Self-Distillation using image-language representation for image classification

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. Itthi Chatnuntawech

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

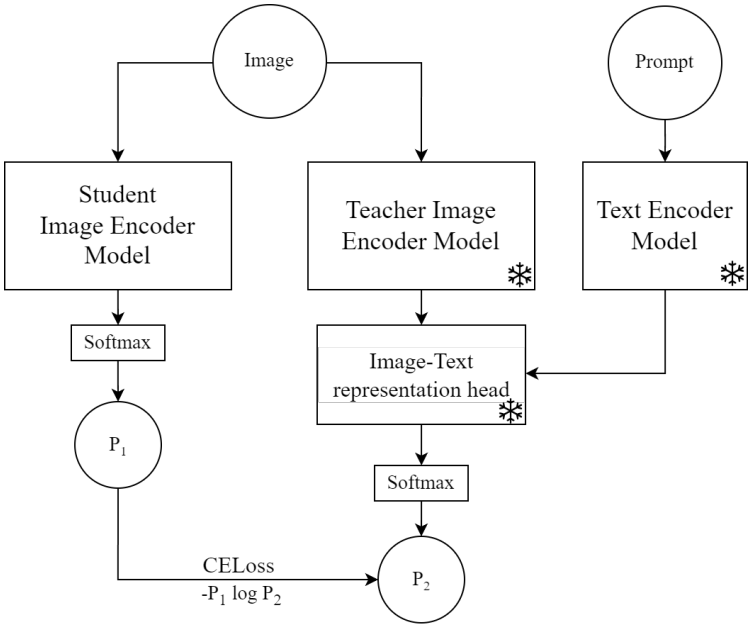
In computer vision, self-distillation (Furlanello et al., 2018; Zhang et al., 2019; Xie et al., 2020) is a technique for improving deep learning models without increasing model size. This paradigm involves training a student model whose parameter size is equal to the teacher model with new parameter initialization. One method from this paradigm can work without any label called Self-**di**stillation with **no** labels (DINO) (Caron et al., 2021). The method has been shown showed to improve the performance of both ResNet (He et al., 2016) and Vision Transformers (ViT) (Dosovitskiy et al., 2021). According to Allen-Zhu and Li (2023), when using the self-distillation technique, the student model is forced to learn soft-label features, which were extracted from the dataset. Additionally, by training the model with difference parameter initialization, the student model acquires knowledge from multiple views of images. The result shows around 2% improvement by the self-distillation method over multiple ResNet models (Zagoruyko & Komodakis, 2016).

In another branch of research, a multimodal approach demonstrates that the model's performance can be improved when combining both image and text data into the model. Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) and **A** **L**arge-scale **I**ma**G**e and **N**oisy-text embedding (ALIGN) (Jia et al., 2021) both achieved performance on par with fully supervised image classification across multiple benchmarks. These models are obtained by training the models with image-text pairs using the contrastive vision language pre-training method. The current state-of-the-art is **Co**ntrastive **Ca**ptioner (CoCa) (Yu et al., 2022). This approach used image-text pairs with contrastive language-image loss and image

captioning loss. It is a clear benefit of the training model in utilizing image and text information. Thus, this work investigated the effectiveness of using both texts and images with the self-distillation method.

Figure 1.1
Overall methodology

Self-distillation training with image text joined representation.



By merging the two paradigms, we proposed a new approach to train an image classification model by distilling knowledge from a multimodal teacher as shown in Figure **??**. A multimodal teacher model is constructed by leveraging a pre-trained language model and a pre-trained image encoder. The output of both encoders is combined using cross-attention and a linear classification layer, called "image-text representation head". The detail of the image-text representation head is described in Figure 3.2. In this work, the encoded text is used as a query to ex-

tract the relevant information from the image encoding. The student model, whose have the same architecture as the teacher image encoder model, is trained using teacher output as a target. Thus, the student learns with high-level semantic information.

The result showed that by combining textual information with images, our approach improved accuracy by 3% in both ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2021) model compared to the baseline self-distillation method. The ablation study shows that the student model achieves 3% higher accuracy by providing detailed descriptions in the training process. This suggests that by using the text encodings with cross-attention, the model extracts higher semantic information and more precise image representations from the images.

To summarize our contribution. Firstly this paper investigates the effectiveness of combining text-image representation by using text as a query to emphasize image representation in the self-distillation method. Secondly, this work proposes a method to efficiently combine textual information and image for self-distillation method. Lastly, this work also investigates the effect of prompt in our methods to create image descriptions for training.

