# Maximizing the Effectiveness of Larger BERT Models for Compression

**Wen-Shu Fan[1,2], Su Lu[3], Shangyu Xing[1,2], Xin-Chun Li[1,2], De-Chuan Zhan[1,2]**
[1]School of Artificial Intelligence, Nanjing University, China,
[2]National Key Laboratory for Novel Software Technology, Nanjing University, China,
[3]Baiont Quant
fanws@lamda.nju.edu.cn, zhandc@lamda.nju.edu.cn

## Abstract

Knowledge distillation (KD) is a widely used approach for BERT compression, where a larger BERT model serves as a teacher to transfer knowledge to a smaller student model. Prior works have found that distilling a larger BERT with superior performance may degrade student's performance than a smaller BERT. In this paper, we investigate the limitations of existing KD methods for larger BERT models. Through Canonical Correlation Analysis, we identify that these methods fail to fully exploit the potential advantages of larger teachers. To address this, we propose an improved distillation approach that effectively enhances knowledge transfer. Comprehensive experiments demonstrate the effectiveness of our method in enabling larger BERT models to distill knowledge more efficiently.

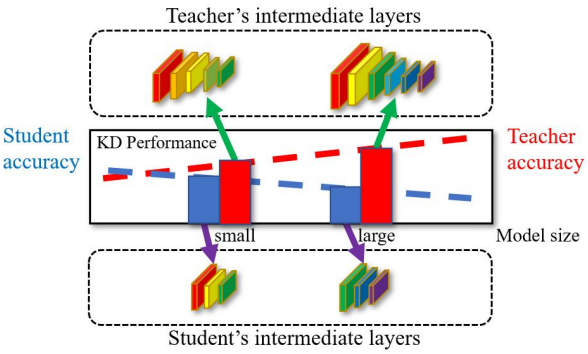


Figure 1: A larger BERT teacher model fails to achieve better distillation as the student's intermediate layers exhibit smaller linear differences. The color variation of intermediate layers represents their linear differences. The student trained by the smaller teacher (red-yellow-green) shows greater overall variation than the one trained by the larger teacher (green-blue-purple).

## 1 Introduction

BERT (Devlin, 2018) has achieved significant success in natural language processing (NLP). However, deploying large BERT models is challenging on resource-constrained platforms such as mobile device. Knowledge distillation (KD) provides an effective approach for compressing large BERT models into smaller ones. Specifically, it leverages a pretrained teacher model to guide the training of a lightweight student model (Hinton, 2015).

KD applied to BERT has made significant progress (Sun et al., 2019; Jiao et al., 2019; Sanh, 2019; Guo et al., 2023). However, the classical PKD method (Sun et al., 2019) reveals an intuitive yet surprising phenomenon: despite their superior performance, utilizing larger BERT models as teacher does not necessarily lead to better KD performance. This reflects the well-known capacity mismatch problem in KD (Wang et al., 2022), where increasing teacher size does not always enhance distillation effectiveness (Cho and Hariharan, 2019). While this issue has been widely studied, existing solutions face limitations in BERT distillation. Some external model-based approaches like TAKD (Mirzadeh et al., 2020) overlook internal data transformations within the model and fail to explain why larger teachers underperform in distillation. Other methods (Huang et al., 2022; Li et al., 2022; Fan et al., 2024) can explain the reason but rely heavily on category information, making them unsuitable for tasks with limited categories, such as GLUE (Wang, 2018) in NLP.

Instead of category information, this paper investigates why larger BERT teacher models fail to achieve better distillation performance from the perspective of linear relationships between intermediate layers. We find larger teacher models exhibit greater overall linear differences across layers, which contribute to improved pre-training performance. However, in previous KD methods, students trained by larger teachers do not show greater linear differences among intermediate layers than those trained by smaller teachers, limiting their generalization ability. As a result, larger teachers fail to distill more effectively, as illustrated in Figure 1.



