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

		Page
ACKNOWLEDGMENTS		**ii**
ABSTRACT		**iii**
LIST OF TABLES		**x**
LIST OF FIGURES		**xi**
CHAPTER 1	**INTRODUCTION**	**1**
1.1	Background	1
CHAPTER 2	**LITERATURE REVIEW**	**4**
2.1	Vision-Language model	4
2.2	Knowledge Distillation and Self-Distillation	4
CHAPTER 3	**METHODOLOGY**	**6**
3.1	Image-Text Representation Head Training	6
3.2	Self-Distillation	7
3.3	Evaluation	8
3.4	Ablation Study	9
3.4.1	Few-shot learning	9
3.4.2	Using Image captioning as a prompt	9
3.4.3	Repeatation self-distillation	9
3.4.4	Image-Text Retrieval	9
CHAPTER 4	**Results**	**11**
CHAPTER 5	**DISCUSSION**	**12**
CHAPTER 6	**CONCLUSION**	**13**
REFERENCES		**14**

LIST OF TABLES

Tables	Page
Table 3.1 Experiment evalutation	9

LIST OF FIGURES

Figures	**Page**
Figure 1.1 Overall methodology	2
Figure 2.1 CLIP Classification example	5
Figure 3.1 Training methodology	7
Figure 3.2 Image-Text Cross Attention Classification head	8

CHAPTER 1

INTRODUCTION

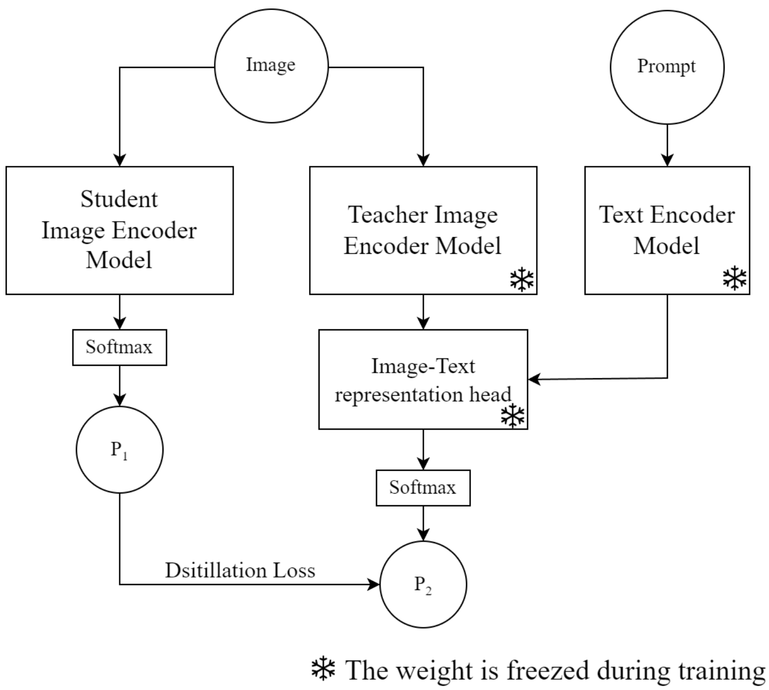
1.1 Background

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 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. Thus, it is a clear benefit of the training model in utilizing image and text information.

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 1.1. Multimodal teacher models were constructed by leveraging a pre-trained language model and a pre-trained image encoder. The output of both encoders was 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 was used as a query to extract the relevant information from the image encoding. The student model, which had the same architecture as the teacher image encoder model, was trained using teacher output as a target. Thus, the student learned 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 showed that the student model achieved 3% higher accuracy by providing detailed descriptions in the training process. This suggested that by using the text encodings with cross-attention, the model extracted higher semantic information and more precise image representations from the images.

To summarize our contribution. Firstly this paper investigated 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 proposed a method to efficiently combine textual information and images for the self-distillation method. Lastly, this work also investigated the effect of prompts in our methods to create image descriptions for training.

