Learning Downstream Task by Selectively Capturing Complementary Knowledge from Multiple Self-supervisedly Learning Pretexts

Jiayu Yao †‡, Qingyuan Wu †, Quan Feng †, Songcan Chen §
† Nanjing University of Aeronautics and Astronautics, Email: jiayu_yao@nuaa.edu.cn
‡ Nanjing University of Aeronautics and Astronautics, Email: wuqingyuan@nuaa.edu.cn
† Nanjing University of Aeronautics and Astronautics, Email: fenquan@nuaa.edu.cn
§ Nanjing University of Aeronautics and Astronautics, Email: s.chen@nuaa.edu.cn

Abstract—Self-supervised learning (SSL), as a newly emerging unsupervised representation learning paradigm, generally follows a two-stage learning pipeline: 1) learning invariant and discriminative representations with auto-annotation pretext(s), then 2) transferring the representations to assist downstream task(s). Such two stages are usually implemented separately, making the learned representation learned agnostic to the downstream tasks. Currently, most works are devoted to exploring the first stage. Whereas, it is less studied on how to learn downstream tasks with limited labeled data using the already learned representations. Especially, it is crucial and challenging to selectively utilize the complementary representations from diverse pretexts for a downstream task. In this paper, we technically propose a novel solution by leveraging the attention mechanism to adaptively squeeze suitable representations for the tasks. Meanwhile, resorting to information theory, we theoretically prove that our scheme significantly exceeds current popular pretext-matching based methods in gathering knowledge and relieving negative transfer in downstream tasks.

Index Terms—Self-supervised Learning, Contrast Learning, Discriminant Representations, Attention Mechanism, Complementary Knowledge.

I. INTRODUCTION

The success of supervised learning as a learning paradigm benefits from large-scale labeled data [1]. In recent years, the progress of supervised learning has gradually been limited by the availability of large amounts of well-labeled data. In the real world, unlabeled data is easier to obtain [2]. These data generally contain generalizable knowledge for learning new tasks. Nevertheless, learning discriminative representations directly from unlabeled data is a very challenging problem. Fortunately, SSL can overcome this problem through automatic annotation of the data’s own attributes. As illustrated in Figure 1, SSL usually follows a two stage pipeline: 1) learning the representation with auto-annotated pretext; 2) transferring the representation to the downstream tasks.

Most current popular works only focus on the first-stage, are dedicated to learning good representation [3], [4]. E.g., [3] the author learns invariant visual representations by contrasting negative pairs stored in a dynamic dictionary. [6] the author proposes to maximize the consistency between different augmented views of the same data sample in the latent space to obtain a more favorable visual representation. Excitingly, the performance of these methods compared with the performance of the supervised learning paradigm has gradually shown comparable results and even better in some specific scenarios.

By contrast, the understudied second-stage pipeline still needs further exploration. From the perspective of knowledge transfer, general transfer learning transfers knowledge from supervised pre-trained tasks to semantically similar target/downstream tasks. SSL typically transfers the representations learned under unsupervised pretext(s) to labeled downstream tasks. The semantics of the self-labeled and/or contrastive information used as a pretext in the SSL scenario is different from the supervision label information of downstream tasks. In addition, the semantics are agnostic of the pretexts to downstream tasks. Such a dilemma may lead to suboptimal or even unfavorable generalization performance for downstream tasks. To the best of our knowledge, a tiny minority of studies tried to focus on the problems derived from this semantic...
difference. E.g., [7] the author defines multiple model pools for existing models to determine which models can be transferred to the current task for fine-tuning. [3] the author proposes to adaptively aggregate the pre-trained parameters of channel alignment that can be learned by multiple source models to promote the transfer of multi-source knowledge to downstream tasks. The similar strategy of these methods is to pick the most similar semantic model from the pre-learned pretexts as the source of knowledge transfer for downstream tasks. However, this type of strategy ignores such problems. For downstream tasks, there may exist more complementary knowledge among different transfer sources. In practice, more likely, the knowledge which the downstream task needs distributes or scatters in (even semantically) different pretexts. Therefore, we need to select some pre-learned models for transfer purposes.