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21975–21990

We propose a method to increase the linear differences among the student's intermediate layers from two perspectives. (1) We argue that if a teacher model's intermediate layer exhibits a larger linear difference with its preceding layer, it extracts more critical information. We should select such layer as teacher. (2) We also directly maximize the linear relationship between the student model's intermediate layers and the teacher's corresponding distillation layers following (Andrew et al., 2013). As a result, the student model's intermediate layers will exhibit significant linear differences, enhancing generalization ability. Consequently, larger teacher models yield better distillation performance.

Our contributions are as follows:

- We claim that larger BERT teachers fail to increase linear differences among the student's intermediate layers, resulting in poorer KD performance.
- We propose MC3KD framework, which can amplify the linear differences among student's intermediate layers by selecting suitable teachers' intermediate layers and minimizing linear differences between selected teachers and student's layers.
- Extensive experimental comparisons demonstrate that our MC3KD framework does maximize the effectiveness of larger BERT models for compression.

## 2 Related Work

**Language Model Compression** Model compression can make deep neural networks more compact (Buciluǎ et al., 2006). Language models can be compressed by network pruning (He et al., 2017), weight quantization (Polino et al., 2018), weight sharing (Dehghani et al., 2018), low-rank approximation (Ma et al., 2019) or knowledge distillation (Sun et al., 2019; Sanh, 2019; Jiao et al., 2019). In this paper, we focus on knowledge distillation.

**Knowledge Distillation for BERT** Similar to leveraging pre-trained models to assist in training (Zhou, 2016), Knowledge Distillation (KD) improves a smaller student model's generalization by learning from a larger teacher model. KD methods can be broadly classified into three types (Gou et al., 2021) based on the knowledge transferred: (1) Logit-based KD, where the student mimics the teacher's logit outputs (Hinton, 2015; Sun et al., 2024; Yang et al., 2024b; Zhang et al., 2025); (2) Feature-based KD, which transfers intermediate-layer representations (Romero et al., 2014; Yang et al., 2021a, 2023a, 2024a,c,d); (3) Relation-based KD, which captures relationships between layers or samples within the teacher model (Tian et al., 2019; Kweon et al., 2021; Yang et al., 2021b). KD has been widely applied to compress BERT models. DistilBERT (Sanh, 2019), a logit-based KD method, uses the teacher model's output as a supervision signal to align the student's predictions. PKD (Sun et al., 2019), a feature-based KD approach, extracts CLS token representations from the teacher's intermediate layers for teaching. TinyBERT (Jiao et al., 2019) combines logit-based and feature-based KD, transferring knowledge through word embeddings, self-attention heads and selected intermediate-layer representations. CoDIR (Sun et al., 2020), a relation-based KD method, captures structural knowledge within intermediate layers. While these approaches leverage different types of knowledge, they overlook how data properties transform throughout the model, leading to suboptimal knowledge extraction.

In PKD, researchers observed that a larger BERT teacher model does not necessarily yield better distillation performance. This reflects a phenomenon in KD named capacity mismatch that as the size of teacher increases, student may perform worse generalization. Existing solutions to capacity mismatch fall into two categories. The first employs external strategies, such as auxiliary models or early stopping (Mirzadeh et al., 2020; Cho and Hariharan, 2019; Wang et al., 2022; Yang et al., 2023b). These methods keep overall model outputs preserved, while they do not explicitly explain why larger teachers underperform in distillation. The second examines intrinsic properties, such as ranking, variance and calibration of outputs (Huang et al., 2022; Li et al., 2022; Fan et al., 2024), but these approaches rely on class diversity and are mainly designed for vision tasks with large label spaces, making them unsuitable for NLP tasks like GLUE (Wang, 2018). Our method addresses both limiations by identifying the root cause of ineffective distillation in large teachers to compress larger BERT models more effectively.

