Seeing What You Miss: Vision-Language Pre-training with Semantic Completion Learning

Yatai Ji1* Rongcheng Tu2* Jie Jiang2* Weijie Kong2 Chengfei Cai2 Wenzhe Zhao2 Hongfa Wang2 Yujiu Yang1† Wei Liu2†

1Tsinghua University 2Tencent

jyt21@mails.tsinghua.edu.cn turongcheng@gmail.com
{zeus, jacobkong, fletchercai, carsonzhao, hongfawang}@tencent.com
yang.yujiu@sz.tsinghua.edu.cn w12223@columbia.edu

Abstract

Cross-modal alignment is essential for vision-language pre-training (VLP) models to learn the correct corresponding information across different modalities. For this purpose, inspired by the success of masked language modeling (MLM) tasks in the NLP pre-training area, numerous masked modeling tasks have been proposed for VLP to further promote cross-modal interactions. The core idea of previous masked modeling tasks is to focus on reconstructing the masked tokens based on visible context for learning local-to-local alignment. However, most of them pay little attention to the global semantic features generated for the masked data, resulting in a limited cross-modal alignment ability of global representations. Therefore, in this paper, we propose a novel Semantic Completion Learning (SCL) task, complementary to existing masked modeling tasks, to facilitate global-to-local alignment. Specifically, the SCL task complements the missing semantics of masked data by capturing the corresponding information from the other modality, promoting learning more representative global features which have a great impact on the performance of downstream tasks. Moreover, we present a flexible vision encoder, which enables our model to perform image-text and video-text multimodal tasks simultaneously. Experimental results show that our proposed method obtains state-of-the-art performance on various vision-language benchmarks, such as visual question answering, image-text retrieval, and video-text retrieval.

1. Introduction

Our real-world contains a wide variety of information, such as texts, images, sounds, etc. For a powerful general artificial intelligence system, it is necessary to capture the se-
language features mainly by using various masked modeling tasks, such as masked language modeling (MLM) and masked vision modeling (MVM). As shown in Fig. 1(a), the basic idea of MLM and MVM is self-reconstructing the masked tokens via leveraging informative visible tokens to realize local-to-local alignment. Specifically, MLM adopted by BERT [24] is to predict the original vocabulary IDs of the masked words. Inspired by the success of MLM in pre-training, there is a flourishing trend to extend it to visual pre-training tasks. Generally, by masking some visual patches, MVM tasks predict their original pixels [13, 17], corresponding discrete tokens [4, 10, 46] generated by the VQ-VAE variants, or Histograms of Oriented Gradients (HOG) features [11], etc.

These masked modeling tasks only focus on reconstructing the local masked tokens, and pay little attention to recovering the missing global semantic information caused by data corruption. The token-level reconstruction may lead to inadequate learning of global representations for cross-modal information. As illustrated in Fig. 1(b), in the situation of token-level reconstructions, the global representation is disordered in its attention on the other modality. It implies that the global-to-local alignment ability of the pre-training model is limited, leading to a degraded global representation. However, the global semantic features have a great impact on the performance of the pre-training model as they are usually used to deal with downstream tasks. Therefore, it is crucial to ensure the global semantic features to learn more accurate global-to-local alignment.

Intuitively, considering that the paired vision and text data are two views of the same semantic information, the missing semantics of masked data can be completed by capturing information from the other modality. From this point of view, we propose a novel pre-training task called Semantic Completion Learning (SCL). Specifically, SCL is composed of dual parts: masked vision semantic completion (MVSC) and masked language semantic completion (MLSC). As shown in Fig. 1(a), MVSC (MLSC) exploits information of complete text (vision) data to recover the global semantic representations of masked vision (text) data. In this way, the model can generate representative global features with accurate global-to-local alignment. For example, as illustrated in Fig. 1(b), compared with the model pre-trained without SCL, the attention maps with SCL pre-training are more discriminative and reasonable.

For the architecture of the vision-language pre-training model, we adopt a general framework that consists of two uni-modal encoders and a fusion encoder. Moreover, we present a flexible vision encoder to enable our model to perform image-text and video-text multimodal tasks simultaneously. Specifically, for video inputs, the vision encoder only adds a few additional learning parameters, and the [CLS] feature of each frame is treated as a bridge associating spatial modeling within the frame and temporal modeling among frames. Inspired by curriculum learning [3], we train the model with image-text and video-text datasets successively to transfer visual knowledge from images to videos.