To this end, this paper focuses on how to squeeze out the favorable representations distributed in diverse learned pretexts. Specifically, we firstly theoretically analyze using the information theory that diverse pretexts can provide more favorable complementary knowledge to the downstream tasks, i.e., squeezing out helpful representations from the constructed representation bank will lead to much more mutual information associated with the downstream tasks to be learned. Next, to guide purposeful squeeze, we propose an attention-based method that extracts diverse knowledge from the constructed representation bank and selectively focuses on the pivotal knowledge beneficial for downstream tasks. Finally, extensive experiment results demonstrate that our method is significantly more competitive compared with state-of-the-art (SOTA).

We summarize our main contributions as follows:

- It’s the first time investigating relationships among diverse representations from multiple pre-trained SSL models.
- A beneficial conclusion that purposefully and strategically squeezing out diverse representational knowledge from the constructed representation bank is conductive for downstream tasks learning.
- A novel attention-based framework that adaptively extracts the most representations from the constructed representations bank without matching the pretexts with downstream tasks.

The rest of this paper is organized as follows: Section 2 reviews some related work; Section 3 introduces our attention-based contrast learning method; Section 4 reports the experimental results, followed by conclusion in Section 5.

II. RELATED WORKS

A. Self-supervised learning (SSL)

Self-supervised learning is a rapidly developing new unsupervised representation learning paradigm, its principal merit is learning representation/knowledge without manual annotations. This paradigm generates supervision signals from intrinsic attributes or regularity of data [3], [9]. Currently, the springing up contrast-based family (like MoCo [5], SimCLR [6], BYOL [10]) even outperforms supervised counterparts in some particular datasets or downstream tasks.

However, the conclusion in [11] shows that there does not exist a best self-supervised representation over all tasks and just those downstream tasks similar to pretext will be generalized well, vice versa. In practice, pretext semantics tends to be agnostic to downstream tasks. Thus seeking a most similar pretext to downstream task [7], [12] or aggregating representations of diverse pretexts [13], [14] becomes a key and has given rise to some methods. The former either estimates the performance on downstream by Log Expected Empirical Prediction or fine-tunes the pre-trained models on downstream task, then transfers the best model to downstream tasks. Doing so tends to overlook the complementarity from other pretexts, especially when the best one just possesses partial knowledge for the target task. While fine-tuning on the model architecture also limits flexibility of the downstream models [15]. And even worse, the details, such as the architecture or weights, of the pre-learned models are all unavailable due to the privacy policy. The latter simply aggregate representations as downstream task-specific representation from all pretexts without selection, which may invoke too much irrelevant information, leading to poor performance of downstream task [16]. Similar to the aggregation methods, but a major difference from them is that ours strategically squeezes out must representations for downstream tasks.
downstream task from the knowledge bank by the attention mechanism without the need of both pre-learned models’ details and searching for single best pretext.

B. Attention mechanism

The attention mechanism is one of the important concepts in the current deep learning field. The attention mechanism is inspired by human biological systems that are more inclined to focus on specific areas of the object when processing information. Therefore, the attention mechanism aims to imitate human visual attention to capture pivotal object-related information. According to the realization of the attention mechanism, it can be briefly divided into the following categories:

1) **Hard attention**, which uses one-hot coding to obtain information about a specific area in the hidden sampling process of the coding layer. E.g., [17] the author proposes a top-down method of traversing the image scale space, which only accesses the multi-scale hard attention network of the image area with the most information along the way for image classification. [18] the author designs an end-to-end hard attention network with encoders and decoders for retinal vessel segmentation. [19] the author proposes a joint hard attention model that couples the input state to the output state separately for natural language processing. However, this type of method requires Monte Carlo sampling or similar methods to estimate the gradient.

2) **Soft attention**, which uses the [0, 1] weight probability distribution to focus on the information of a specific area in the coding process of hidden layer sampling. E.g., [20] the author proposes an attention model combining bidirectional long-term and short-term memory, soft maximum attention layer and convolutional neural network for real-time sarcasm recognition. [21] the author designs a deep network that uses trainable weights as soft multiplication masks and trains adaptive weights and network weights at the same time. This type of method solves the problem of hard attention not being differentiable, and makes it utilize the forward propagation and backward feedback of the neural network to learn the attention weight.

3) **Self-attention**, an attention variant, focuses more on internal correlations in data features. E.g., [22] the author designs a VisText self-attention module that combines relevant multi-modal features and clinical features to predict the risk of esophageal fistula. [23] the author designs a separable self-attention module that sequentially models spatial and temporal correlations for video representation learning. This type of method reduces dependence on external information, thereby improving the reasoning time of the model.

