Part of speech masking training vision language model

by

Pasit Tiwawongrut

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Data Science and Artificial Intelligence

Examination Committee:	Dr. Chaklam Silpasuwanchai
	Dr. Mongkol Ekpanyapong
	Dr.

Nationality:	Thai
Previous Degree:	Bachelor of Computer Engineering
	Khon Kaen University
	Thailand
Scholarship Donor:	Asian Institute of Technology

Asian Institute of Technology
School of Engineering and Technology
Thailand
December 2023

ACKNOWLEDGMENTS

ABSTRACT

Five sentences:

1) background - very specific background; hint the problem 2) problem - very very measurable problem; start with a signal word like ”However”, ”Anyhow”, ”Despite” 3) solution/what you do - Use verb wisely; explore/investigate/develop/compare 4) key findings (2-3 sentences) - summarize ONLY the key findings - it means interesting findings 5) contributions - why this is important to be solved; what impact it can bring

Exercise: within 15 mins, write down these five sentences, and then put on the chat.

NONE of you are qualified to stay away from this format. Who is qualified: very very competent writer.

Let’s analyze

Adding salt enhances the positive sensory attributes of foods (subjective).

However, consuming too much salt can raise your chances of enlarged heart muscle, headaches, heart failure, high blood pressure, kidney disease, kidney stones, osteoporosis, stomach cancer, and stroke (you write too much).

This study compared 18 flavorful salt alternatives. (but why?, where are these 18 things come from? what is the objective? no link with the problem)

The results showed that lemon juice outperformed other alternatives to brighten up the flavor of dishes (performance of what? how did you measure flavor? what does ”brighten up” mean?). Further discussion and implications were made.

This study comprehensively compares salt alternatives which can be applied to existing menus (this author does not think about real scenario....).

Let's analyze

Deep learning models are considered as blackbox and really hard to interpret (what is NOT considered as blackbox?) (when you write, NO emotion......We never say always, never, really) (what does "interpret" mean???)

Specifically, when model makes mistakes in dialog system as model misuderstand the intent of users (huh? dialog system? it's of course.....i don't understand what you want to say)

Most model do this task are lack of interpretability (overclaim....i don't even understand what does interpretability means.....and the author even said "MOST"....).

Thus, it really is hard to track or imporve the system accordingly (until now....i still don't what the authors want to do....).

However, in dialog system, we can trace model by probing component play roles based on decisions (what does "trace" means, what does "probing" means? what does "play roles" mean? what does "decision" mean?.

The benefits of this work is to help debuging models (you never talk about debuging...).

Let's analyze

Several paradigms have been used to develop BCI Spellers to help people with neurological disorders (so i assume you will make a speller and test with people with disorders; i also assume you will develop new paradigms or new speller).

However, researchers are still working on various techniques to make the Spellers efficient (too general problem.....a good problem usually help us imagine a good solution.....)

This paper developed a speller combining P300 potential and Steady State Visually Evoked Potential (SSVEP) paradigms, which is faster and more reliable (what does reliable means?) than the existing spellers (because your problem is not clear, i don't know whether it's a new work or not....).

We found that the hybrid speller improves the performance tremendously by improving the ITR (when you first write abbreviations, need full name) to 120 bits/min. (compare to what?)

This finding brings forward a new approach of developing of an efficient BCI Speller (what does efficient means?).

Let's analyze

Competitive online action video games have become quite popular among a large fraction of people, given its popularity, their effect on cognition have become an important topic for research (what does action mean? what does cognition mean? can be shorter....).

Despite many studies investigating the cognitive impacts of competitive online action video games (very long) exist, majority of them are cross-sectional (what does cross-sectional mean?) and lacking an active control group to compare with.

This study compares the executive cognitive functions (what are executive cognitive functions...), curiosity and aggression of a candidate online action video games (PUBG) (why PUBG????) against a brain training game (why brain training game) in a longitudinal (how long?) approach.

The study found that the cognitive effect of both of the concerned games is increasing (i still not sure how you measure....) upon the training for 2 weeks (you should put two week in the above sentence) in terms of processing speed, working memory, task-switching ability and fluid intelligence (why these four....how about curiosity, aggression?).

contribution sentence?

