Contrastive Learning for Sentence Embeddings in BERT and its Smaller Variants

Stanford CS224N Custom Project

Vrishab Krishna
Department of Computer Science
Stanford University
vrishab@stanford.edu

Rohan Bansal
Department of Computer Science
Stanford University
robansal@stanford.edu

Abstract

Contrastive learning is a method of learning representations using invariances in the data under augmentations and encouraging the resultant embeddings of augmented samples to remain close together. An interesting property of such approaches is that they enable models to perform better on different tasks even when trained on smaller amounts of data and also enables smaller models to perform as well as their larger counterparts. In this project, we demonstrate that both supervised and unsupervised contrastive learning approaches provide improved semantic performance for smaller BERT architectures (including $\text{BERT}_{\text{small}}$, and $\text{BERT}_{\text{mini}}$) both in pre-training and downstreaming objectives, while improving the representational uniformity of the word embeddings and retaining widespread downstream flexibility. Our results indicate that we can continue to maximize performance in smaller transformer architectures and produce comparable results to larger state-of-the-art architectures at a fraction of the computing cost and training time. We conclude by offering new areas of research that may provide even larger boosts to semantic performance, including supervised applications in computer vision that have shown to perform well for comparable objectives.

1 Key Information to include

External collaborators (if you have any): **None**, External mentor (if you have any): **None**, Sharing project: **False**

2 Introduction

Self-supervised learning (SSL) has resulted in tremendous improvements in NLP models and representations of data without the need for intensive and noisy labeling on ill-defined tasks. Next Sentence Prediction (NSP), Masked Language Modeling (MLM) have been used to inculcate language specific priors into models removing the need for extremely large amounts of labelled data for a particular task as well as generating more general models that can be used as a precursor to a fine-tuned model for classification, regression, as well as generative tasks.

A promising sub-field of of self-supervised learning is contrastive learning, where the goal of the optimization is to distinguish between similar and dissimilar samples in the data. This involves capitalizing on the fundamental stuctures within the data to develop compressed, expressive, and robust representations. One way this is done is using priors on invariances within the data and using them to label pairs of datapoints as similar (positive) or dissimilair (negative). In computer vision, such methods have been found to be very successful as image transformations like affine disturbances, color jitters and noise are easy to apply and do not have a significant impact on the semantics of an image when used in moderation. The positive samples would be two augmented versions of the same starting image and negative samples would be augmented versions of different images. SSL methods

like SimCLR Chen et al. (2020) and DINO Caron et al. (2021) have provided significant boosts when trained on fractions of label data.

In NLP, we look at similar approaches but in the context of sentences and their representations. The issue is that augmentations and invariances are more complex in natural language than with images - the structure makes it difficult to generate alternate views of the same sentence which do not perturb the semantics. However, the addition of contrastive approaches have resulted in significant improvements in few-shot and fine-tuning accuracies as well as generalization. Hence, even smaller models contrastively fine tuned on such methods could potentially result in similar performance as much large models trained in the vanilla fashion. Distillation methods between larger and smaller models have shown that equivalent performance can be reached with an order of magnitude or fewer parameters. SSL methods could provide ways for smaller models to be trained and achieve better accuracy without the need for training larger models at all.

3 Related Work

In the context of sentence embeddings, recent works have shown that different views of the same sentence can be generated with different dropouts in the model (Yao et al., 2021; Yan et al., 2021). Using the InfoNCE contrastive loss function (very similar to that used by SimCLR in images (Chen et al., 2020)), these methods optimize generated embeddings across a corpus, obtaining state of the art results with BERT-base and BERT-large on datasets of semantic and textual similarity (STS) (Rethmeier and Augenstein, 2021; Gao et al., 2021) (with a positive pair (x, a^+) and K negative pairs (x, a_i^-)):

$$\mathcal{L}_{\text{InfoNCE}} = \frac{e^{s(x,a^+)}}{e^{s(x,a^+)} + \sum_{i=1}^{K} e^{s(x,a_i^-)}} \tag{1}$$

One issue with such contrastive schemes is feature suppression. In the case of the SimCSE paper above, the embeddings sometimes fail to discern textural and semantic components. Early last year, Wang et al. (2022) made an addition of soft negative samples to force a difference between textual and semantic similarity. They then suggest to constrain the cosine similarity difference between positive pairs and soft negative pairs Δ by proposing a bidirectional margin loss to constrain this value in the interval $[-\beta, -\alpha]$:

$$\mathcal{L}_{\text{BML}} = \text{ReLU}(\Delta + \alpha) + \text{ReLU}(-\Delta - \beta) \tag{2}$$

Thus, the final objective function for soft negative contrastive training is defined as:

$$\mathcal{L}_{\text{SNCSE}} = \mathcal{L}_{\text{InfoNCE}} + \lambda \mathcal{L}_{\text{BML}} \tag{3}$$

where λ is used to control the weight of these soft negative samples. This approach was shown to alleviate some of the issues with feature suppression for larger BERT models, and the authors also experimented with treating soft negative examples as purely negative (in an identical objective to $\mathcal{L}_{\text{InfoNCE}}$), however found no marked improvement (Wang et al., 2022).