In a nutshell, our contributions are three-fold. (1) To enhance the global-to-local alignment of global representations, we propose a new pre-training task called Semantic Completion Learning (SCL), which recovers missing semantic information from unmasked data, promoting learning more representative global features. (2) We design an adaptive vision encoder, which can transfer multimodal pre-training knowledge between images and videos readily. (3) We conduct multiple vision-language downstream tasks to demonstrate the generalization of semantic completion learning, and the vision encoder, including visual question answering, visual reasoning, image-text retrieval, and video-text retrieval. Our model SCL achieves state-of-the-art performance based on a similar pre-training data scale. Our code is available at https://github.com/IIGROUP/SCL.

2. Related Works

2.1. Vision-Language Pre-training

Existing vision-language pre-training works can be divided into two categories: dual-tower and cross-fusion architecture.

The dual-tower architecture based methods [1–3, 12, 22, 39, 48] employ two individual encoders to separately extract the features for the visual data (images or videos) and textual data, and then map these features into a common semantic space. Among them, CLIP [39] exploits contrastive learning with a huge quantity of noisy image-text pairs directly collected from the Internet, achieving remarkable results on plenty of vision-language tasks. Similarly, FROZEN [3] proposes a curriculum learning schedule to train the vision-language model on both image-text and video-text datasets by treating an image as a single-frame video. Although these two-stream architecture based methods perform well on cross-modal retrieval tasks with high efficiency, their performances on the more complex multimodal downstream tasks are not inspirational due to the insufficient interaction between local vision and text features.

To overcome this limitation, the cross-fusion architecture based methods [5, 9, 30, 31, 43] have been proposed, which employ a cross-modal fusion encoder to enhance the interactions between vision and text features. For example, ALBEF [1] not only aligns the image and text features with contrastive learning but also feeds them into a cross-modal attention-based encoder to obtain the fused features. Clover [18] improves cross-modal feature alignment and fusion via a tri-modal alignment pre-training task. Our model also conducts multimodal feature fusion to achieve encouraging performance on more downstream tasks.
2.2. Masked Modeling Tasks

Recently, various masked modeling tasks have been proposed, whose strategy is self-reconstructing the masked data. Masked Language modeling (MLM) adopted by BERT [24] is the most classical one. It randomly masks some tokens of the input and then predicts the original vocabulary IDs of the masked words based on their context. By pre-training with the MLM, BERT achieves state-of-the-art results on eleven natural language processing (NLP) tasks. Inspired by the success of MLM in NLP, some works extend it into the visual domain and propose masked vision modeling (MVM). For example, VLMAE [17] proposes the Regional Masked Image Modeling (RMIM) task to facilitate the fusion of multimodal features. The RMIM masks some patches of an input image and then reconstructs the original pixels depending on the visible patches and the corresponding text. Similarly, VIOLET [10] proposes a masked visual-token modeling task, which first maps the original video frame patches into discrete visual tokens and then recovers the corresponding visual tokens of masked patches to train a joint encoder for the vision-language fusion. However, these tasks focus on reconstructing local masked tokens, ignoring the recovery of global semantic information of the masked data after cross-modal interactions. Hence, we propose a novel semantic completion learning (SCL) task.

3. Approaches

In this section, we first introduce our pre-training objectives in Sec. 3.1, and then describe the model architecture in Sec. 3.2. Please refer to Appendix A for the whole architecture figure and the details of previous pre-training tasks.

3.1. Pre-training Tasks

3.1.1 Previous Pre-training Tasks

Contrastive Learning (CL). The input images and texts are projected into vision and language embedding spaces with two uni-modal encoders, respectively. We utilize contrastive learning to adjust the positions of semantic features, enforcing the paired image-text features close and negative samples far apart. Then the token-wise fusion is employed for features of different modalities in the unified semantic space.

Vision-Text Matching (VTM). VTM aims to determine the correspondence of an image-text pair. The model conducts a binary classification on the concatenation of vision and text global representations generated by the fusion encoder, which contributes to the overall alignment of different modalities.