Let's analyze

Diabetes is one of the highest chronic disease (but i think you can use more words to be more specific you want to do....i think is too general...i cannot imagine what you want to do).

Research found that regular self-measurement of blood glucose enhances the patient's ability to self-regulate.

but self monitor of blood sugar, especially, continuous is impossible because current feasible methods are invasion method (a little bit headache.....i still could imagine your solution).

However, existing methods are mostly invasive which do not enable continuous and easy self-measurement methods

This paper develop a mean to measure blood glucose with a non-invasive continuous method by exploiting the advancement of Raman technology (not so bad...). (he did not mention about he will find the best spot....)(he did not mention anything about ML (he did not mention what is this paradigm he will work on...) (did not mention wearable.....)

we found that the performance of nour method is not only reach the clinical level but also comparable to the old invasive fashion. the best spot to measure raman for this task is ... (secret) which out performance all previous finding in the past.

5. this research contributes in 1) best measuring spot 2) suitable ml for this task 3) exploring calibration paradigm 4) newly develop wearable device for blood

glucose monitoring. (please avoid the word ”best”....)

Let’s analyze

Question Answering (QA) systems enable users to retrieve exact answers for questions posed in natural language (you don’t need to say this....this is general knowledge). High-resource language e.g. English, Chinese etc. apporach good performance (so what? i still don’t know what you want to say....i could not imagine what would be the problem).

However, There are some gap in low-resource language (too general.....). Another challenges on thai language is how to tokenize word because this language do not have a white-space to seperate a word (there are many works - maybe 20 years already - we ALREADY know how to tokenize thai words very very well.....).

This paper explores how to improve prefomance on thai language (you should talk thai language since background....what is the problem with thai language).

Thus we will compare A,B,C which one is suitable for thai language. (why A, B, C???)(what experiments you will do)

Finally, we found that augment model with XXX techinque with A tokenizer can achive better perfomance.

Let’s analyze

Large pretrained transformer models using self-supervised learning such as BERT has attracted a lot of researchers (not so bad....maybe ok...but you can punch more if you are more specific...).

However, for low-resource language like Nepali, due to its fairly complex linguistic structures (no....you are subjective...), several feature extraction and preprocessing needs to be considered (many works already been done?) while training

traditional machine learning and deep learning models, but it lacks tools like a generic stemmer or a list of proper stop words (this problem does not sound challenging or interesting....).

no link between your problem and solution

This paper compares Nepali pre-trained language models with multilingual variants (why multilingual variants?) such as mBERT and xlm-RoBERTa models (why mBERT? why xlm-RoBERTa) with a very minimal pre-processing steps (i don't know what is your contribution....) and evaluate them to a Nepali text classification task.

Results show that, transformer models outperform traditional machine learning techniques by significant margin when given adequate amount of data. (not interesting findings......)

This research contributes in 1) Creation of a well-balanced text classification dataset for Nepali language with more data. 2) Finding the better model by fine-tuning Nepali transformers models on text classification tasks

Exercise: Read this abstract quickly. Try to identify what is 1) background, 2) problem, 3) solution, 4) results, 5) contributions - 15 mins....

Keywords: keyword1, keyword2.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Vision language (VL) models have gained significant attention due to their ability to perform both zero-shot and transfer learning, achieving high performance across numerous downstream tasks through pre-training with web-scale image-text pairs (Mo, Kim, Lee, & Shin, 2024; Z. Wang, Wu, Agarwal, & Sun, 2022; J. Zhang, Huang, Jin, & Lu, 2024). Many VL models incorporate masked language modeling (MLM) as a pre-training task, making it an important method to train VL models (J. Li et al., 2021; C. Li et al., 2022; Chen et al., 2020; W. Wang et al., 2023; Tan & Bansal, 2019). Typically, a subset of word tokens is randomly masked at a percentage during training, and the model is tasked with predicting these masked tokens using information from both visual and language modality. This masking approach has proven to enhance the alignment between visual and linguistic representations, boosting performance in VL tasks (Tan & Bansal, 2019).