These contrastive methods have provided significant performance increases across different datasets and tasks when tested on large models like $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$. However, another set of experiments of interest that, to our knowledge, has not been previously tested include observing how the performance of smaller models like $\text{BERT}_{\text{small}}$ and $\text{BERT}_{\text{mini}}$ hold up under SimCSE and SNCSE pretraining. We hope to see that smaller models can reach the accuracy of $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ on downstream tasks with only a fraction of the training data. This is incredibly useful as it opens up the use of highly accurate models in resource constrained settings.

4 Approach

Our goal is to study the improvement provided by the previously described pretraining methods on top of (after) regular NSP that is used to train the vanilla BERT variants. We also contextualize these improvements in terms of the training time taken (number of epochs of pretraining to achieve a result)

as well as amount of data used for training in downstream tasks. We run a comprehensive set of experiments, including on different architectures sizes, datasets, and augmentation methods, to best assess the maximal performance that these techniques can provide.

We are using the SimCSE and SNCSE repositories for our experiments, adapting and significantly simplifying their original code with the huggingface transformers module to make it more modular for the experiments we are hoping to run in the future. We also include 1 which visualizes the supervised and unsupervised approaches proposed by Gao et al. (2021). This figure highlights that the supervised task is essentially a binary classification task where embeddings are drawn closer together for entailment pairs and separated for all other pairs. In contrast, the unsupervised task uses two different dropout masks to attempt to predict whether a sentence is identical to itself and does not require annotated data. For the soft negative unsupervised task, we follow the best performing variable selection in the original work Wang et al. (2022) by setting $\alpha = 0.1, \beta = 2.0$ and $\lambda = 1 \times 10^{-3}$.

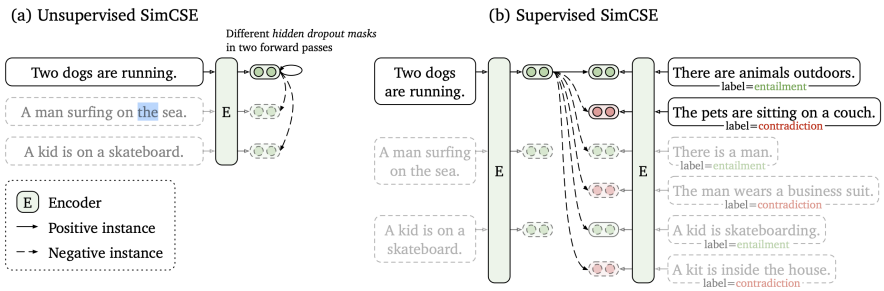

Figure 1: A depiction of the supervised and unsupervised SimCSE approaches for pre-training Gao et al. (2021)

As baselines we use each of the vanilla BERT models - in particular, we consider the vanilla, pretrained and available $\text{BERT}_{\text{large}}$, $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{small}}$, and $\text{BERT}_{\text{mini}}$ models, each of which is trained on both NSP and MLM tasks on 2.5 billion words from Wikipedia (Devlin et al., 2018). We also write custom code to further pre-train BERT on the random sample of Wikipedia sentences we use for contrastive learning in order to provide a more quality baseline and eliminate potential differences in precision. The checkpointed models are provided from the HuggingFace transformers repository and run on sentences from each dataset to generate the embeddings which are subsequently evaluated.

5 Experiments

5.1 Data

Our data use cases, we have a broad classification into those used for the contrastive pretraining and those used in downstream tasks as a pseudometric on the quality of embeddings.

5.1.1 Contrastive Pretraining

Supervised, a collection of $570k$ annotated sentence pairs from the SNLI dataset which are labeled either entailment (similar) or contradiction (different).

General Unsupervised, a collection of 1 million randomly sampled sentences from Wikipedia where each sentence is considered similar to itself and separate from all other sentences.

Soft Unsupervised, again generated from a collection of 1 million randomly sampled sentences from Wikipedia, however soft samples are drawn by using explicit parser negation for relevant sentences and hard samples are simply the remaining sentences.