Masked Language Modeling (MLM). MLM was first used as a pretext task in natural language processing and was later introduced to multi-modal pre-training. Following the text tokens masked out with a probability of 15%, the model attempts to predict the original words based on visual information and textual context. The token-level reconstruction task plays an important role in the way that the model learns to associate linguistic words and visual entities, realizing local-to-local semantic alignment.
3.1.2 Semantic Completion Learning (SCL)

It is significant for the model to learn multi-modal information fusion, that is, to extract knowledge from the other modality. Instead of local information reconstruction in former masked modeling tasks, we expect that the model can also recover the global semantics of masked images or texts after cross-modal interaction.

As shown in Fig. 2, for each data pair, we first randomly mask the image and text separately to get \{I_{mask}, T\} and \{I, T_{mask}\}, so that the masked one manages to learn semantic information from the other complete modality. Then the two couples of data are sent to the model respectively. The recovered features of masked data are obtained by leveraging information from the other modality to complete its missing semantic information:

\[
\begin{align*}
I_{Re}, T_{Co} &= \text{Model}(I_{mask}, T), \\
I_{Co}, T_{Re} &= \text{Model}(I, T_{mask}),
\end{align*}
\]

where \(I_{Re}, T_{Re}\) are recovered global features of masked data, and \(I_{Co}, T_{Co}\) refer to global features of complete data. Then we conduct masked vision semantic completion (MVSC) and masked language semantic completion (MLSC) simultaneously. Specifically, we bridge the gap between recovered global features of masked data, the global representations will learn supplementary knowledge from the other modality to complete its missing semantic information:

\[
\text{NCE}_V = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s(I_{Re}, I_{Co})/\tau)}{\sum_{n=1}^{N} \exp(s(I_{Re}, I_{Co})/\tau)},
\]

\[
\text{NCE}_L = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s(T_{Re}, T_{Co})/\tau)}{\sum_{n=1}^{N} \exp(s(T_{Re}, T_{Co})/\tau)},
\]

where \(s\) denotes cosine similarity and \(\tau\) serves as the temperature hyper-parameter. The negative samples are global features of other complete images or texts in a batch. Note that \(I_{Co}\) and \(T_{Co}\) are detached for gradient backward, which makes the model more focused on the learning of recovering global features. Finally, the SCL loss is defined as:

\[
\mathcal{L}_{SCL} = \text{NCE}_V + \text{NCE}_L.
\]

By minimizing Eq.(3), it will make the global feature \(I_{Re}\) of the masked image similar to \(I_{Co}\) of the complete image \(T_{Re}\) similar to \(T_{Co}\). To recover the semantic information of masked data, the global representations will learn supplementary knowledge from the corresponding tokens of the other modality, i.e., accurate global-to-local alignment.

3.2. Model Architecture

Our model consists of three components: Vision Encoder, Text Encoder, and Fusion Encoder.

3.2.1 Vision Encoder

Input. The vision encoder takes visual data (a video or image) \(I \in \mathbb{R}^{M \times H \times W}\) containing \(M\) frame(s) of resolution \(H \times W\) as input, and when \(I\) is an image, \(M = 1\). The visual data \(I\) is first split into \(M \times N\) patches \(x \in \mathbb{R}^{M \times N \times D}\), where \(P \times P\) is the size of patches and \(N = HW/P^2\). Then, the patches \(x\) are transformed into \(M\) sequences of vision tokens \(V = \{v_i\}_{i=1}^{M} \in \mathbb{R}^{M \times N \times D}\), where \(v_i \in \mathbb{R}^{N \times D}\) denotes the sequence of tokens for the \(i\)th frame in the visual data and \(D\) denotes the dimension of vision tokens. Next, a learnable [CLS] token is concatenated to every token sequence \(v_i\), and we obtain \(V = \{v_i\}_{i=1}^{M} \in \mathbb{R}^{M \times (N+1) \times D}\). Finally, the tokens \(V\) are summed with learnable spatial positional embeddings \(\mathbf{E}^s \in \mathbb{R}^{(N+1) \times D}\) and temporal positional embeddings \(\mathbf{E}^t \in \mathbb{R}^{M \times D}\):

\[
g_{ij}^t = v_{ij} + \mathbf{E}_i^t + \mathbf{E}_j^s,
\]

where all patches in the same spatial location of different frames are given the same spatial positional embedding \(\mathbf{E}_j^s\), and all patches in the same frame share the same temporal positional embedding \(\mathbf{E}_i^t\).