Despite the widespread adoption of MLM in VL training, its effects on model performance, efficiency and training loss remain underexplored. Bitton, Stanovsky, Elhadad, and Schwartz (2021) demonstrated that many of the randomly masked tokens are often stop-words or punctuation, which the model can easily learn without any need for masking. Another study by Wilf et al. (2023) demonstrated that selectively masking infrequent words from the pre-training dataset can boost model performance on out-of-domain datasets during continued training. Additionally, Tou and Sun (2024) suggested that random masking causes the model to rely heavily on local text signals, and it result in inefficient and inconsistent interactions between modalities, leading to suboptimal performance. These findings emphasize the importance of strategic token selection in MLM to enhance VL model performance and efficiency.

In this work, we aim to address the gap in understanding how masking each part of speech (POS) impacts VL models. Each POS contributes distinctively to sentence meaning: nouns typically denote objects, while verbs describe actions and often demand contextual comprehension. By selectively masking different parts of speech, we can better understand how each category influences the alignment between visual and linguistic information. We furthur expand the experiment into fine-tuning situation, where the model had pre-trained already. The experiment is designed to answer the following questions:

1. How does masking each POS impact the performance, efficiency and training loss of VL pre-training models?
2. Which types of questions in visual question answering (VQA) task benefit most from masking specific parts of speech in VL models?
3. What are the effects of part-of-speech masking during the fine-tuning phase of VL models?

1.2 Objective

The objectives for our experiment are as listed.

1. Develop a pre-trained VL model to evaluate the impact of masking each POS on performance and training dynamics.
2. Benchmark the performance of our masking approach using specialized datasets (Parcalabescu et al., 2022; Shekhar et al., 2017), to gain a deeper understanding of masking effects.
3. Analyze the effects of part-of-speech masking in the fine-tuning setup to understand its effect and performance in VL task.

1.3 Scope

1. The training and testing datasets are both natural images.
2. The model architecture is a cross-attention model, chosen for its ability to jointly predict answers based on information from multiple modalities.
3. The fine-tuning dataset is in the same domain as the training dataset.

CHAPTER 2

LITERATURE REVIEW

This section of the literature review is organized around two key topics relevant to our study. The first topic addresses VL models, providing an overview of the model architectures recently used in VL models and discussing the choice of the base architecture for the VL model used in this research. The second topic MLM, an important pre-training approach that has improved VL model performance. Together, these sections provide a comprehensive overview of the methodological foundations of this study.

2.1 Vision-Language model

In the early stage of VL learning, the goal of training is to align fine-grained features of the image with text. Many works have adopted object detection to create fine-grained labels for the training images (Chen et al., 2020; Bao et al., 2022). However, the focus of VL training has shifted to using web-scale image-text pairs as a training set, demonstrating competitive performance, as shown by CLIP (Radford et al., 2021). Radford et al. (2021) proposed contrastive training for VL with a large-scale image-text pairs dataset by optimizing the alignment of image and text encodings from the same pair, which was proven to be scalable by Jia et al. (2021). This approach has become a foundational model for VL tasks (Bommasani et al., 2021).

Recent advancements in VL model training can be roughly categorized into three main methods. The first approach is a separate unimodal encoder for each modality, as seen in models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). This method is trained with the objective of aligning the intermediate outputs of each modality's encoding. The second method uses a cross-attention layer to fuse multimodal inputs, e.g., Flamingo (Alayrac et al., 2022), mPLUG (C. Li et al., 2022), LXMERT (Tan & Bansal, 2019), and ALBEF (J. Li et al., 2021). The cross-attention layer enables the model to fuse each modality more deeply. Finally, the third approach uses a single large attention model with concatenated image and text tokens as input, as in BEIT-3 (W. Wang et al., 2023), OSCAR (X. Li et al., 2020). This approach allows for early-stage fusion of each modality, though it requires the highest amount of computational resources. In this work, we adopt the cross-attention method as the base model due to its effectiveness in fusing multimodal inputs. Additionally, this approach allows the model to be trained using the MLM task.