**Analysis of Similarity of Representation** Several methods measure neural network similarity. Canonical Correlation Analysis (CCA) is a classical statistical approach for assessing similarity

between two sets of multivariate data (Hotelling, 1936; Anderson, 1985; Hardoon et al., 2004). Singular Value Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017) extends CCA by applying Singular Value Decomposition (SVD) to reduce dimensionality. Recently, CKA (Kornblith et al., 2019), a kernel-based method, has been introduced to capture nonlinear relationships between networks. However, the computational expense of CKA is high so it is impractical for resource-heavy models like BERT. We use CCA, which also offers a stable theoretical foundation, to probe BERT. The study by (Hao et al., 2020) closely aligns with our work, it employs SVCCA to analyze how BERT layers change during fine-tuning. Unlike them, we use SVCCA to explore the layer-wise correlations that well-generalizing models exhibit and propose a method to enhance distillation performance based on this analysis.

## 3 Background and Notations

### 3.1 Vanilla KD

In our setting, each sample in training set is defined as $\{\mathbf{x}, y\}$, where $\mathbf{x}$ refers to text input and $y$ refers to label. Suppose there is a teacher network denoted as $\mathcal{T}$ and a student network $\mathcal{S}$. The outputs of teacher network and student network can be denoted as $\mathcal{T}(\mathbf{x})$ and $\mathcal{S}(\mathbf{x})$, respectively.

The vanilla Knowledge Distillation (KD) loss consists of two parts (Hinton, 2015). The first part minimizes the difference between the teacher and the student, allowing the student to acquire knowledge from teacher. This difference is typically measured using the Kullback-Leibler (KL) divergence. The second part is to account for potential errors in teacher by aligning the student's output with the labels via Cross-Entropy (CE) loss. The total loss of vanilla KD can be written as:

$$\mathcal{L}_{\text{KD}} = \alpha \text{KL}(\mathcal{S}(\mathbf{x}), \mathcal{T}(\mathbf{x})) + (1-\alpha)\text{CE}(\mathcal{S}(\mathbf{x}), y) \quad (1)$$

where $\alpha$ is the weight balancing the two parts, with a value range from 0 to 1.

### 3.2 Comparison of Canonical Correlation

The Singular Vector Canonical Correlation Analysis (SVCCA) matrix measures the similarity between the representations of different layers in a neural network (Raghu et al., 2017). We regard the elements in SVCCA matrix as **Canonical Correlation Coefficients (CCC)**. Suppose there is an SVCCA matrix $S$ for network $f$, a CCC $S_{ij}$ in $S$ indicates linear relationship between $i$-th layer and $j$-th layer in $f$. Larger $S_{ij}$ indicate stronger correlations between layer $i$ and layer $j$.

Suppose the $m$-dimensional SVCCA matrices of models distilled (or trained) by method A and method B are respectively denoted as $A$ and $B$. If almost all CCCs in matrix A is smaller than all corresponding CCCs in matrix B, we can say model A shares larger linear differences among its intermediate layers than B. From all our experimental data in Figure 9, we find that for the same model, the relative magnitudes of corresponding CCC in the SVCCA matrix remain almost consistent across different training methods. That is, given $1 \leq p \leq m$ and $1 \leq q \leq m$, if $A_{pq} > B_{pq}$, then $A_{ij} > B_{ij}$ for almost all $i \in [1, m]$ and $j \in [1, m]$. So comparing two SVCCA matrices can be sufficed to select a representative element from each matrix for comparison. In the example above, we choose the CCC element $A_{1m}$ as our **Representative Canonical Correlation Coefficient (RCCC)** to represent linear differences among the model's intermediate layers.

In all the tables that follow in this paper, **RCCC represents the whole SVCCA matrix for comparison**, with smaller values indicating greater linear differences among the model's intermediate layers.

## 4 Why Larger BERT Teachers Fail to Teach Well?

We aim to identify a metric that reflects both *the superior finetuning performance of the larger BERT teacher* and those are *not well inherited by the BERT student taught by larger teacher*.