VisualBlock. The pre-processed vision tokens \(G^0 = \{g_i^0\}_{i=1}^{M} \in \mathbb{R}^{M \times (N+1) \times D}\) are fed into the vision encoder which can process image and video data. The vision encoder is a modified ViT [8], containing a stack of VisualBlocks.

The detail of each VisualBlock is shown in Fig. 3. Specifically, each VisualBlock will perform temporal attention to exploit the global temporal information of the visual data, and conduct the spatial attention to capture sufficient local spatial semantic information. For temporal attention, we perform multi-head attention for the [CLS] tokens \(\{g_{0j}^i\}_{j=1}^{M}\) of all frames through attending to all \(M \times (N+1)\) tokens to produce [CLS] tokens \(\{g_{ij}^0\}_{j=1}^{M}\). Regarding spatial attention, it is the multi-head attention within each frame. Taking the \(i\)th frame as an example, we use \(\{g_{ij}^{0} - 1\}_{j=1}^{N}\) without [CLS] token as queries and all the \(N + 1\) tokens in the frame as keys and values to conduct attention and obtain the output.
Table 1. Performance comparison on VQA2.0 and NLVR2.

| Model          | VQA2.0 | NLVR2 |
|----------------|--------|-------|
|                | test-dev | test-std | dev | test-p |
| Pre-trained with >10M images |         |        |     |
| ALBEF(14M) [29] | 75.84 | 76.04 | 82.55 | 83.14 |
| SimVLM [48]    | 77.87 | 78.14 | 81.72 | 81.77 |
| OFA [45]       | 78.0  | 78.1  | -    | -     |
| BLIP [28]      | 78.25 | 78.32 | 82.15 | 82.24 |
| Pre-trained with <10M images |         |        |     |
| Oscar [31]     | 73.16 | 73.44 | 78.07 | 78.36 |
| UNITER [5]     | 72.70 | 72.91 | 77.18 | 77.85 |
| ViLT [25]      | 71.26 | -     | 75.70 | 76.13 |
| TCL [52]       | 74.90 | 74.92 | 80.54 | 81.33 |
| VLMo [47]      | 76.64 | 76.89 | 82.77 | 83.34 |
| METER [9]      | 77.68 | 77.64 | 82.33 | 83.05 |
| Ours           | 78.72 | 78.78 | 83.63 | 84.27 |

Table 2. Performance comparison of zero-shot image-text retrieval on Flickr30K.

| Model          | IR@1 | IR@5 | IR@10 | TR@1 | TR@5 | TR@10 |
|----------------|------|------|-------|------|------|-------|
| Evaluate pre-trained models directly |         |        |       |      |      |       |
| UNITER [5]     | 66.16 | 88.40 | 92.94 | 80.70 | 95.70 | 98.00 |
| ViLT [25]      | 55.00 | 82.5  | 89.8  | 73.24 | 93.6  | 96.5  |
| ALIGN [23]     | 75.70 | 93.80 | 96.80 | 88.60 | 98.70 | 99.70 |
| METER [9]      | 79.60 | 94.96 | 97.28 | 90.90 | 98.30 | 99.50 |
| Ours           | 79.74 | 95.46 | 97.86 | 91.70 | 99.30 | 99.90 |

Evaluate models fine-tuned on COCO

| Model          | IR@1 | IR@5 | IR@10 | TR@1 | TR@5 | TR@10 |
|----------------|------|------|-------|------|------|-------|
| ALBEF [29]     | 76.8  | 93.7 | 96.7  | 90.5 | 98.8 | 99.7  |
| ALBEF [39]     | 82.8  | 96.3 | 98.1  | 94.1 | 99.5 | 99.7  |
| TCL [52]       | 79.6  | 95.1 | 97.4  | 93.0 | 99.1 | 99.6  |
| Ours           | 81.74 | 96.72 | 98.54 | 94.80 | 99.60 | 100.00 |