5.1.2 Downstream

Sentiment Classification (SST-2), a collection of $65k$ movie review sentences, labeled as positive (similar) or negative (different).

Question Answering (QNLI), a collection of $110k$ question-answer pairs randomly sampled from Wikipedia (and answered by human experts).

Semantic Similarity (STSB), semantic analysis on the STS benchmark consisting of $8.5k$ annotated sentence pairs with human annotated similarity scores.

5.2 Evaluation method

For the pre-training evaluation, make use of the similar methods of as SimCSE and utilize the SentEval package with different poolers. To check the quality of embeddings, we evaluate them zero-shot - there is no training on the actual dataset itself. The matching is done using the cosine similarity of embeddings to predict across multiclass and binary tasks. This fundamentally evaluates whether the embeddings generated by the model have the predictive capacity to solve this problem independently, with no training. These outputs are then evaluated with Pearson's and Spearmans correlation to check the degree of similarity between predictions and the ground truth. We report these scores individually for datasets STS12-15 and SICK and well as averaged across all datasets. Because this evaluation is done zero-shot and the models do not see these semantic datasets during pre-training, we utilize the comparable vanilla BERT checkpoints as baselines. Since our goal is to measure whether this pretraining method provides additional benefits on top of NSP pretraining, the embeddings from these checkpoints are a fair baseline.

We have 3 different downstream tasks that we use to evaluate the quality of embeddings. While SST-2 and QNLI are binary classification tasks (where we use accuracy as our metric), STSB is a discrete regression task and we use the average Pearson's and Spearmans correlation score for evaluation.

5.3 Experimental details

The models ($\text{BERT}_{\text{large}}$, $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{small}}$, , and $\text{BERT}_{\text{mini}}$) were first pretrained starting with a BERT uncased checkpoint for each variant on one of the three datasets described above using the `transformers` library (Wolf et al., 2020). We use a batch size of 64 with a learning rate of 3×10^{-5} in the Adam optimizer for 1 and 3 epochs for the general unsupervised and supervised tasks respectively as suggested by Gao et al. (2021). These experiments use the traditional infoNCE loss described above and we also incorporate further MLM to make the task more difficult, and our final objective function is simply the sum of these two tasks. For the soft unsupervised task, we use the bidirectional margin loss added with the MLM task as well, and perform a grid search over learning rates of $\{2 \times 10^{-4}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$ while training for 3 epochs as suggested by Oord et al. (2018). After the conclusion of this pre-training, we performed our sentence embedding based STS evaluation, with different poolers (including the CLS token, averaging the embeddings, and random sampling), to quantify the improvements of this contrastive learning on an unseen dataset and compared to the original baselines. Finally, we trained each model on the downstream tasks presented above by performing a comprehensive gridsearch over learning rates of $\{2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$, batch sizes of $\{16, 32, 64\}$ and epochs $\{2, 3, 4\}$ and chose the best performing combination for each model for comparison. All experiments are performed on 4 A100 GPUs with distributed training.

5.4 Results

Model	STS12	STS13	STS14	STS15	SICK	Avg.
BERT-Mini	30.52	32.97	27.11	7.40	51.93	26.05
BERT-Small	33.80	35.39	30.93	11.55	54.68	29.28
BERT-Base	17.19	29.06	19.55	7.16	35.11	17.60
BERT-Large	18.82	28.53	23.69	9.29	35.09	17.92
BERT-Small (Sup.)	73.81	76.42	74.60	14.19	79.62	56.80
BERT-Small (Unsup.)	60.38	75.08	66.56	14.67	68.76	50.42
BERT-Small (Soft)	53.76	63.12	56.11	14.80	63.15	44.56
BERT-Mini (Sup.)	69.51	70.28	69.26	13.08	76.27	53.91
BERT-Mini (Unsup.)	55.13	63.76	56.23	13.95	60.45	44.68
BERT-Mini (Soft)	56.15	63.31	55.80	13.77	62.15	44.82

Table 1: Aggregated Results after Pre-Training for custom SentEval Evaluation

Model	SST-2	QNLI	STSB
BERT-Base	93.5	90.5	85.8
BERT-Large	94.9	92.7	86.5
BERT-Mini	85.9	84.1	75.4
BERT-Small	89.7	86.4	78.8
BERT-Small (Sup.)	87.5	85.0	84.6
BERT-Small (Unsup.)	86.9	85.2	84.5
BERT-Mini (Sup.)	82.9	82.88	78.3
BERT-Mini (Unsup.)	82.0	82.7	77.5