In NLP, the category space for many tasks is quite small. Take the classical GLUE benchmark as an example, there are only two classes in many classification tasks. To answer why larger BERT teachers fail to teach well, we cannot analyze the category information contained in the output logits as some methods (Huang et al., 2022; Li et al., 2022; Fan et al., 2024) for addressing capacity mismatch in the vision field do. A natural idea is to study the corrlation of hidden outputs from the intermediate layers of the BERT teacher. To implement this idea, we select teacher models that have been fine-tuned on the task and compute the SVCCA matrix (Raghu et al., 2017) for all their intermediate layers.

From the heatmap of SVCCA matrices of pretrained models on RTE task in Figure 2, we observe

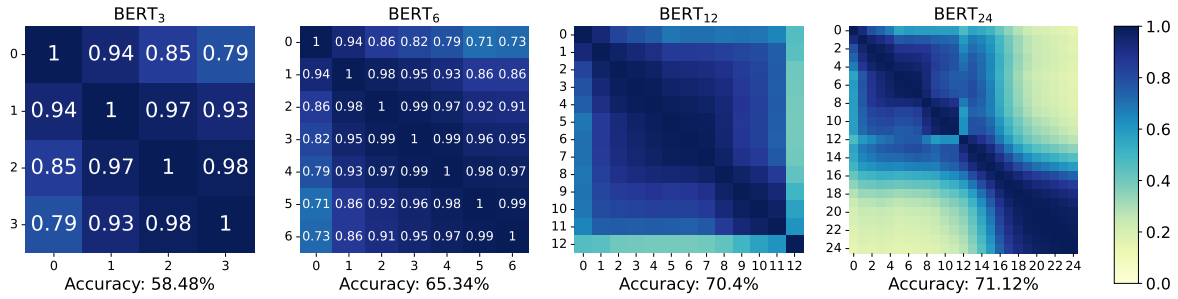


Figure 2: Heatmaps of SVCCA for BERT pre-trained models of varying sizes, with the RTE task as an example.

that as the teacher model size increases, linear differences between intermediate layers become more pronounced. For instance, in the first column (or row) of a matrix, which represents the relationship between the first intermediate layer and all others, a 3-layer model maintains strong linear correlations across layers. However, in larger models, the correlation between the first and last few layers weakens significantly. As shown in Figure 6, this pattern also holds for other tasks. This indicates that increasing model size amplifies linear differences between intermediate layers, thereby improving generalization performance. From this perspective, better generalization can be attributed to the *greater diversity* of features extracted across layers.

Building on this, we investigate whether greater layer differences in the student model correlate with better distillation performance. Table 1 shows the KD performance and SVCCA matrix (represented by RCCC) for student models taught by different teacher models on the RTE task, with results for other tasks in Table 3.

| Teacher-Student | Acc | RCCC |
|---|---|---|
| $BERT_{12}$-$BERT_6$ (KD) | **65.70%** | **70.60%** |
| $BERT_{24}$-$BERT_6$ (KD) | 64.98% | 70.74% |
| $BERT_{12}$-$BERT_6$ (PKD) | **65.34%** | **70.85%** |
| $BERT_{24}$-$BERT_6$ (PKD) | 64.62% | 71.25% |

Table 1: KD performance and SVCCA matrix for student models on the RTE task (with different KD methods in parentheses). Better results are bolded.

From Table 1, we observe that the distillation performance is positively correlated with the linear differences between the student model's intermediate layers. Additionally, we find that under classical methods such as KD and PKD, *a larger teacher model does not effectively enhance the diversity of representations across the student model's intermediate layers* (the reason will be explained further in Section 5.1). As a result, the larger model fails to achieve improved distillation performance.