4) **Multi-headed attention**, which performs parallel computing on multiple attentions. E.g., [24] the author proposes an unsupervised multi-head attention network for cross-media hash retrieval of pictures and texts of semantic information. [25] the author proposes a cascaded multi-head attention network composed of multi-head local self-attention and feature relations for action recognition. The advantage of this type of method is that it can pay attention to the information of multiple subspaces at the same time.

5) **Local attention**, which only focuses attention to the context information of the small window (the small area), but does not focus attention on the state of all coding hidden layers. E.g., [26] the author proposes a model that uses supervised dynamic local attention to explicitly capture structural information for continuous sentence classification. [27] the author proposes to compare the local attention in the long short-term memory (LSTM) model to generate abstract text summarization. The characteristic of this kind of method is that more cost can be used to obtain miniaturized information on some key objectives.

6) **Global attention**, which takes into account both the object decoding hidden state and all encoded hidden states in the calculation process. E.g., [28] the author proposes a global attention network composed of a point-independent global attention module and a point-dependent global attention module for point cloud semantic segmentation. [29] the author proposes an order sensitive global attention network for text and image feature enhancement. This type of method is computationally costly and easy to distract when dealing with long text sequences.

Since this paper focus is how to use the existent attention mechanism to adaptively squeeze out the complementary knowledge from the bank, rather than itself. Therefore, just reviewing a few closely-related works as above.

III. Method

In practice, the semantics of the pre-learned pretext features are unnecessarily sufficiently relevant with the supervised information of the downstream task, leading to a semantic gap between pretext(s) and specific downstream task to great extent so that knowledge required by learning downstream task is NOT sufficient or deficient. Therefore, squeezing out knowledge from only one single pretext for a specific task very likely gives rise to a serious knowledge deficiency. In this section, our work is to compensate for the deficiency by supplying the specific task with its must complementary knowledge from diverse learned pretexts. Specifically, we propose an attention-based framework (Figure 2) to selectively extract or collect the representations distributed in diverse pretexts instead of searching for the single best pretext. The more details are described in the following subsections.

A. Preliminaries

Given a set of N pre-learned models \( \mathcal{F} = \{f_1, ..., f_N\} \) via some contrastive learning such as MoCo [3]. For them we need not know their details such as data used in pretext training and the architecture of the model except for its penultimate layer. We aim to squeeze out knowledge from \( \mathcal{F} \) to assist learning a downstream task \( \mathcal{T} = \{(x, y)\} \). With the \( \mathcal{F} \), we get \( x \) representation bank \( \mathcal{B} = \{b_i|b_i = f_i(x), i = 1, ..., N\} \), then we squeeze out so-needed representation \( z \) from \( \mathcal{B} \) for the task \( \mathcal{T} \) by aggregating \( z = agg(\mathcal{B}) \).

An ideal representation \( z \) should contain the maximal task-relevant knowledge associate with the specific downstream task \( \mathcal{T} \) (maximized information) and minimal task-irrelevant knowledge (minimized irrelevant information) [16].
Fig. 3: The illustration of complementary knowledge using the information diagram. \(1\) and \(2\) represent pretext 1 and pretext 2, respectively, and \(3\) represent downstream tasks. \(A\) and \(C\) are complementary knowledge; \(B\) is the knowledge downstream task shares with both pretext 1 and pretext 2.

It is clear that for any two different \(i \neq j\) pre-trained models \(f_i\) and \(f_j\), the knowledge they provide for a specific downstream task \((x, y)\) is complementary, meaning Eq. (3), where \(b_i = f_i(x)\) and \(b_j = f_j(x)\).

\[
I(b_i; y|b_j) > 0, \tag{3}
\]

For the pre-trained model \(f_i\), we squeeze out downstream task-specific representation \(\hat{z}_i\) from \(b_i\). According to Eq. (2), the \(\hat{z}_i\) has the property as following:

\[
I(y; b_i) = I(y; \hat{z}_i^{suf}) = I(y; \hat{z}_i) \geq I(y; \hat{z}_i). \tag{4}
\]

Based on Assumption 1, for a specific downstream task \((x, y)\), the aggregated representation \(z\) from \(B\) will be more sufficient than the \(\hat{z}_i\) from a single pre-trained model as shown in Theorem 1:

**Theorem 1.** When \(z = agg(B)\) is the sufficient representation of \((b_1, b_2, ..., b_N)\) for \(y\), then

\[
I(y; z) > \max \{I(y; \hat{z}_1), I(y; \hat{z}_2), ..., I(y; \hat{z}_N)\}. \tag{5}
\]

**Proof:** Based on DPI, we get following mutual information inequality of \(z\) and \(y\):

\[
I(y; z) \leq I(y; b_1, b_2, ..., b_N), \tag{6}
\]

And Eq. (2) implies that \(z^{suf}\) is the sufficient representation of \((b_1, b_2, ..., b_N)\) for \(y\), thus

\[
I(y; z^{suf}) = I(y; b_1, b_2, ..., b_N). \tag{7}
\]

So for any \(b_i (1 \leq i \leq N)\), based on the chain rule of mutual information, we have

\[
I(y; z^{suf}) = I(y; b_1, b_2, ..., b_N) = I(y; b_i) + I(b_1, ..., b_{i-1}, b_{i+1}, ..., b_N; y|b_i), \tag{8}
\]

Again, from the Eq. (3) in Assumption 1, we get

\[
I(b_1, ..., b_{i-1}, b_{i+1}, ..., b_N; y|b_i) > 0, \tag{9}
\]

and

\[
I(y; z^{suf}) > I(y; b_i), \tag{10}
\]

Incorporating the Eqs. (4), (6), (7), (10), we get

\[
I(y; z^{suf}) > I(y; b_i) = I(y; \hat{z}_i^{suf}) \geq I(y; \hat{z}_i), \tag{11}
\]

Hence,

\[
I(y; z^{suf}) > \max \{I(y; \hat{z}_1), I(y; \hat{z}_2), ..., I(y; \hat{z}_N)\}. \tag{12}
\]

**Remark:** Eq. (5) shows that an ideal representation \(z\) squeezed from bank \(B\) will contain more downstream knowledge than the \(\hat{z}_b m\) learned from the best matching pretext \(\hat{z}_b m\), where

\[
\hat{z}_b m = \arg \max_{\hat{z}_i} \{I(y; \hat{z}_1), ..., I(y; \hat{z}_i), ..., I(y; \hat{z}_N)\}. \tag{13}
\]
C. Self-Attention block & Cross-Attention block

In general, directly aggregating (e.g. adding and concatenating [32], [33]) all representations of the bank \( B \) will unfavorably introduce relatively much irrelevant information, resulting in poor performance on downstream task. To alleviate the negative effect, we would like to squeeze out the must representation with both maximal task-relevant knowledge and minimal task-irrelevant knowledge as described in Definition 1. The existent attention mechanism [32], [33] is able to minimal task-irrelevant knowledge as described in Definition 1. Accordingly, the output of the CAB is generated by Eq.(17).

\[
w_{ca}(b_i, b_j) = \begin{cases} 
\text{sigmoid}(q(b_i)k(b_j)^T), & N = 2, i \neq j \\
\sum_{j=1}^{N} \exp(q(b_i)k(b_j)^T), & N > 2, i \neq j 
\end{cases}
\]

Accordingly, the output of the CAB is generated by Eq.(17).

\[
z_i^{\text{cross}} = b_i + \sum_{j=1, j \neq i}^{n} w_{ca}(b_i, b_j)v(b_j).
\]

D. Architecture

We adaptively squeeze out the representations from the bank for the downstream task by the architecture equipped with SAB and CAB as Figure 2. We get the representation bank \( B = \{b_1, b_2\} \) of \( x \) by \( F = \{f_1, f_2\} \). For \( b_1 \) and \( b_2 \), we (1) firstly enlarge their own task-related knowledge by SA Blocks, and then (2) extract the complementary knowledge from the another by CA Blocks. The logits are obtained by a fully-connected (FC) layer for downstream classification task.

| Data set      | Train  | Test  | Features | classes |
|--------------|-------|------|----------|--------|
| CIFAR10      | 50,000| 10,000| 1024     | 10     |
| CIFAR100     | 50,000| 10,000| 1024     | 100    |
| SVHN         | 73,257| 26,032| 1024     | 10     |
| TinyImagenet | 100,000| 20,000| 4,096    | 200    |
| Chars74k     | 7,705 | 4,348| 1024     | 64     |