CHAPTER 2

LITERATURE REVIEW

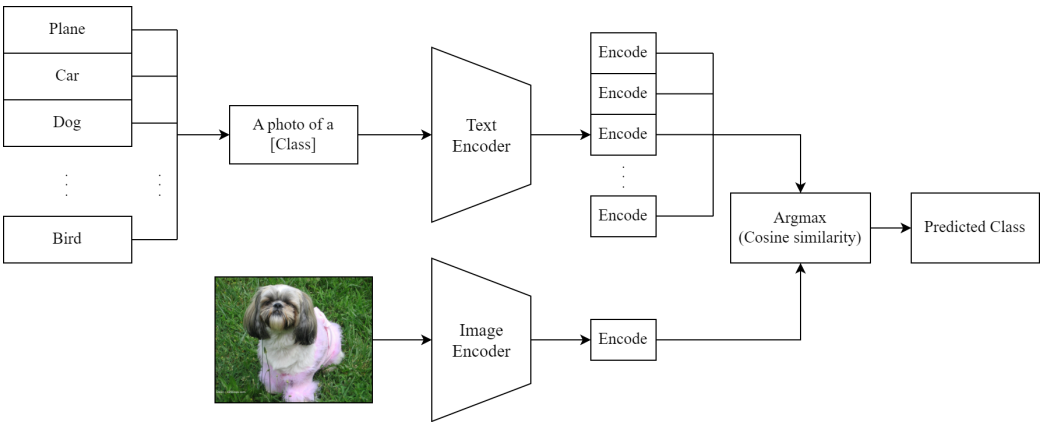
2.1 Vision-Language model

In the past few year, many work have shown the ability to utilize textual information with image task by training with image text pair, such as, Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), **A** **L**arge-scale **I**ma**G**e and **N**oisy-text embedding (ALIGN) (Jia et al., 2021) By training with huge amount of the image-text pair dataset, the ALIGN model can make up with the noisy image description and surpass the model, which is trained with benchmark dataset in zero shot image classification task. Recently **Co**ntrastive **Ca**ptioner (CoCa) (Yu et al., 2022) have propose vision-language encoder-decoder model which is training with image-text contrastive loss and captioning loss and add cross attention to join image-text modality could perform linear probing image classification on ImageNet with top-1% 90.6% accuracy. In this research we adopt two stream encoder same as CLIP and add cross attention to create image-text representation for classification. todo: Add detail about blip model

2.2 Knowledge Distillation and Self-Distillation

Knowledge Distillation was firstly proposed by Hinton et al. (2014) to compress the model size while maintain the model performance as much as possible. The method contains a smaller student model and a single or multiple larger teacher models. The knowledge is transferred by optimize the student model output to match the teachers output. Furlanello et al. (2018) has investigate knowledge distillation using a student model size same as teacher model, but and the result showing improvement of the student model. Such a method is call self-distillation. The self-distillation has been widely use in semi-supervised image

Figure 2.1
CLIP Classification example



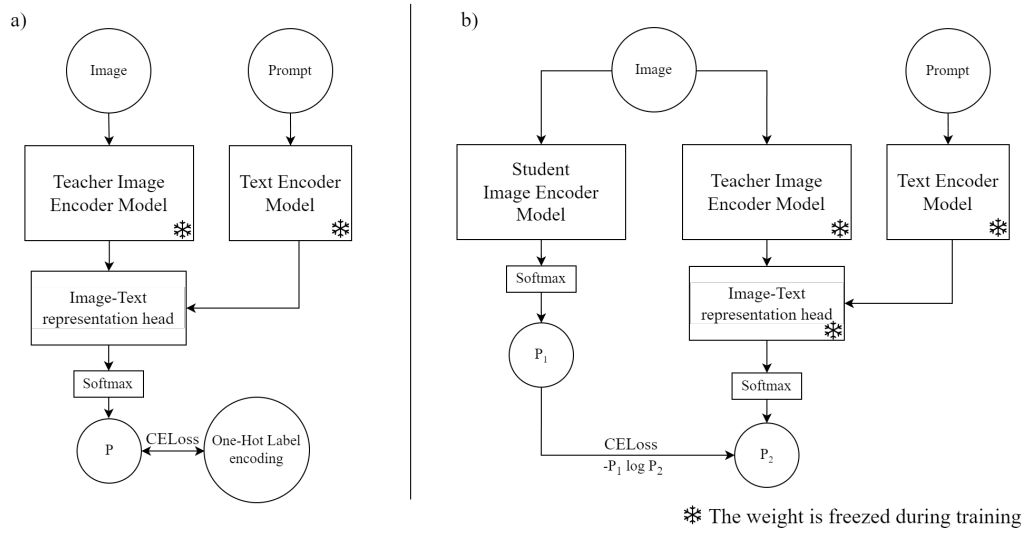
classification task, such as Mean Teacher (Tarvainen & Valpola, 2017), EMAN (Cai et al., 2021) and FixMatch (Sohn et al., 2020). DINO Caron et al. (2021) propose self-distillation pre-training without using any label method, which result in performance improvement. In this paper, we extended the self-distillation by creating representation which is image-text combined representation and we train student model to match teacher softmax outputs.