**Do larger BERT teachers have the potential to teach better?** If we can more effectively extract the "dark knowledge" from the teacher model, a larger teacher could indeed teach better. Given that the logits in BERT models contain limited information for tasks such as GLUE, we focus on identifying intermediate layers that encode more knowledge. We analyze the magnitude of changes in linear relationships between each layer and its preceding layer to identify key layers that perform significant feature transformations. Therefore, in Figure 2, we focus on the elements below the diagonal, specifically $A_{i,i-1}$ in the given SVCCA matrix $A$. These elements reflect the magnitude of CCC changes between adjacent intermediate layers. As the model size increases, these values become smaller or more dispersed. Larger figures in Figure 8 provide a clearer illustration of this phenomenon. This suggests that as the model size increases, adjacent intermediate layers exhibit more abrupt changes, making it easier to identify layers that extract critical information. Thus, for larger models, if we can select intermediate layers practically, it is possible to utilize more knowledge contained in larger BERT for distillation, *increasing the potential for better distillation performance with a larger teacher*.

## 5 Proposed Method

Building on the previous findings, enabling a larger teacher BERT to teach more effectively requires increasing the differences between the student model's layers. To achieve this, we adopt a two-step approach. First, we select the teacher's intermediate layers that exhibit the largest linear differences from their preceding layers to guide the student. Second, we employ a distillation objective

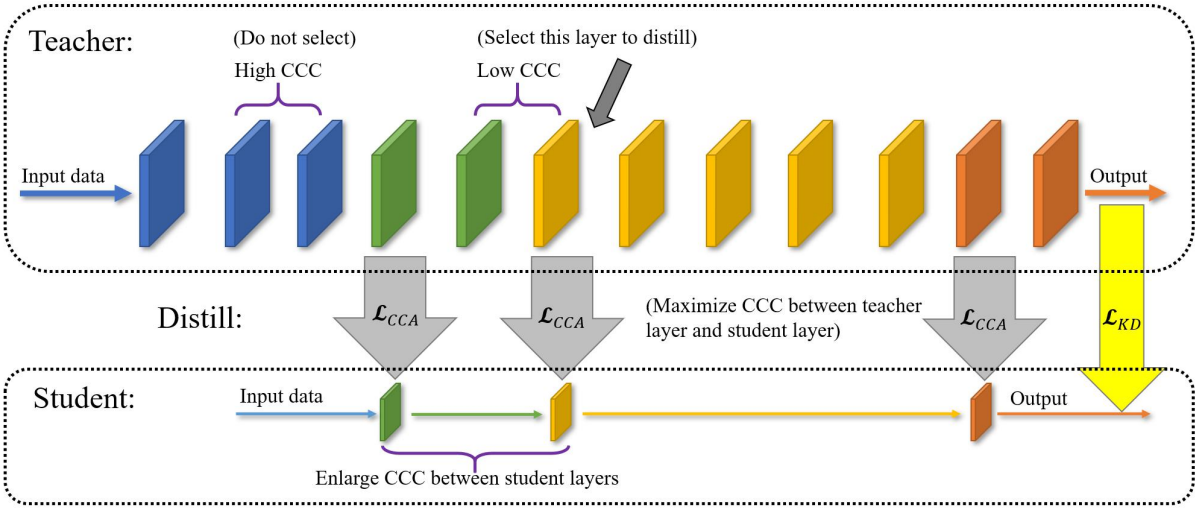


Figure 3: The framework of our proposed MC3KD.

that strengthens the linear relationships between the corresponding teacher and student layers, thereby amplifying the linear differences across the student model's layers. We refer to our method as **M**aximizing **C**anonical **C**orrelation **C**oefficients **K**nowledge **D**istillation (MC3KD). The overall framework is illustrated in Figure 3.

### 5.1 Find Proper Teacher

We first answer why KD and PKD fail to reduce RCCC, as observed in Table 1. KD relies solely on the teacher model's logits, which are unrelated to RCCC. PKD selects intermediate layers either at fixed intervals (PKD-Skip) or from the last few layers (PKD-Last), which may also neglect suitable teacher layers. For example, as shown in the SVCCA matrix $A$ (Figure 7) of teacher models in Table 1, $A_{1,0} = 0.93$ is the second largest among $A_{i,i-1}$, which means that the first layer has the second largest linear difference from the previous layer and is well-suited as a teacher layer. However, neither PKD selection strategies will select the first layer. So PKD fails to select the most linearly distinct layers, thereby restricting the effective utilization of the larger BERT teacher's intermediate layer differences to reduce RCCC.