4.2. Evaluation Results

4.2.1 Image-Text Understanding

We conduct multimodal understanding tasks on VQA2.0 and NLVR2, which require the model to exploit vision and language semantic fusion. As the results shown in Table 1, SCL achieves new state-of-the-art performance compared with previous models, implying that cross-modal fusion benefits from our new pre-training task. Specifically, when pre-trained with fewer than 10M images, our model outperforms METER [9] by +1.04 and +1.14 scores on VQA2.0 test-dev and test-std. On NLVR2, we gain +0.86 and +0.93 score improvements over VLMo [47], respectively. Moreover, our model pre-trained with 4M images also surpasses some models with more than 10M images, for instance, SimVLM [48], BLIP [28].

4.2.2 Image-Text Retrieval

We evaluate image-text retrieval in both zero-shot and fine-tuning scenarios. Our model achieves substantial perfor-
4.2.3 Video-Text Retrieval

Due to our adaptable vision encoder, the image-text pre-trained model can be readily transferred to video-text pre-training. We evaluate text-to-video retrieval on two popular datasets, MSRVTT and LSMDC, to prove the performance of the video pre-training model. Table 4 summarizes results under both fine-tuning and zero-shot settings. In the fine-tuning situation, compared with the previous SOTA model, SCL achieves notable performance improvements with +4.6% and +10.1% in R@1 on MSRVTT and LSMDC. When doing zero-shot retrieval, SCL also gains remarkable improvements over the existing methods with +4.8% and +3.4% in R@1 on MSRVTT and LSMDC, respectively. These results demonstrate that the knowledge of a comprehensive boost from previous methods, reaching 79.74% and 91.7% in terms of IR@1 and TR@1. When evaluated with the model fine-tuned on COCO, SCL outperforms models pre-trained on datasets of similar sizes, including ALIGN [23] and ALBEF [29]. Moreover, compared with ALBEF pre-trained on 14M images, SCL also has a more impressive performance in five out of six recall metrics, further demonstrating the effectiveness of our proposed strategies.

For fine-tuning experiments, our model surpasses previous models by a large margin, as shown in Table 3. TCL [52] has a distinguished retrieval performance with triple contrastive learning, which is cross-modal, intra-modal, and global-local. Compared with TCL, our method brings +1.14% and +10.1% in R@1 on COCO and Flickr30K, respectively. It is worth noting that our model also has higher scores than ALIGN [23] with 1.8B image-text pairs pre-trained. Thanks to semantic completion learning, the global features capture more cross-modal information, leading to an encouraging performance on retrieval.

| Model | COCO | Flickr30K |
|-------|------|----------|
|       | IR@1 | IR@5 | IR@10 | TR@1 | TR@5 | TR@10 | IR@1 | IR@5 | IR@10 | TR@1 | TR@5 | TR@10 |
| ALIGN [23] | 59.9 | 83.3 | 89.8 | 77.0 | 93.5 | 96.9 | 84.9 | 97.4 | 98.6 | 95.3 | 99.8 | 100.0 |
| ALBEF(14M) [29] | 60.7 | 84.3 | 90.5 | 77.6 | 94.3 | 97.2 | 85.6 | 97.5 | 98.9 | 95.9 | 99.8 | 100.0 |

Table 3. Performance comparison of fine-tuned image-text retrieval on Flickr30K and COCO datasets.

The results are also competitive with models pre-trained on larger datasets, such as ALIGN [23] and ALBEF [29]. In the fine-tuning phase, the model is trained with CL and VTM losses. During inference, first filter top-k candidates with vision and language encoders and then compute VTM scores for ranking.

To investigate the generalization ability of our model, we conduct zero-shot experiments on the Flickr30K dataset. As shown in Table 2, SCL achieves the best performance in both settings of zero-shot retrieval on Flickr30K. When we evaluate with the pre-trained model directly, SCL gains performance improvements on Flickr30K and COCO datasets with similar model sizes and pre-training data scales. The results are also competitive with models pre-trained on larger datasets, such as ALIGN [23] and ALBEF [29]. In the fine-tuning phase, the model is trained with CL and VTM losses.