IV. EXPERIMENT

To further verify the effectiveness of the method proposed in Section 3, we conduct extensive image classification downstream tasks on benchmark datasets.
TABLE II: Classification accuracy (%) of transferring representation from single pre-trained model to different downstream tasks. The best result is indicated as bold.

| pre-trained model | SA Block | CIFAR10 | CIFAR100 | SVHN | Chars74K | CIFAR10 | CIFAR100 | SVHN |
|-------------------|----------|---------|----------|------|---------|---------|----------|------|
|                   | w/o SA   | 82.40   | 46.42    | 60.62| 31.99   | 79.95   | 40.01    | 54.65|
|                   | w/ SA    | 85.78   | 57.49    | 72.47| 44.54   | 80.57   | 41.73    | 66.87|
| CIFAR100          | w/o SA   | 75.64   | 51.90    | 65.98| 35.37   | 72.25   | 42.75    | 60.70|
|                   | w/ SA    | 81.54   | 60.61    | 76.33| 45.62   | 74.62   | 45.31    | 70.87|
| Tiny              | w/o SA   | 72.98   | 43.36    | 60.84| 30.58   | 69.23   | 34.77    | 53.67|
|                   | w/ SA    | 79.60   | 57.71    | 76.96| 50.27   | 74.16   | 42.05    | 68.48|
| SVHN              | w/o SA   | 62.90   | 32.90    | 92.37| 40.75   | 58.34   | 25.80    | 90.90|
|                   | w/ SA    | 71.62   | 46.82    | 94.40| 44.83   | 66.08   | 33.41    | 91.27|

TABLE III: Classification accuracy (%) of different aggregation feature scheme, add, concatenate and our attention-based. The best result is indicated as bold and the second is underlined.

| pre-trained models | CIFAR10+CIFAR100 | CIFAR10+Tiny | CIFAR100+Tiny | CIFAR10+SVHN | CIFAR100+SVHN | Tiny+SVHN |
|--------------------|------------------|---------------|---------------|--------------|--------------|-----------|
| add                | 83.23            | 83.15         | 77.65         | 82.71        | 77.03        | 73.57     |
| concatenate        | 83.60            | 83.38         | 78.34         | 83.01        | 77.48        | 74.41     |
| ours               | **86.08**        | 86.72         | **85.28**     | **86.21**    | **83.21**    | **82.53** |
| supervised         |                  |               |               |              |              |           |
|                   | **95.32**        |               |               |              |              |           |
| SVHN               |                  |               |               |              |              |           |
| add                | 68.62            | 65.71         | 68.83         | 92.37        | 92.27        | 92.37     |
| concatenate        | 69.48            | 66.84         | 69.38         | 92.52        | 92.65        | 92.49     |
| ours               | **80.61**        | **80.22**     | **81.18**     | **94.26**    | **94.66**    | **94.37** |
| supervised         |                  |               |               |              |              |           |
|                   | **96.55**        |               |               |              |              |           |
| CIFAR100           |                  |               |               |              |              |           |
| add                | 54.9             | 50.8          | 54.03         | 49.15        | 53.23        | 44.91     |
| concatenate        | 55.01            | 51.27         | 54.3          | 49.43        | 53.32        | 45.36     |
| ours               | **60.94**        | **60.26**     | **60.62**     | **60.62**    | **61.16**    | **58.72** |
| supervised         |                  |               |               |              |              |           |
|                   | **77.35**        |               |               |              |              |           |
| Chars74k           |                  |               |               |              |              |           |
| add                | 38.70            | 36.99         | 38.06         | 43.64        | 44.23        | 44.27     |
| concatenate        | 38.80            | 36.81         | 38.60         | 44.27        | 44.69        | 43.79     |
| ours               | **47.56**        | **48.79**     | **48.29**     | **50.5**     | **51.29**    | **51.63** |
| supervised         |                  |               |               |              |              |           |
|                   | **66.8**         |               |               |              |              |           |

A. Experimental Settings

1) Datasets: We adopt four datasets, CIFAR10 [35], CIFAR100 [35], SVHN [36] and TinyImagenet [37], for pretext training. And CIFAR10, CIFAR100, SVHN and Chars74k [38] are used as different four downstream task datasets. CIFAR10 has a total of 60000 color images from 10 natural image classes, with 6000 images per class. Each image size is 32 × 32, and the training and test sets are respectively 50000 and 10000. CIFAR100 is like CIFAR10, and it has 100 classes containing 600 images each. SVHN is a real world digit datasets containing 10 classes. It has 73257 images for training and 26032 digits for testing, and size is also like 32 × 32. TinyImagenet has a total of 200 classes with 500 images for each class for training, which is drawn from Imagenet ILSVRC 2012 and down-sampled to 64 × 64. Chars74k, we only use Englishlmg subset, has 62 classes of letters and numbers, with 7705 images training set and 4348 images testing set, and images are resized to like 32 × 32. Table-I reports detailed information on datasets.