Table 2: Downstream GLUE Experiments

We observe that this brief period of pre-training (< 5 minutes on 2 A100 GPU) allows the small and tiny BERT models to improve drastically on the majority of the STS datasets. For sentence embedding based evaluation, we tried different pooling outputs–including taking the CLS token, averaging the embeddings and randomly sampling from each sentence–however, we saw little discrepancy in results, and thus we provide relevant numbers for these smaller variants using the traditional CLS pooling in Table 1. These results indicate that even unsupervised pre-training approaches on these smaller architectures produce better results than the vanilla $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ embeddings while being $\frac{1}{4}$ and $\frac{1}{10}$ the size of bert-base. It is also important to note that conducting further standard pre-training on these vanilla models did not improve their semantic performance. In fact, the numbers highlight that the released checkpoints for the larger BERT models exhbit particularly poor semantic performance. Additionally, we see that the models trained with soft negative samples generally lag behind and do not seem to outperform the traditional supervised and unsupervised InfoNCE objectives.

We then provide the downstream results in Table 2 where we observe that the contrastive pre-trained models perform slightly worse than their traditional counterparts on the sentiment classification (SST-2) and question answering (QNLI) tasks, but actually outperform on the pearson score for semantic similarity after being trained on the STSB dataset. The $\text{BERT}_{\text{small}}$ model actually provides nearly identical performance to both $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ on the STSB task which indicates the ability of contrastive learning to provide a massive improvement despite the significant difference in model size. Additionally, all smaller variant runs take under 12 minutes to complete downstream training on 2 A100 GPUs.

6 Analysis

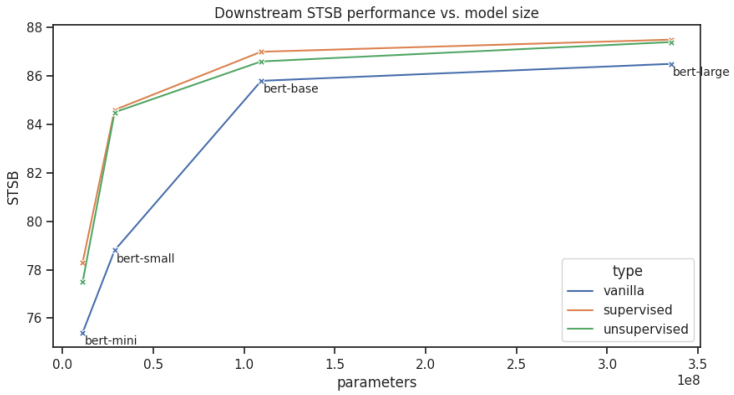

Figure 2: Downstream STSB scores for supervised, unsupervised, and standard BERT models

The pretraining results produced above indicate that the contrastive learning approaches are highly effective in improving semantic representations, and the performance of different self-supervised and contrastive methods are helpful in generating smaller models with representations on par with larger models on some tasks.

The drastic increase in performance compared to the BERT baselines also emphasize the poor nature of these released embeddings and their tendency to predict high similarity for the vast majority of sentences. We provide 3 examples of sentence pairs which highlight the types of improvements that the supervised, contrastive $\text{BERT}_{\text{small}}$ makes when compared to the vanilla $\text{BERT}_{\text{base}}$:

- We observe an improvement in score from $.967 \rightarrow .863$ for the pair **I have this book to read** and **I have to read this book**. This example highlights how changes in word order can often lead to different sentence meanings, but the traditional BERT models struggle to contextualize this difference.
- We observe an improvement in score from $.802 \rightarrow .853$ for the pair **There's a man on a bicycle** and **That man is riding a bike**. Here, the sentences are nearly identical in meaning but use different phrasing which leads to a lower score in $\text{BERT}_{\text{base}}$.
- We observe a large improvement in score from $.932 \rightarrow .167$ for the pair **Have you seen my cat** vs **I play the piano**. These sentences are completely unrelated outside of the first-person subject, however $\text{BERT}_{\text{base}}$ assigns an unreasonably high similarity score.

A known issue with BERT embeddings is that and indicate that they are highly clustered and lack uniformity - this can be seen in the poor performance of BERT in the STS tasks. The above examples help illustrate how contrastive learning leads to better uniformity and reduces both the outsized influence of wording and the anisotropic nature of the vanilla BERT embeddings. These results also confirm the findings of Gao et al. (2021) and Wang et al. (2022) indicating that contrastive learning produces higher quality similar sentence retrieval for given query sentences.