2.2 Masked Language Modelling

MLM is a widely used pre-training method in language model (LM) training (Devlin, Chang, Lee, & Toutanova, 2018; Lan, 2019; Yu et al., 2022; S. Zhang et al., 2022; Guu, Lee, Tung, Pasupat, & Chang, 2020) as a self-supervised task. BERT (Devlin et al., 2018) proposed MLM as a pre-training task, which has been proven effective for pre-training language models. The MLM task involves replacing some input tokens with a special [MASK] token, and the model must predict the masked tokens based on the given unmasked tokens. In the field of VL models, many VL models have also adopted MLM as a training task to train the model to predict masked text based on visual information (J. Li et al., 2021; C. Li et al., 2022; Chen et al., 2020; W. Wang et al., 2023).

In the field of selective masking strategies in natural language processing, several works have further refined MLM to enhance training efficiency. ERNIE (Sun et al., 2019), SpanBERT (Joshi et al., 2020), and n-gram Masking (Levine et al., 2021) propose span masking instead of single-token masking, which forces the model to rely more on long-range dependencies rather than adjacent tokens, resulting in better performance compared to BERT (Devlin et al., 2018). Considering linguistic features, Yang, Zhang, and Zhao (2023) conducted a training analysis based on POS masking focused on LM training. The results showed that focusing the masking of non-function words (ADJ, ADV, NOUN, PROPN, and VERB) in the later stages of training can encourage the LM model to develop a better contextual understanding.

For selective masking in VL training, Bitton et al. (2021) introduced an object token masking strategy, selectively masking object tokens in image captions and pre-training the model. This approach achieved superior performance compared to random masking. Another study by Wilf et al. (2023) showed that selectively masking infrequent words from the pre-training dataset during continued training enhances model performance on out-of-domain datasets. Additionally, (Tou & Sun, 2024) proposed a curriculum-based masking strategy in which a reinforcement learning agent dynamically selects masking spans based on cross-modal interactions. This method improved the model's mulitmodalities understanding while reducing the dataset size needed for effective training. In this work, we conduct experiments to analyze the impact of each POS on results within a VL setting.

2.3 Part of speech tagging

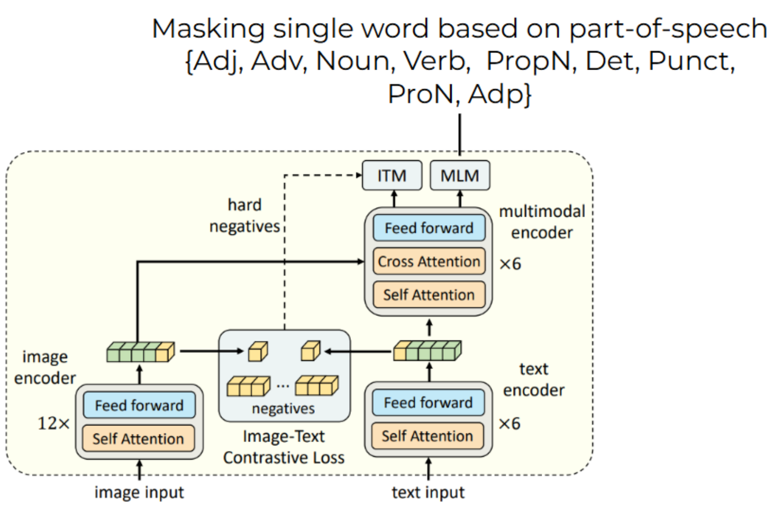
CHAPTER 3

METHODOLOGY

In this chapter, the methodology is detailed as follows. First, we describe the architecture of the model. Second, we explain the MLM pre-training loss functions used in this experiment. Third, the details of POS tagging are provided. Fourth, we outline the datasets used in this experiment. Fifth, we provide details on the visual question answering setup. Lastly, we provide implementation detail for the pre-training model.

Figure 3.1
Overall methodology

Pre-training the model with a MLM task by masking tokens based on the POS in the image captions.



3.1 Model architecture

As shown in Figure 3.1, our model includes three main components: an image encoder, a text encoder, and a multimodal encoder. The first component is the image encoder, for which we use ViT (Dosovitskiy et al., 2021), modified following (Radford et al., 2021), as the image encoder in this experiment. The second component is the text encoder, which employs a transformer architecture as BERT (Devlin et al., 2018) to encode image captions. The final component is the multimodal encoder, where VL interactions occur.