2) Pretext and Pre-trained model: For the first stage of SSL, we take ResNet50 as the backbone and learn the semantic representation with MoCo.v2 (pretext task). For learning the CIFAR10, CIFAR100 and SVHN (and TinyImagenet) representation, we use mini-batch of 256 (64 for TinyImagenet), a dictionary number of 8192, and an initial learning rate of 0.1. We then train for 400 epochs with learning rate deweighted 0.1 at 200 and 300 epochs.

3) Downstream tasks: For the downstream classification, we leverage 4 datasets, natural images(CIFAR10 and CIFAR100) and numbers and letters (SVHN and the Chars74k Englishlmg subset). We trained downstream networks with the standard cross-entropy loss in the end-to-end way, where the mini-batch size is 128, the SGD weight decay is 0.0001, the SGD momentum is 0.9, and the initial learning rate is 0.1. We transfer the penultimate layer features of the pre-trained models. We further train for 200 (100) epochs for natural (numbers and letters) images datasets with learning rate deweighted 0.1 at 100 and 150 (75) epochs.

4) Supervised training: For the supervised paradigm baseline, we take ResNet50 as the backbone and train it with the standard cross-entropy loss in the end-to-end way for 500 epochs, where the mini-batch size is 128, the SGD weight...
B. Comparison Results

1) Transferring with single pre-trained model: We first study the performance of transferring knowledge from a single pre-trained model to downstream tasks. We transfer the representation to downstream tasks by training a task-specific linear classifier (w/o SA). Furthermore, to validate the effectiveness of the self-attention block, we train an additional self-attention block (w/ SA) with a linear classifier. The results are shown in Table II, where we draw the following two conclusions.

- When the specific downstream task is dissimilar with the pretext, simply training a task-specific classifier will significantly drop the performance (e.g. transferring from CIFAR10 to SVHN).
- Self-attention blocks can effectively extract downstream task-related knowledge, which improves the performance of downstream tasks.

2) Transferring with the representation bank: In this subsection, we conduct experiments on transferring representation from the bank by (a) adding the representations (add), (b) concatenating the pre-trained representation (concatenate), (c) our proposed attention-based representation aggregation (ours). Table-III summarizes the experimental results on two different categories (natural image and numbers/letters) downstream datasets. Table-IV summarizes the result on downstream tasks with small amounts of labeled data (10%). We conclude by following several observations.

Observation 1. Comparing experiment results in Table-II and Table-III, it strongly verifies that squeezing out representation from the bank (e.g CIFAR10+Tiny to CIFAR10) is better than a single model (CIFAR10/Tiny to CIFAR10).

Observation 2. Table-III and Table-IV results show that our method is superior to other methods. Meanwhile, we observe that the squeeze representation or downstream task performance from knowledge heterogeneous bank (CIFAR100+SVHN to CIFAR100) is almost consistent with that from knowledge homogeneous bank (CIFAR100+CIFAR10 to CIFAR100), and even exceeds the best single pre-training model in the bank. In the end, promising results demonstrate the powerful ability of our method to adaptively squeeze out representations from the bank.

Observation 3. The results in Table-IV show that when the downstream task has only a small amount of labeled data, SSL significantly surpasses the supervision paradigm. The experimental results again prove that our method is superior to other methods.

In summary, we draw a beneficial conclusion that selectively squeezing out representations from the representation bank can collect more complementary knowledge to assist downstream tasks learning.

To study the performance of our framework on object localization and object recognition, we apply Grad-CAM [39] to 4 different networks: (1) supervised training; (2) pre-trained Tiny net; (3) pre-trained CIFAR10 net; (4) pre-trained CIFAR10 and Tiny. All these networks are trained on the downstream CIFAR10 with 10% images and labels. We show our experimental results by Figure 5 and analyse them as follows:

- Image 1-2 shows that the supervised network focuses on the background classes, and our approach effectively overcomes this problem.
Image 3-12 shows that our network is able to focus on object localization information more precisely and is more robust to irrelevant background information.