CHAPTER 3

METHODOLOGY

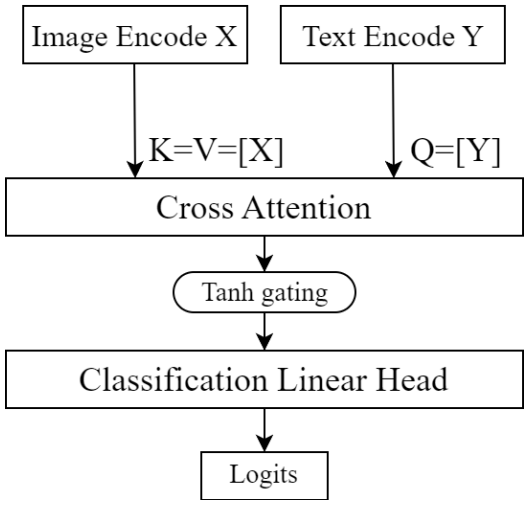
The proposed self-distillation method training process is shown in Figure 3.1. The first step in training process is to train the image-text representation head by freezing both image and text encoder model as shown in Figure 3.1 a). The second step is self-distillation with combined text and image representation output from the image-text representation head as shown in Figure 3.1 b). We experiment with multiple image and text encoder models. We compare our self-distillation method with the self-distillation from the teacher model directly without any text encoder model. The detail in each part of the expeirment is provided in this section.

Figure 3.1
Training methodology



a) Training image-text representation head using cross entropy loss b) Self-distillation training by freezing all teacher model

Figure 3.2
Image-Text Representation Head



3.1 Image-Text Representation Head Training

todo:change method image to use sub figure a and b In the first step as shown in Figure 3.1a), the image-text representation head is trained with image-text pairs (x_i, t_i), where x_i is an image input and t_i is a text created with a propmt "This is an image of [Class]". The teacher image encoder θ_i and the text encoder θ_t in the training are pre-trained and freezed. The image and text encoding are obtained by a mapping function $\hat{x}_i = f(x_i; \theta_i)$ and $\hat{t}_i = f(t_i; \theta_i)$ repectively. The image-text representation head as shown in 3.2 which based on an cross-attention and linear classification layer, is produce logits output as 3.1. Then, the logits output from the image-text representation head transform into probability distribution output with a softmax function.

$$test \tag{3.1}$$

The loss function for training the image-text representation head is a cross-entropy as 3.2, where y is a one-hot label.

$$\mathcal{L}_{Cls} = -y \log \hat{y} \tag{3.2}$$

3.2 Model Achitecture

3.2.1 Image-text representation head

By using two encoder model for vision and language individually, the image-text representation have be created to get a single image representation for every single images. In this experiment, the cross-attention (Alayrac et al., 2022) architecture with the linear classification head is used to create image-text representation as illustrated in Figure 3.2.

3.2.2 Teacher student

For the teacher model, we will use two stream encoder based model same as CLIP model (Dosovitskiy et al., 2021). In this experiment, the teacher vision encoder model will be ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2021) version. For the student model we used the same architecture as teacher vision encoder model, which are ResNet and ViT. todo: Add table describes both image and text encoders.

3.3 Training Objectives

In the first step, we trained the image-text representation head with benchmark datasets by using Cross Entropy loss as describe in **??** a). The image and text encoder was freezed during the first step training. For text input, we used "This is the image of [Class]" as a prompt (Radford et al., 2021). After the first image-text representation head were trained, we create a new student model which have the same architecture as a image encoder model with a linear classification head. The student model was randomly initialized parameters. The objective for self-distillation with teacher and student is

3.4 Evaluation

3.5 Ablation Study

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.

Allen-Zhu, Z., & Li, Y. (2023). Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The eleventh international conference on learning representations.* Retrieved from `https://openreview.net/forum?id=Uuf2q9TfXGA`

Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C., Tu, Z., & Soatto, S. (2021). Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 194–203).

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9650–9660).

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`

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., & Anandkumar, A. (2018). Born again neural networks. In *International conference on machine learning* (pp. 1607–1616).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G., Dean, J., & Vinyals, O. (2014, 03). Distilling the knowledge in a neural network. In (p. 1-9).

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).

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).

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, *33*, 596–608.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, *30*.

Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10687–10698).

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*. Retrieved from `https://openreview.net/forum?id=Ee277P3AYC`

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. In *Bmvc*.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 3713–3722).