CHAPTER 2

LITERATURE REVIEW

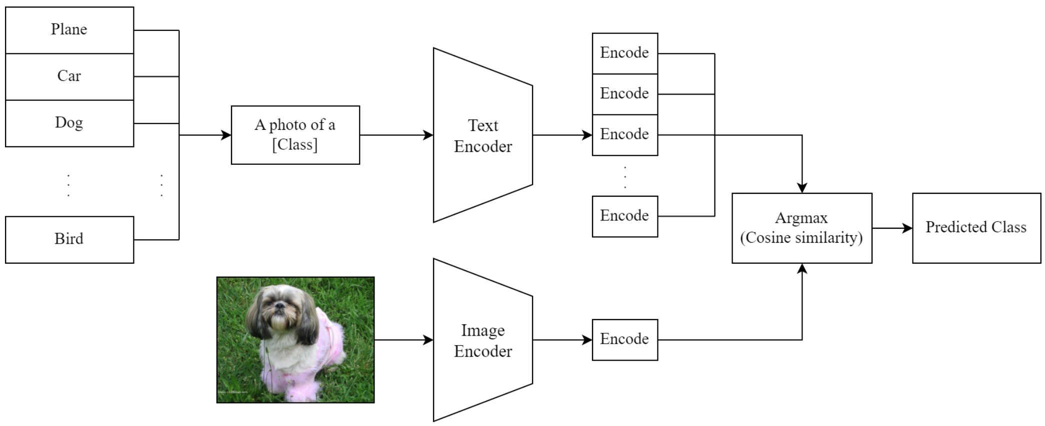
2.1 Vision-Language model

In the past few years, many works have shown the ability to utilize textual information with the 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 a large amount of the image-text pair dataset, the ALIGN model could make up for the noisy image description and surpass the model, which was trained with the benchmark dataset in the zero shot image classification task. Recently **Co**ntrastive **Ca**ptioner (CoCa) (Yu et al., 2022) proposed a vision-language encoder-decoder model which was trained with image-text contrastive loss and captioning loss Cross attention layers were added to join image-text modality. The CoCa model performed linear probing image classification on ImageNet with top-1% 90.6% accuracy. In this research, we adopted the two stream encoder method same as CLIP, and we also used a cross attention layer to create image-text representation for classification.

2.2 Knowledge Distillation and Self-Distillation

Knowledge Distillation was firstly proposed by Hinton et al. (2014) to compress the model size while maintaining the model performance as much as possible. The method contained a smaller student model and a single or multiple larger teacher model. The knowledge was transferred by optimizing the student model output to match the teacher's output. Furlanello et al. (2018) investigated knowledge distillation using a student model size the same as the teacher model, showing improvement in the student model. Such a method is called self-distillation.

Figure 2.1
CLIP Classification example



The self-distillation has widely adopted in semi-supervised image classification tasks, such as Mean Teacher (Tarvainen & Valpola, 2017), EMAN (Cai et al., 2021) and FixMatch (Sohn et al., 2020). DINO Caron et al. (2021) proposed self-distillation pre-training without using any label, which resulted in performance improvement. In this paper, we extended the self-distillation by creating representation which was image-text combined representation, and we trained the student model to match teacher softmax outputs.

CHAPTER 3

METHODOLOGY

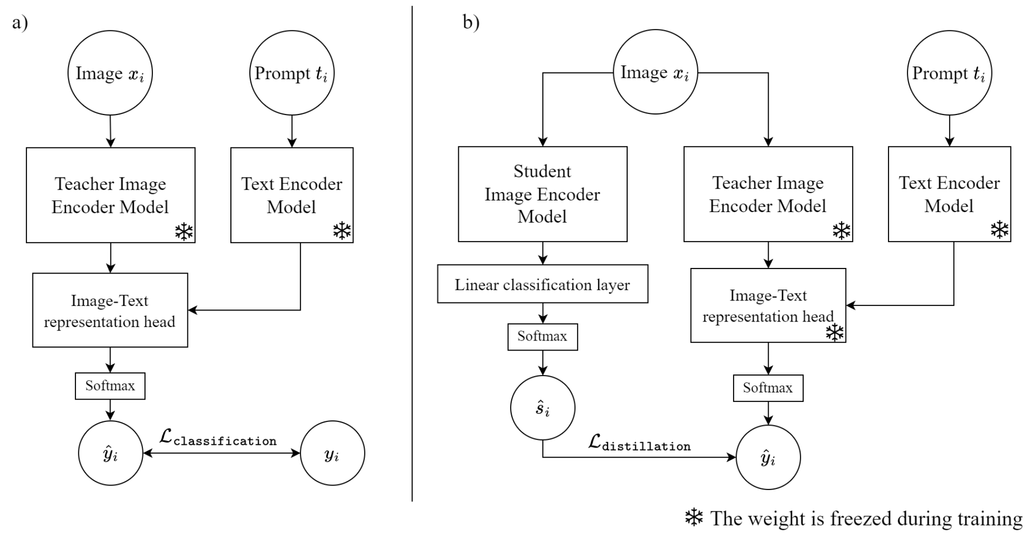
The proposed self-distillation method training process is described in Figure 3.1. The first step in the training process was to train the image-text representation head by freezing both the image and text encoder model as shown in Figure 3.1 a). The second step was self-distillation with combined text and image representation output from the image-text representation head as shown in Figure 3.1 b). Difference image and text encoder models pair were chosen to demonstrate the benefit of our method. We compared our approach with other self-distillation methods (Furlanello et al., 2018; Xie et al., 2020). The detail for each part of this experiment is provided in this section.