The C1AIFR10+Tiny is better than Tiny and CIFAR10 which means that the multiple pre-trained models are better than the single pre-trained model.

C. Ablation Study

In this section, we further conduct ablation studies to validate the effectiveness of each component of the proposed model.

1) Attention block: To explore the merit of SAB and CAB in the framework, we respectively remove the SAB (w/o SA) and the CAB (w/o CA), observing its performance in downstream tasks (CIFAR100+Tiny to CIFAR10 & Chars74k). As shown in Figure 6, removing SAB or CAB will drop the performance. As for the CAB, it can aggregate complementary representation in different pretexts (Figure 6-(a), w/o SA is better than w/o CA). However, we conjecture taking the knowledge heterogeneous bank as the source, SAB is able to filter the irrelevant information (Figure 6-(b), w/o SA falls behind w/o CA).

We also conduct extensive comparison study on connection order of both attention blocks to validate the optimality of our architecture. Details are as follows:

we conduct extensive framework comparison experiments with SA Block and CA Block downstream (SVHN with 10% images and labels). We train 4 different type networks: (1) using SA block only (SA only), (2) using CA block only (CA only), (3) using SA first and CA later (SA2CA), (4) using CA first and SA later (CA2SA), (5) using SA and CA meanwhile (SCA). The results are shown in the Table-V. From the results, we conclude that using SA first and CA later is the best framework.

In addition, to improve the representation capacity, as [32], we also apply multi-head attention in our experiments. The experimental results are shown in Figure 5. With the increase of the number of heads, the aggregated representation gets stronger representation capacity and performs better on downstream tasks.
2) Diverse pretexts: In order to further investigate the performance of our model, it is verified by utilizing a more diverse representation bank. We first apply the multi-head attention mechanism to our framework and construct the multi-head attention (MHA) block. Then, we investigate the diverse pre-trained models’ impact on downstream tasks and conduct comparison experiments of transferring from triple pre-trained models. Finally, the results in Table-VI show that adding diverse pre-training knowledge can further improve the learning performance of downstream tasks. The above results once again prove our conclusion that selectively squeezing out representations from the representation bank can collect more complementary knowledge to assist downstream task learning.

V. CONCLUSION

In this paper, we explored how to adaptively squeeze the beneficial knowledge for downstream tasks learning from various pretexts. To this end, we propose a new solution that utilizes the attention mechanism to adaptively squeeze diverse complementary representations from the constructed representation bank to learn downstream tasks. Then, we leverage information theory to prove that the representation/knowledge from diverse pre-trained models can be used as a supplement to downstream tasks. Therefore, we come to a beneficial conclusion that purposefully and strategically squeezing out diverse representation knowledge from the representation bank is helpful to the learning of downstream tasks. For SSL, we provide a novel and effective method for solving how to transfer knowledge from various pretexts. Finally, a large number of experimental results prove that our proposed solution shows the performance of SOTA in collecting various pretexts knowledge and alleviating the negative migration of downstream tasks.

REFERENCES

[1] Y. Wang, Z. Ni, S. Song, L. Yang, and G. Huang, “Revisiting locally supervised learning: an alternative to end-to-end training,” in International Conference on Learning Representations, 2020.

[2] G.-J. Qi and J. Luo, “Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[3] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” arXiv preprint arXiv:1803.07728, 2018.

[4] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, “Contrastive learning with hard negative samples,” arXiv preprint arXiv:2010.04592, 2020.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” arXiv preprint arXiv:2002.05709, 2020.

[7] C. Bengio, A. S. Pinto, L. Rimanic, J. Puigcerver, C. Riquelme, C. Zhang, and M. Lucic, “Which model to transfer? finding the needle in the growing haystack,” arXiv preprint arXiv:2010.06402, 2020.

[8] Y. Shi, Z. Kou, Z. Cao, J. Wang, and M. Long, “Zoo-transfer: Adaptive transfer from a zoo of models,” in International Conference on Machine Learning. PMLR, 2021, pp. 9626–9637.