The lack of impact on different pooling choices for the sentence evaluation tasks also lends credence to the idea of improved distributions and indicates that the model is learning intrinsically better representations for frequent words in the training corpus. Additionally, the performance of the unsupervised models (which were trained on Wikipedia corpus) and the lack of improvement demonstrated by the vanilla models when trained for more steps through the standard pre-training approach serves as a partial ablation result and seems to indicate that the contrastive learning objective alone is responsible for the lift, as opposed to the data or further training. Separately, the lagging performance of soft negative approaches may imply that the smaller models are not flexible enough to capture the more nuanced differences through a purely unsupervised approach, such as the addition of a single negation term, because they have been shown to outperform InfoNCE loss for both $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ in Wang et al. (2022).

The downstream performance of the models also tends to align with prior literature and our expectations. We observe incrementally worse performance on both the sentiment classification and question answering tasks which indicate that the traditional BERT is able to better fit these specific datasets but is not necessarily capturing the true semantics of the English language. However, the minor discrepancy seems to actually indicate that the pre-trained contrastive embeddings are still sufficiently flexible to be used in general tasks and would be appropriate as an alternative to the vanilla checkpoints released in the original BERT paper. On the contrary, for the downstream STSB task we see that the contrastive pre-trained approaches produce better performance than their standard counterparts and we provide a visualization of these changes in Fig 2. Generally, the supervised approaches perform slightly better than the unsupervised approach, which is expected and in line with the pre-training STS results, however this difference seems to be nearly negligible. Additionally, we see that the pearson curves seem to flatten significantly beginning with $\text{BERT}_{\text{small}}$ and the improvements for scaling the parameter count are relatively minimal. This indicates that the contrastive approaches do have a limit on the added value they can provide, and we hypothesize that access to more high-quality annotated data may allow for even larger improvements in $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$.

7 Conclusion

In this work, we demonstrate that contrastive learning approaches in pre-training, as presented by Gao et al. (2021) and Wang et al. (2022), provide significant semantic improvements to high-quality BERT embeddings, particularly in its smaller variants such as $\text{BERT}_{\text{small}}$ and $\text{BERT}_{\text{mini}}$. Through ablation experiments and downstream training, we also highlight that these semantic embeddings are a viable alternative to the current state-of-the-art approaches and can provide the necessary flexibility needed to adapt them to a wide variety of fine tuning objectives. Additionally, the smaller size of

these models allows for quicker and cheaper training and reduces the hardware overhead associated for utilization in general use cases. The improvements are demonstrated even with unsupervised objectives using easily accessible datasets, such as the Wikipedia corpus, further eliminating the need for more human annotated semantically accurate data.

An avenue we would like to further explore is the use of different augmentations and methods to produce the two different views of the data in contrastive learning. SimCSE makes use of different dropout masks in two forward passes (Gao et al., 2021) but there exist a host of other transforms that could be applied in tandem for training, like reordering, corruption, etc. (see Bhattacharjee et al. (2022) for a more comprehensive list). For future work, we would intend to combine several of these transforms such as dropout, latent/embedding perturbation, word deletion and reordering, and test which combinations result in marked metric improvements. We would be particularly interested in analyzing whether these types of augmentations can boost the performance of soft contrastive sampling, an approach which provided significant boosts to larger BERT architectures but had disappointing performance for the smaller variants. We would also wish to determine whether these types of corruption methods could allow for linear probing to provide high-quality results, where the semantic word embeddings could remain fixed and only a classification layer head would need to be trained for various finetuning tasks. This would further reduce the training cost and allow for retention of semantic details even in other domains.

Finally, another aspect of future work we would like to consider is the use of more modern constrastive self-supervised approaches from computer vision. DINO is a method that involves self distillation with a teacher and student network where the student is trained to mimic the activations of the teacher under augmented samples but the teacher's weights are updated as an exponential moving average of the student's weights Caron et al. (2021). This method showed emergent properties in vision transformers as well as obtaining state-of-the-art results across several tasks–including image retrieval, copy detection, and semantic layout–which indicate that they may further improve semantic performance in textual domains. To our knowledge, such method of self-distillation for pretraining has not been applied to BERT and has not been tested in smaller variants either. It would be interesting to see how methods that were highly effective in computer vision fare in NLP and what are the potential reasons for their success or failure.

References

Amrita Bhattacharjee, Mansooreh Karami, and Huan Liu. 2022. Text transformations in contrastive self-supervised learning: A review.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding.

Nils Rethmeier and Isabelle Augenstein. 2021. A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples. *arXiv preprint arXiv:2201.05979*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger,

Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models.