrieval task, and the larger scale dataset gave a significant boost in performance. We believe that if our model is pretrained on other large datasets, it has the potential for further development.We believe that our model may be improved further if it is pre-trained on more huge datasets.We look into the significance of the momentum coefficient m in more detail and find that the optimal performance is reached when m = 0.5 achives the best results.

X. CONCLUSION

We describe a new vision-language pre-training system using triple contrastive learning (TCL) in this research study. To guarantee that the acquired representations are meaningful inside each modality, TCL additionally takes intra-modal supervision into account, in contrast to previous research that solely employed a cross-modal contrastive loss to align picture and text representations. Joint multi-modal embedding learning and cross-modal alignment gain from this. Additionally, TCL presents the local MI, which optimizes the mutual information between the local information derived from image patches or text tokens and the global representation, in order to use the structural and localized information in representation learning. Based on widely-known benchmarks, experimental results show that TCL performs far better than existing SOTA techniques.

XI. REFERENCE

1) Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910, 2019.
2) Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Finch.
3) Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3558–3568, 2021.
4) Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. arXiv preprint arXiv:2202.09061, 2022.
5) Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.
6) Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.
7) Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057, 2021.
8) Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng,

and Jingjing Liu. Uniter Universal image-text representation learning. In European conference on computer vision, pages 104–120. Springer, 2020.

9) Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702–703, 2020.
10) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
11) Github Reference for code Implementation `https://github.com/uta-smile/TCL`