During inference, for the sake of efficiency, we first filter top-k candidates with vision and language encoders and then compute VTM scores for ranking.

To investigate the generalization ability of our model, we conduct zero-shot experiments on the Flickr30K dataset. As shown in Table 2, SCL achieves the best performance in both settings of zero-shot retrieval on Flickr30K. When we evaluate with the pre-trained model directly, SCL gains
4.3. Ablation Studies

We conduct empirical ablation experiments on pre-training tasks and the vision encoder. Since pre-training is time-consuming, we use COCO and VG as pre-training datasets, which is also a common setting in previous works [19, 43, 50].

4.3.1 Different Pre-training Tasks

There are four pretext tasks in our method, including CL, VTM, MLM, and SCL. As summarized in Table 5, we explore the impact of each task on both retrieval and understanding datasets. The first row shows the results of our model with all pre-training tasks, and the second to fifth rows reflect the effect of removing each task separately. According to the chart, we observe that the retrieval performance drops most due to the lack of SCL when conducting retrieval with feature fusion. Specifically, SCL brings +3.38% and +6.20% boost in IR@1 and TR@1 on F30K-ZS. The model without MLM loses 2.13% in the accuracy of VQA2.0, which indicates that MLM has a great effect on multimodal understanding tasks. As for NLVR2, VTM has a relatively large impact. However, contrastive learning is only effective for retrieval in our model, which is perhaps because the other three pre-training tasks have already learned cross-modal fusion sufficiently for understanding tasks. Overall, comparing the first row with the fifth row, the model with SCL makes progress on all downstream tasks, which demonstrates that the model learns more accurate cross-modal alignment to generate representative global features.

Furthermore, SCL comprises MVSC and MLSC, whose effects we showcase in Table 6. According to the first three rows, either MVSC or MLSC can improve the performance of downstream tasks. We find that MVSC has a superior impact on retrieval tasks, which is probably because it improves the robustness of visual information understanding. In VQA2.0 and NLVR2, MLSC plays a more important role. Additionally, when combining the two sub-tasks, our model performs better in most metrics, which indicates that they are in synergy.

4.3.2 Mask Ratio in SCL

As shown in Table 7, we observe that the mask ratios of image and text affect downstream tasks, especially on zero-shot retrieval. VQA2.0 is less sensitive to the mask ratio because the model has been fine-tuned with a large amount of data. Considering the second to fourth rows, when the image mask ratio is fixed, the model with a text mask ratio of 0.4 has almost the best performance. Moreover, when the text mask ratio is set to 0.4, the results of the image mask ratio of 0.8 are the highest. We speculate that when the mask ratio is lower, semantic completion will rely more on intra-modal information and lack learning across modalities, leading to inferior performance. When the mask ratio is too high, the
small number of remaining tokens can only perform very limited cross-modal interactions. In conclusion, we choose 0.4 and 0.8 as text and image mask ratios, respectively.

4.3.3 Vision Encoder Design

To investigate the effectiveness of our designed vision encoder in processing video data, we compare it with two other variants: (1) The first variant, termed Mean Pooling, directly treats a video as $M$ separate images and then uses the mean pooling of $M$ [CLS] tokens as the video representation. (2) The second variant, termed Global CLS, is the vision encoder proposed by MCQ [14]. In this experiment, we pre-train the vision encoder and text encoder on the WebVid [3] dataset via contrastive learning and then conduct zero-shot cross-modal retrieval on the MSRVTT dataset. The experimental results are shown in Table 8, where the Frame CLS denotes our designed vision encoder. It can be found that Frame CLS achieves the best performance on both video-to-text and text-to-video retrieval tasks, which demonstrates the outstanding capability of video temporal modeling.

4.4. Visualization Analysis

To demonstrate that SCL boosts cross-modal alignment for global representations, we visualize cross-attention maps between text [CLS] and the whole image in the last layer of the fusion encoder. We conduct max-pooling on attention maps of 12 heads to draw heatmaps, as shown in Fig. 1(b) and Fig. 4. Compared with pre-trained by CL, VTM and MLM, the model with SCL can recognize relevant regions more precisely. For example, in the first image of Fig 1(b), the attention distribution of global text representation to the image is scattered without SCL, while after semantic completion learning, [CLS] pays attention to the fish, lemons, asparagus in the image. Taking the second image of Fig 4 as another example, the model pre-trained with SCL identifies the dishes and sink in the kitchen, which indicates a desirable global-to-local alignment ability.