To address this, we aim to select intermediate layers that capture the most valuable knowledge, characterized by minimal RCCC with their preceding layers. Given a student model with $m$ intermediate layers and a teacher model with $N$ intermediate layers, our goal is to identify the $m$ layers from the $N$ available in the teacher model that satisfy these criteria. Formally, we define the output of the $l$-th intermediate layer of the teacher model as $\mathcal{T}_l(\mathbf{x})$, where $l \in [0, 1, 2, ..., N]$ (abbreviated as $\mathcal{T}_l$). Note that $\mathcal{T}_N(\mathbf{x})$ is equivalent to $\mathcal{T}(\mathbf{x})$ as defined in Section 3.1. The "intermediate layers" we refer to here are the Transformer encoders in BERT, with the embedding layer typically considered as layer 0, which is output alongside the hidden features in the official implementation[1]. Therefore, the index $l$ starts from 0, but this does not affect our results, as we are selecting appropriate Transformer layers, not the embedding layer. Starting from $l = 1$, we compute CCC between each intermediate layer and its preceding layer, forming a CCC sequence:

$$C_{\mathcal{T}}(l) = \text{CCA}(\mathcal{T}_{l-1}, \mathcal{T}_l). \quad (2)$$

We denote the set of selected intermediate layer indices as $S = \{l_i \mid i \in [1, m]\}$ and define $R$ as the set of remaining layer indices, where $R = [1, N] \setminus S$. The selected intermediate layers satisfy the following conditions:

$$\max\{C_{\mathcal{T}}(l_1), ..., C_{\mathcal{T}}(l_m))\} < \min_{j \in R}\{C_{\mathcal{T}}(j)\} \quad (3)$$

where $1 \leq l_1 \leq l_2 \leq ... \leq l_m \leq N$ ensures that the selected $m$ teacher intermediate layers maintain their original order in teacher model.

### 5.2 Maximizing CCC Between Intermediate Layers

After selecting the most suitable teacher intermediate layers, we need to design a more effective distillation objective. The MSE loss used in PKD is not optimal for maximizing linear relationship between

[1]https://github.com/huggingface/transformers

teacher's and student's corresponding intermediate layer(Hastie, 2009; Trigeorgis et al., 2016; Tzirakis et al., 2017; Köprü and Erzin, 2020; Zhou et al., 2024). To address this, we define the loss function directly based on the core objective—maximizing the CCC between corresponding layers. However, CCC computation involves the Pearson correlation coefficient, which is non-differentiable and obstructs gradient backpropagation, making it unsuitable as an optimization objective. Therefore, we adopt the surrogate function used in (Andrew et al., 2013) to compute CCC between the corresponding intermediate layers of the teacher and student models.

We denote the selected intermediate layer outputs of the teacher model as $\mathcal{T}_{l_i}$ and the corresponding intermediate layer outputs of the student model as $\mathcal{S}_i$ for each layer, where $i \in [1, m]$. Then, the centered matrix data of $\mathcal{T}_{l_i}$ is calculated as $\bar{\mathcal{T}}_{l_i} = \mathcal{T}_{l_i} - \frac{1}{m}\mathcal{T}_{l_i}\mathbf{1}$ (resp. $\bar{\mathcal{S}}_i$). We define a whitened cross-covariance matrix $T_i$ as

$$T = (\hat{\Sigma}^i_{tt})^{-1/2}\hat{\Sigma}^i_{ts}(\hat{\Sigma}^i_{ss})^{-1/2} \tag{4}$$

where

$$\begin{aligned} \hat{\Sigma}^i_{tt} &= \frac{1}{m-1}\bar{\mathcal{T}}_{l_i}(\bar{\mathcal{T}}_{l_i})^\top + r_t I \\ \hat{\Sigma}^i_{ts} &= \frac{1}{m-1}\bar{\mathcal{T}}_{l_i}(\bar{\mathcal{S}}_i)^\top \\ \hat{\Sigma}^i_{ss} &= \frac{1}{m-1}\bar{\mathcal{S}}_i(\bar{\mathcal{S}}_i)^\top + r_s I \end{aligned} \tag{5}$$