3.1 Image-Text Representation Head Training

In the first step as shown in Figure 3.1a), the image-text representation head was trained with image-text pairs (x_i, t_i), where x_i was an i^{th} image input and t^{th} was an i^{th} text created with a prompt "This is an image of [Class]". The teacher image encoder θ_{IE} and the text encoder θ_{TE} in the training were pre-trained and frozen. The image and text encoding were obtained by a mapping function $x_i' = f(x_i; \theta_{IE})$ and $t_i' = f(t_i; \theta_{TE})$ respectively. The image-text representation head as shown in Figure 3.2 which is based on a cross-attention and linear classification layer, produced logits output as Eq.3.1. Then, the logits output from the image-text representation head transformed into probability distribution output with a softmax function.

$$\hat{y}_i = \texttt{Softmax}(\texttt{Attention}(K = x_i', Q = t_i', V = x_i')) \tag{3.1}$$

Figure 3.1
Training methodology



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

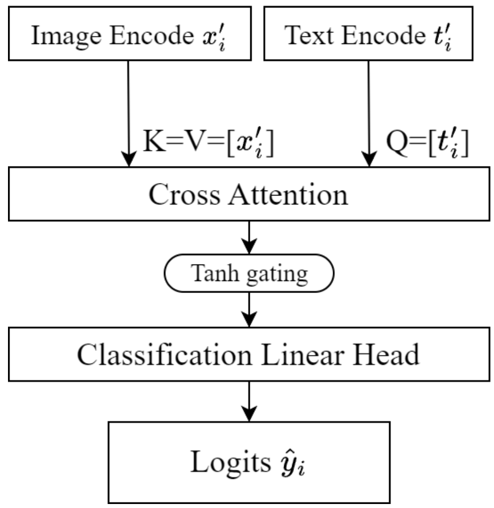
The loss $\mathcal{L}_{\texttt{classification}}$ for training the image-text representation head was a cross-entropy as Eq.3.2, where $y_i \in \{0,1\}^C$ is a one-hot encoded label, C is the number of classes and N is the number of training samples.

$$\mathcal{L}_{\texttt{classification}} = -\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{3.2}$$

3.2 Self-Distillation

After the image-text representation head was trained, the image-text representation head was frozen during the self-distillation process. For the student model, we created a new image encoder model with the same architecture as the teacher image encoder model, but with different initialized parameters. A linear classifi-

Figure 3.2
Image-Text Cross Attention Classification head



cation and softmax layer was added on top of the student image encoder model to produce output distribution $\hat{s}_i$ for the self-distillation process. The target for training self-distillation was the softmax output $\hat{y}_i$ from the image-text representation head with cross-entropy loss as a loss function. The loss $\mathcal{L}_{\texttt{distillation}}$ for self-distillation was the cross-entropy loss as shown in Eq 3.3.

$$\mathcal{L}_{\texttt{distillation}} = -\sum_{i=1}^{N} \hat{y}_i \log \hat{s}_i \tag{3.3}$$

3.3 Evaluation

In this work, we evaluated the student model with accuracy using an image classification task. The benchmarks for evaluation were ImageNet, CIFAR-10 and CIFAR-100. The student model was evaluated compared to the teacher image encoder model using linear probing and student model trained with self-disillation using a single image encoder as a teacher model as shown in Table 3.1.

Table 3.1
Experiment evalutation

Teacher Image Encoder	Image Encoder Parameters	Text Encoder	Self-Distillation without Text				Self-Distillation with Text			
			CIFAR10	CIFAR100	ImageNet Top1%	ImageNet Top5%	CIFAR10	CIFAR100	ImageNet Top1%	ImageNet Top5%
ViT-B/32	86M	RoBERTa								
ViT-B/32	86M	CLIP								
ViT-B/16	86M	RoBERTa								
ViT-B/16	86M	CLIP								
ResNet-50	102M	RoBERTa								
ResNet-50	102M	CLIP								

3.4 Ablation Study

3.4.1 Few-shot learning

As this method provided texts for training student image encoder models, the texts provided additional information for better image representations. Consequently, the student model benefited from our method in few-shot learning situations. In this part, we provided benchmark results for few-shot learning situations.

3.4.2 Using Image captioning as a prompt

For a better understanding of the effect of text prompts in our self-distillation method, we experimented by providing better descriptive prompts. The image captioning model was used to create image descriptions for the self-distillation process. Multiple image captioning models were tested for a prompt generation.

3.4.3 Repeatation self-distillation

By using the student as a teacher model for training another student model repeatedly, the performance increased gradually over each generation of the student model (Furlanello et al., 2018; Xie et al., 2020). In this work, we also investigated the performance increase over each generation of the student model using our self-distillation method.

3.4.4 Image-Text Retrieval

By increasing performance in the student model using our method with textual information, we suggest that the student would be a good image encoder which also has information about text. Such that, we can use our method to improve image-text retrieval tasks. In this ablation study, we provided a result of using a student as an image encoder for the image-text retrieval benchmark, which could test the image encoder's higher semantic understanding and multimodal capability.

CHAPTER 4

Results

CHAPTER 5

DISCUSSION

CHAPTER 6

CONCLUSION

REFERENCES

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).