Observing attention maps of 12 heads in Fig. 4, we find that the attention maps without SCL are basically the same for an image, but for the model pre-trained with SCL, the attention maps of different heads are distinctive, which means that each head learns various information from the image. Overall, semantic completion learning encourages global representations to learn cross-modal interactions, extracting useful knowledge from the other modality. There are more visualization cases in Appendix C.

5. Conclusion

In this paper, we proposed a new vision-language pre-training task called Semantic Completion Learning (SCL). Different from previous pre-training tasks that reconstruct masked local tokens, SCL leverages cross-modal interactions to recover global semantic information of masked data, promoting cross-modal alignment for global representations. Ablation studies and visualization analysis demonstrate the effectiveness of SCL. Moreover, we introduced a flexible vision encoder, which adapts to image-text and video-text multimodal tasks readily. We conducted image-text and video-text pre-training sequentially and applied our model to various challenging downstream tasks. The extensive evaluations validate the great superiority of our SCL method.

6. Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61991450) and the Shenzhen Science and Technology Program (JSGG20220831093004008; ZDSYS202008111142605016).
References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In NeurIPS, 2021. 2

[2] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In CVPR, 2022. 2

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In ICCV, 2021. 1, 2, 5, 6, 8

[4] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. VI-beit: Generative vision-language pretraining. CoRR, abs/2206.01127, 2022. 2

[5] 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 ECCV, 2020. 2, 5, 6

[6] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. CoRR, abs/2109.04290, 2021. 1

[7] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In NeurIPS, 2020. 12

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at 6797 scale. In NeurIPS, 2020. 12

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning with noisy text supervision. In ICML, 2021. 2

[10] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET : End-to-end video-language transformers with masked visual-token modeling. CoRR, abs/2111.12681, 2021. 1, 2, 3, 6, 12

[11] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. CoRR, abs/2209.01540, 2022. 2

[12] Valentin Gabeur, Chen Sun, Katheek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In ECCV, 2020. 2

[13] Hongyu Gao, Chao Zhu, Mengyin Liu, Weibo Gu, Hongfa Wang, Wei Liu, and Xu-Cheng Yin. Calic: Accurate and efficient image-text retrieval via contrastive alignment and visual contexts modeling. In ACM Multimedia, 2022. 2

[14] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In CVPR, 2022. 1, 6, 8, 12

[15] Yuying Ge, Yixiao Ge, Xihui Liu, Alex Jinpeng Wang, Jianting Wu, Ying Shan, Xiaohu Qie, and Ping Luo. MILES: visual BERT pre-training with injected language semantics for video-text retrieval. CoRR, abs/2204.12408, 2022. 6

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Proc. of CVPR, 2017. 5

[17] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Chen Wu, Xiujun Shu, and Bo Ren. VLMASE: vision-language masked autoencoder. CoRR, abs/2208.09374, 2022. 1, 2, 3

[18] Jingjia Huang, Yinan Li, Jiashi Feng, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards A unified video-language alignment and fusion model. CoRR, abs/2207.07885, 2022. 2, 6, 12

[19] Zhicheng Huang, Zhaoyang Zeng, Yupon Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In CVPR, 2021. 6, 7

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. CoRR, 2020. 6

[21] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujii Yang. MAP: modality-agnostic uncertainty-aware vision-language pre-training model. CoRR, abs/2210.05335, 2022. 1

[22] Chao Jia, Yinfeng Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsun Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021. 2

[23] Chao Jia, Yinfeng Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsun Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021. 5, 6

[24] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019. 2, 3, 11

[25] Wonjae Kim, Bokyoung Son, and Ilبدو Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In ICML, 2021. 5

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis., 2017. 5

[27] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C. H. Hoi. Align and prompt: Video-and-language representation learning with entity prompts. In CVPR, 2022. 6

[28] Jannan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, 2022. 5

[29] Jannan Li, Ramprasath Selvaraju, Akhilone Gomtare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS, 2021. 5, 6