As mentioned in (Andrew et al., 2013), the total correlation of the top $K$ components of $\mathcal{T}_{l_i}$ and $\mathcal{S}_i$ is the sum of the top $K$ singular values of $T_i$. So calculation of CCC can be modified as:

$$\operatorname{corr}(\mathcal{T}_{l_i}, \mathcal{S}_i) = \|T_i\|_{\mathrm{tr}} = \operatorname{tr}\left(T_i^\top T_i\right)^{1/2} \tag{6}$$

and the CCC loss is defined as:

$$\mathcal{L}_{CCC} = -\sum_{i=1}^{m} \operatorname{corr}(\mathcal{T}_{l_i}, \mathcal{S}_i) \tag{7}$$

### 5.3 MC3KD

Similar to PKD, we use the hyper-parameter $\beta$ to weight the importance of the CCC loss. The complete loss definition is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{KD}} + \beta\mathcal{L}_{CCC} \tag{8}$$

The entire process described in Sections 5.1 and 5.2 constitutes the full MC3KD framework. The algorithm is shown in Algorithm 1.

**Algorithm 1** MC3KD

**Input:** A sequence of words $\mathbf{x}$ with label $y$
**Params:** $\mathcal{T}_i$ Teacher output at the $i$-th layer
$\mathcal{S}_i$ Student output at the $i$-th layer
$N$ Number of layers of teacher
$m$ Number of layers of student
$\alpha$ Hyper-parameter for KD Loss
$\beta$ Hyper-parameter for CCC Loss
**Output:** Total loss function $\mathcal{L}$

**for** $i = 1$ to $N - 1$ **do**
    Calculate $C_{\mathcal{T}}(i)$ according to Equation (2).
**end for**
Pick out the $m$ smallest teacher layers $\{\mathcal{T}_{l_1}, ..., \mathcal{T}_{l_m}\}$ that satisfies Equation (3).
Calculate $\mathcal{L}_{CCC}$ according to Equation (7).
Calculate $\mathcal{L}_{KD}$ according to Equation (1).
Calculate $\mathcal{L}$ according to Equation (8).
**return** $\mathcal{L}$

## 6 Experiments

This section presents a comprehensive evaluation of MC3KD from multiple perspectives.

### 6.1 Datasets

We conduct evaluation experiments on the GLUE (Wang, 2018) benchmark. Specifically, we evaluate our proposed approach on tasks including Sentiment Classification, Paraphrase Similarity Matching, Natural Language Inference, and Linguistic Acceptability. For Sentiment Classification, we test on the Stanford Sentiment Treebank (SST-2) (Socher et al., 2013). For Paraphrase Similarity Matching, we use the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) and Quora Question Pairs (QQP) datasets. For Natural Language Inference, we evaluate on Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2017), Question-answering Natural Language Inference (QNLI) (Rajpurkar, 2016), and Recognizing Textual Entailment (RTE). For Linguistic Acceptability, we use Corpus of Linguistic Acceptability (CoLA) (Warstadt, 2019).

