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

Tables **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Vision language (VL) models have gained significant attention due to their ability to perform zero-shot and transfer learning, achieving high performance across numerous downstream tasks through pre-training with web-scale image-text pairs. Many VL models incorporate Masked Language Modeling (MLM) as a pre-training task, making it a fundamental approach for training VL models (J. Li et al., 2021; C. Li et al., 2022; Chen et al., 2020; Wang et al., 2023). Typically, a subset of word tokens is randomly masked at a fixed percentage during training, and the model is tasked with predicting these masked tokens using information from the visual modality. This masking approach has proven to enhance the alignment between visual and linguistic representations, significantly boosting performance in vision-language tasks.

However, the impact of MLM on VL training remains underexplored. Bitton, Stanovsky, Elhadad, and Schwartz (2021) demonstrated that many of the masked tokens are often stop-words or punctuation, leading the model to rely more on language patterns that do not require visual understanding. By focusing on masking objects instead, this approach showed improvements in both performance and efficiency compared to random masking (Bitton et al., 2021). Another study by Wilf et al. (2023) found that selectively masking infrequent words from the pre-training dataset during continued training enhances model performance on out-of-domain datasets. Furthermore, Tou and Sun (2024) introduced a curriculum masking scheme where a reinforcement learning agent selects masking spans based on cross-model interactions, leading to improved relational understanding with a reduced training dataset. These works underscore the importance of selecting

appropriate tokens for masking.

Despite the widespread adoption of MLM in VL training, its effects on model performance and learning dynamics remain underexplored. Bitton et al. (2021) highlighted that many of the randomly masked tokens are often stop-words or punctuation, which encourages the model to rely on linguistic patterns that require minimal visual understanding. To address this, they proposed masking object-related tokens, which led to notable improvements in model performance and training efficiency compared to random masking (Bitton et al., 2021). Similarly, Wilf et al. (2023) demonstrated that selectively masking infrequent words from the pretraining dataset can boost model performance on out-of-domain datasets during continued training. Additionally, Tou and Sun (2024) introduced a curriculum-based masking strategy, where a reinforcement learning agent dynamically selects masking spans based on cross-model interactions. This approach improved the model's relational understanding while reducing the dataset size required for effective training. These findings collectively emphasize the importance of strategic token selection in MLM to enhance VL model learning and efficiency.

In this work, we explore the effects of masking each part-of-speech category by masking each part-of-speech of each image captions as shown in figure 1.1. As each part-of-speech contributes uniquely to the meaning of a sentence, for instance, nouns typically represent objects, while verbs describe actions, which often require contextual understanding. We hypothesize that masking verbs is the best way to increase the VL model understanding, as verbs in a sentence represent interactions between objects and require the model to rely more on visual information. By selectively masking different part-of-speech, we can gain insight about how each part-of-speech category affects the alignment between vision and language modalities. The experiment is designed to answer the following questions:

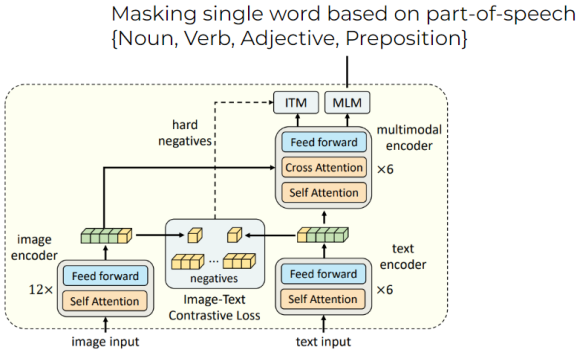
1. How does selective part-of-speech masking affect the alignment of vi-

sion and language modalities?

2. How does part-of-speech masking change the contribution of the vision and language modalities to the model's output?
3. How does specific masking impact the performance of vision question answering tasks, particularly in terms of improvements based on the type of question?

Figure 1.1
Overall methodology

Pre-training model with MLM task by masking token based on part-of-speech of the image captions.



1.2 Objective

The objectives for our experiment are as listed.

1. Pre-trained VL model for the experiment to identify the performance of masking in each part-of-speech.
2. Benchmarking our method against specialize dataset based on linguistic feature (Parcalabescu et al., 2022) for better understanding of the masking effect.
3. Analyze contribution from each modality of vision and language to the

prediction output based on MM-Shap (Parcalabescu & Frank, 2023).

1.3 Scope

1. The training and testing dataset are both natural image.
2. The model architecture is cross-attention model due to the ability of cross attention to jointly predicted answer based on another modality.

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 aim to align fine grain feature of the image to text. Many work have adopt object detection to create fine grain label of the training images (Chen et al., 2020; Bao et al., 2022). However, the idea of traning VL have shifted to web-scale image-text pairs as a training target with a competetive performance as demonstrated from CLIP (Radford et al., 2021). Radford et al. (2021) proposed contrastive training for VL with large scale image-text pairs dataset by optimize alignment of image and text encoding from the same pair, which is proved to be scalable by Jia et al. (2021), and has become a foundation model for VL task (Bommasani et al., 2021).

Recent advancements in VL model training can be roughly categorized into three main methods. The first approach is an individual unimodal model encoder for each modality, such as CLIP (Radford et al., 2021) and Align (Jia et al., 2021). This method is trained with the objective to align the intermediate output of each modality encoding. The second method utilizes a cross-attention layer to fuse multimodal input, e.g., Flamingo (Alayrac et al., 2022), mPlug (C. Li et al., 2022),

LXMERT (Tan & Bansal, 2019), and ALBEF (J. Li et al., 2021). With the cross-attention layer, the model can fuse each modality more deeply. Finally, the third approach is a single large attention model with the concatenation of image and text tokens as input, such as BEIT (Wang et al., 2023). This approach allows each modality to be fused in the early stage, although it requires the highest amount of computational resources. In this work, we adopt the cross-attention method as the base model due to its ability to fuse each modality input. Additionally, this approach enables the model to be trained using the MLM task.

2.2 Masked Language Modelling

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; Zhang et al., 2022; Guu, Lee, Tung, Pasupat, & Chang, 2020) as a self-supervised task. Devlin et al. (2018) had proposed MLM as a pre-training task, which had been proved to be effective for pre-training. MLM task is a task where some of the input tokens replace with special [MASK] token, and the model have to predict the masked tokens based on the given unmasked tokens. In the field of VL model, many VL model also adopted MLM as a training task to train the model to predict masked text to based on visual information (J. Li et al., 2021; C. Li et al., 2022; Chen et al., 2020; Wang et al., 2023). MLM have shown to be effective for training both

In the field of selective masking strategy, Yang, Zhang, and Zhao (2023) had present a training analysis based on POS masking focus on LM training, and proposed masking ratio decay to control the percentage of categories POS masking ratio. The result showed that, focusing on masking Non-function word (ADJ, ADV, NOUN, PROPN and VERB) in the late stage of training can forced the LM model to get better understanding of the context. For selective masking in VL training, Bitton et al. (2021) introduced an object token masking strategy,

where object tokens in image captions are selectively masked, and the model is pre-trained from scratch. This approach led to superior performance compared to random masking. Furthermore, Wilf et al. (2023) demonstrated that selectively masking infrequent words from the pre-training dataset during continued training enhances model performance on out-of-domain datasets. In this work, we focus on analysis effect and behavior of the VL model when training with masking each POS.

CHAPTER 3

METHODOLOGY

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

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*.

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).

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

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.

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

Parcalabescu, L., & Frank, A. (2023, July). MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4032–4059). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-long.223 doi: 10.18653/v1/2023.acl-long.223

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).

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).

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, 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`