Given a training dataset D consisting of image-text pairs $(I_i, T_i) \in D$, where I_i is the image and T_i is the image caption of the i-th image, each image is first encoded as a sequence of tokens $\{v_{cls}, v_1, \ldots, v_n\}$ using ViT (Dosovitskiy et al., 2021). Here, v_{cls} represents the embedding of the [CLS] token prepended to the image patch sequence. In this experiment, the image encoder was initialized with ViT-B-32 pre-trained on ImageNet-21K (Deng et al., 2009). Next, we use a 6-layer transformer, randomly initialized, to encode the image caption T_i into text embeddings $\{w_{cls}, w_1, \ldots, w_n\}$, where w_{cls} is the embedding of the [CLS] token. Finally, both text and image encodings are passed through the multimodal encoder to fuse both inputs, producing multimodal encodings. For the multimodal encoder, a cross-attention layer is used, where both keys and values are the image encodings, and the text encoding serves as the query in the cross-attention layer.

3.2 Pre-training objectives

In this work, we pre-train our model with three objectives: masked language modeling (MLM), image-text constrative learning (ITC) and image-text matching (ITM).

3.2.1 Mask language modelling

Our model is trained with the MLM task. Typically, a percentage of tokens $\{w_1, \ldots, w_T\}$ are replaced with a special [MASK] token to create a masked caption T^{mask}. However, in this work, the masked tokens are selected based on POS type instead of randomly masking. The model is then trained to predict the original tokens at the masked positions, conditioned on both the unmasked tokens in T^{mask} and the visual features of I as $p^{\text{mask}}(I, T^{\text{mask}})$. Let y^{mask} be a one-hot vector representing the ground-truth vocabulary for the masked token, where the masked token has a probability of 1. The model's objective is to minimize the cross-entropy $\mathbf{H}$, given by:

$$\mathcal{L}_{\text{MLM}} = \mathbf{H}(y^{\text{mask}}, p^{\text{mask}}(I, T^{\text{mask}})S))$$

3.2.2 Image-text contrative learning

To improve each unimodal encoders representation, we use ITC to improve alignment of each modality. ITC aims to improve alignment by maximize similarity score of image and text from the same pair with score function $s(I, T) = v_{cls}^{\top} w_{cls}$, and minimize similarity score of image and text not from its pair. We then calculate softmax-normalized similiarity score for each image to any text and each text to any image, identified as image-to-text $p^{i2t} \in \mathbb{R}^M$ and text-to-image $p^{t2i} \in \mathbb{R}^M$ score as:

$$p_i^{i2t}(I) = \frac{\exp\left(s(I, T_i)\right)/\tau}{\sum_{m=1}^{M} \exp\left(s(I, T_m)/\tau\right.}, \quad p_i^{t2i}(T) = \frac{\exp\left(s(T, I_i)\right)/\tau}{\sum_{m=1}^{M} \exp\left(s(T, I_m)/\tau\right.}$$

where τ is a learnable temperature parameter. Let $y^{i2t}(I) \in \{0,1\}^M$ and $y^{t2i}(T) \in \{0,1\}^M$ be a ground truth with probability of 1 at a position of same pair, and probability of 0 on the otherhand. The ITC loss is calculated as cross-entropy $\mathbf{H}$ between p and y:

$$\mathcal{L}_{\text{ITC}} = \frac{1}{2}(\mathbf{H}(y^{i2t}, p^{i2t}) + \mathbf{H}(y^{t2i}, p^{t2i}))$$

3.2.3 Image-text matching

3.3 Part of speech tagging

3.4 Pre-training dataset

We pre-trained the model on the Conceptual Captions dataset (Sharma, Ding, Goodman, & Soricut, 2018), which consists of 3.3 million image-text pairs. In Conceptual Captions dataset, an automated process was used to select, filter, and refine these image-caption pairs to ensure they are clear, informative, and suitable for effective model training.

3.5 Evaluation

In this work, we evaluate each model trained with different types of POS masking using image-text retrieval and visual question answering tasks. The details of the evaluation methods and datasets are described in this section.

