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

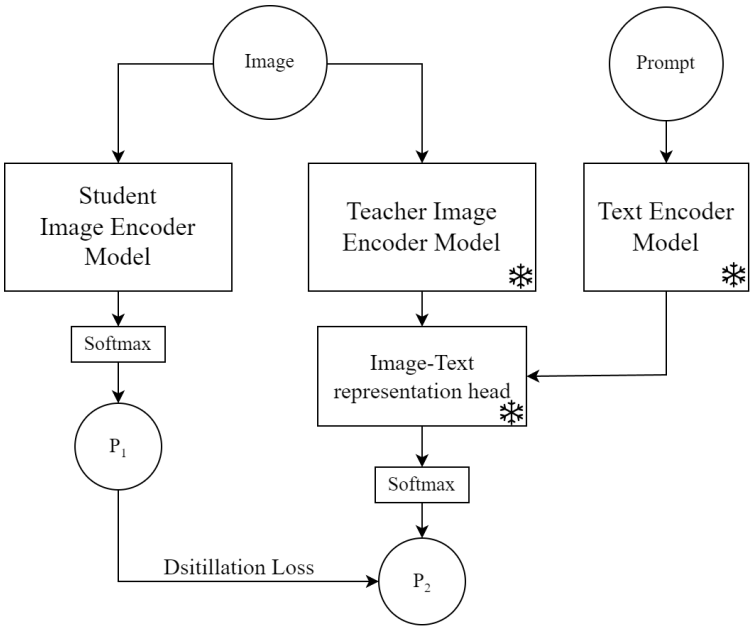
1.1 Background

Masked Language Modeling (MLM) is key task for training vision-language (VL) models(J. Li et al., 2021; C. Li et al., 2022; Chen et al., 2020; Wang et al., 2023). Most work randomly masked some word token to a fixed percentage in the training process, while expected model to predict the missing token based on vision modal. Bitton, Stanovsky, Elhadad, and Schwartz (2021); Wilf et al. (2023) suggest that by masking specific words in the training process often yield better performance.

However,

Figure 1.1
Overall methodology

Self-distillation training with image text joined representation.

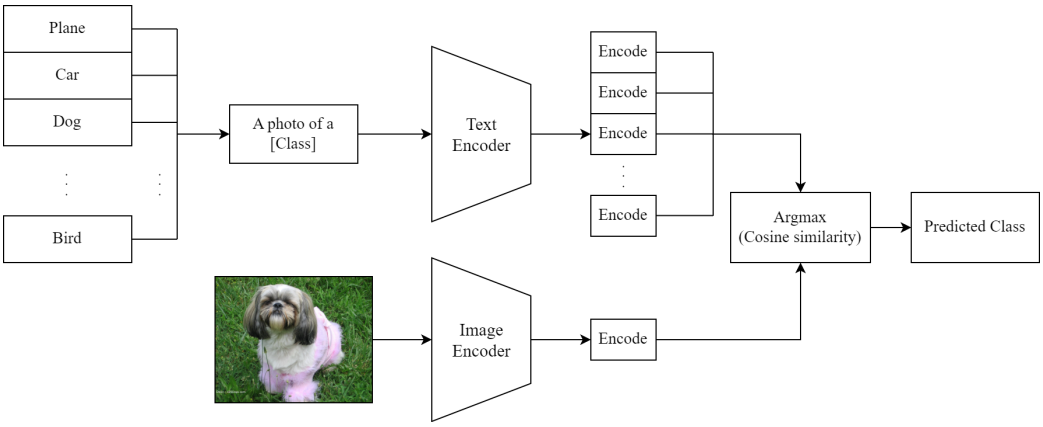


CHAPTER 2

LITERATURE REVIEW

2.1 Vision-Language model

Figure 2.1
CLIP Classification example



2.2 Knowledge Distillation and Self-Distillation

CHAPTER 3

METHODOLOGY

CHAPTER 4

Results

CHAPTER 5

DISCUSSION

CHAPTER 6

CONCLUSION

REFERENCES

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

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

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.

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.