[9] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 132–140.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheishlaghi Azar et al., “Bootstrap your own latent-a new approach to self-supervised learning,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[11] L. Ericsson, H. Gouk, and T. M. Hospedales, “How well do self-supervised models transfer?” arXiv preprint arXiv:2011.13377, 2020.

[12] C. V. Nguyen, T. Hassner, C. Archambeau, and M. Seeger, “Leep: A new measure to evaluate transferability of learned representations,” arXiv preprint arXiv:2002.12462, 2020.

[13] J. Ngiam, A. Khasla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in ICMlr, 2011.

[14] N. Srivastava and R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” The Journal of Machine Learning Research, vol. 15, no. 1, pp. 2949–2980, 2014.

[15] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, “Boosting self-supervised learning via knowledge transfer,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9359–9367.

[16] N. Tiehby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in 2015 IEEE Information Theory Workshop (ITW). IEEE, 2015, pp. 1–5.

[17] A. Papadopoulos, P. Korus, and N. Memon, “Hard-attention for scalable image classification,” arXiv preprint arXiv:2102.10212, 2021.

[18] D. Wang, A. Haytham, J. Pottenburgh, O. Saeedi, and Y. Tao, “Hard attention net for automatic retinal vessel segmentation,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3384–3396, 2020.

[19] S. Shankar, S. Garg, and S. Sarawagi, “Surprisingly easy hard-attention for sequence to sequence learning,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 640–645.

[20] D. Jain, A. Kumar, and G. Garg, “Sarcasm detection in mash-up language using soft-attention based bi-directional lstm and feature-rich cnn,” Applied Soft Computing, vol. 91, p. 106198, 2021.

[21] M. McClenny and U. Braga-Neto, “Self-adaptive physics-informed neural networks using a soft attention mechanism,” arXiv preprint arXiv:2009.04544, 2020.

[22] Y. Guan, H. Cui, Y. Xu, Q. Jin, T. Feng, H. Tu, P. Xuan, W. Li, L. Wang, and B.-L. Duh, “Predicting esophageal fistula risks using a multimodal self-attention network,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 721–730.

[23] X. Guo, X. Guo, and Y. Lu, “Scan: Separable self-attention network for video representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12618–12627.

[24] Z. Li, F. Ling, C. Xu, C. Zhang, and H. Ma, “Cross-media hash retrieval using multi-head attention network,” in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 1290–1297.

[25] J. Wang, X. Peng, and Y. Qiao, “Cascade multi-head attention networks for action recognition,” Computer Vision and Image Understanding, vol. 192, p. 102898, 2020.

[26] X. Shang, Q. Ma, Z. Lin, J. Yan, and Z. Chen, “A span-based dynamic local attention model for sequential sentence classification,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 198–203.

[27] P. M. Hanunggul and S. Suyanto, “The impact of local attention in lstm for action recognition,” Language Processing (Volume 2: Short Papers) of the 59th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 72–77.

[28] Z. Li, F. Ling, C. Xu, C. Zhang, and H. Ma, “Cross-media hash retrieval using multi-head attention network,” in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 1290–1297.

[29] J. Wang, X. Peng, and Y. Qiao, “Cascade multi-head attention networks for action recognition,” Computer Vision and Image Understanding, vol. 192, p. 102898, 2020.

[30] X. Shang, Q. Ma, Z. Lin, J. Yan, and Z. Chen, “A span-based dynamic local attention model for sequential sentence classification,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 198–203.

[31] P. M. Hanunggul and S. Suyanto, “The impact of local attention in lstm for action recognition,” Language Processing (Volume 2: Short Papers) of the 59th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 72–77.
[29] D. Ren, J. Li, M. Han, and M. Shu, “Scga-net: Skip connections global attention network for image restoration,” in *Computer Graphics Forum*, vol. 39, no. 7, Wiley Online Library, 2020, pp. 507–518.

[30] A. Achille and S. Soatto, “Emergence of invariance and disentanglement in deep representations,” *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1947–1980, 2018.

[31] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[33] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[34] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 6281–6290.

[35] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.

[36] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” 2011.

[37] Y. Le and X. Yang, “Tiny imagenet visual recognition challenge,” *CS 217N*, vol. 7, 2015.

[38] T. E. De Campos, B. R. Babu, M. Varma *et al.*, “Character recognition in natural images.” *VISAPP (2)*, vol. 7, 2009.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.