3.5.1 Image-text retrieval

For the image-text retrieval task, we evaluate the effect of masking on each POS category by performing zero-shot evaluations on the Flickr30K (Plummer et al., 2015) and VALSE (Parcalabescu et al., 2022) datasets for both image-to-text and text-to-image retrieval. The Flickr30K dataset is used to assess the model's overall performance in retrieval tasks. For a deeper understanding, the VALSE dataset provides a breakdown of linguistic phenomena into six categories: existence, plurality, counting, relation, action, and coreference. Each image caption in the VALSE dataset also includes a ”foil” version, in which words related to each caption category are altered. This approach enables us to analyze how different POS masking strategies impact the model retrieval performance and the alignment between visual and textual representations.

3.5.2 Visual question answering

For visual question answering task, we assess each model with

CHAPTER 4

Results

CHAPTER 5

DISCUSSION

CHAPTER 6

CONCLUSION

REFERENCES

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... others (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, *35*, 23716–23736.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., ... Wei, F. (2022). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, *35*, 32897–32912.

Bitton, Y., Stanovsky, G., Elhadad, M., & Schwartz, R. (2021). Data efficient masked language modeling for vision and language. *arXiv preprint arXiv:2109.02040*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., ... Liu, J. (2020). Uniter: Universal image-text representation learning. In *Computer vision – eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xxx* (p. 104–120). Berlin, Heidelberg: Springer-Verlag. Retrieved from `https://doi.org/10.1007/978-3-030-58577-8_7` doi: 10.1007/978-3-030-58577-8_7

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=YicbFdNTTy`

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). *Realm: Retrieval-augmented language model pre-training.* Retrieved from `https://arxiv.org/abs/2002.08909`

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (pp. 4904–4916).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, *8*, 64–77. Retrieved from https://aclanthology.org/2020.tacl-1.5 doi: 10.1162/tacl_a_00300

Lan, Z. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., & Shoham, Y. (2021). {PMI}-masking: Principled masking of correlated spans. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=3Aoft6NWFej

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., ... others (2022). mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, *34*, 9694–9705.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... others (2020). 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* (pp. 121–137).

Mo, S., Kim, M., Lee, K., & Shin, J. (2024). S-clip: Semi-supervised vision-language learning using few specialist captions. *Advances in Neural Information Processing Systems*, *36*.

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., & Gatt, A. (2022, May). VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8253–8280). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.567

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the ieee international conference on computer vision* (pp. 2641–2649).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of acl.*

Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., & Bernardi, R. (2017). "foil it! find one mismatch between image and language caption". In *Proceedings of the 55th annual meeting of the association for computational linguistics (acl) (volume 1: Long papers)* (pp. 255–265).

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., ... Wu, H. (2019). ERNIE: enhanced representation through knowledge integration. *CoRR*, *abs/1904.09223*.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing.*

Tou, K., & Sun, Z. (2024, June). Curriculum masking in vision-language pretraining to maximize cross modal interaction. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 3672–3688). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.naacl-long.203 doi: 10.18653/v1/2024.naacl-long.203

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., ... others (2023). Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 19175–19186).

Wang, Z., Wu, Z., Agarwal, D., & Sun, J. (2022). *Medclip: Contrastive learning from unpaired medical images and text.* Retrieved from https://arxiv.org/abs/2210.10163

Wilf, A., Akter, S. N., Mathur, L., Liang, P. P., Mathew, S., Shou, M., ... Morency, L.-P. (2023). Difference-masking: Choosing what to mask in continued pretraining. *arXiv preprint arXiv:2305.14577*.

Yang, D., Zhang, Z., & Zhao, H. (2023). *Learning better masking for better language model pre-training.*

Yu, W., Zhu, C., Fang, Y., Yu, D., Wang, S., Xu, Y., ... Jiang, M. (2022). *Dict-bert: Enhancing language model pre-training with dictionary.* Retrieved from https://arxiv.org/abs/2110.06490

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625-5644. doi: 10.1109/TPAMI.2024.3369699

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... Zettlemoyer, L. (2022). *Opt: Open pre-trained transformer language models.* Retrieved from https://arxiv.org/abs/2205.01068