The label space in tasks of the GLUE benchmark is relatively small. For example, QQP is designed to predict whether a pair of questions is a duplicate, based on data from the popular online question-answering website Quora. As a result, all samples in QQP only belong to one of two categories.

| Teacher-Student | CoLA Matt | RTE (Acc) | MRPC (F1/Acc) | STS-B (Pear/Spear) | SST-2 (Acc) | QNLI (Acc) | QQP (F1/Acc) | MNLI-m (Acc) | MNLI-mm (Acc) |
|---|---|---|---|---|---|---|---|---|---|
| $BERT_{12}$ (teacher) | 57.79 | 70.40 | 99.29/99.02 | 87.42/87.18 | 91.86 | 89.58 | 87.67/90.73 | 84.49 | 84.72 |
| $BERT_{24}$ (teacher) | 59.40 | 71.12 | 99.46/99.27 | 88.19/87.99 | 92.89 | 91.74 | 87.98/91.08 | 84.90 | 85.27 |
| $BERT_{12}$-$BERT_3$ | 16.97 | 60.29 | 91.57/87.5 | 83.21/82.72 | 88.19 | 84.57 | 84.79/88.19 | 76.32 | 76.58 |
| $BERT_{24}$-$BERT_3$ | **20.05** | **61.73** | **92.28/88.97** | **83.71/83.50** | **88.53** | **85.10** | **85.09/88.48** | **76.85** | **77.35** |
| $BERT_{12}$-$BERT_6$ | 44.12 | 66.43 | 98.58/98.04 | 88.26/87.97 | 91.17 | 88.67 | **87.22/90.50** | 81.93 | 82.00 |
| $BERT_{24}$-$BERT_6$ | **44.81** | **67.15** | **99.64/99.51** | **88.37/87.98** | **91.40** | **88.94** | 87.17/90.43 | **82.14** | **82.21** |

Table 2: KD performance (%) of teacher models of different capacity on GLUE benchmark.

| Method | CoLA (Matt) | CoLA (RCCC) | RTE (Acc) | RTE (RCCC) | MRPC (F1/Acc) | MRPC (RCCC) | STS-B (Pear/Spear) | STS-B (RCCC) |
|---|---|---|---|---|---|---|---|---|
| Teacher: $BERT_{12}$ Student: $BERT_3$ | | | | | | | | |
| KD | 16.29 | 21.51 | 57.40 | 82.15 | 86.71/79.41 | 86.10 | 80.99/80.85 | 83.27 |
| PKD | 16.72 | 19.06 | 59.21 | 80.43 | 90.46/85.78 | 84.99 | 82.48/82.11 | 79.78 |
| MC3KD | **16.97** | **17.40** | **60.29** | **79.71** | **91.57/87.50** | **80.82** | **83.21/82.72** | **68.81** |
| Teacher: $BERT_{24}$ Student: $BERT_3$ | | | | | | | | |
| KD | 14.89 | 38.78 | 57.76 | 81.67 | 88.05/82.11 | 85.73 | 82.38/82.30 | 70.31 |
| PKD | 15.28 | 34.37 | 58.84 | 81.26 | 86.00/79.41 | 86.54 | 81.43/80.69 | 84.29 |
| MC3KD | **20.05** | **15.32** | **61.73** | **78.41** | **92.28/88.97** | **70.75** | **83.71/83.50** | **64.11** |

Table 3: The relationship between distillation performance (%) and RCCC (%).

| Method | CoLA (Matt) | CoLA (RCCC) | RTE (Acc) | RTE (RCCC) | MRPC (F1/Acc) | MRPC (RCCC) | STS-B (Pear/Spear) | STS-B (RCCC) | SST-2 (Acc) | SST-2 (RCCC) |
|---|---|---|---|---|---|---|---|---|---|---|
| PKD | 15.28 | 34.37 | 58.84 | 81.26 | 86.00/79.41 | 86.54 | 81.43/80.69 | 84.29 | 87.73 | 0.52 |
| TAKD | 14.61 | 55.67 | 59.21 | 80.38 | 88.23/82.60 | 85.47 | 82.22/81.84 | 77.36 | 86.81 | 3.41 |
| MC3KD | **20.05** | **15.32** | **61.73** | **78.41** | **92.28/88.97** | **70.75** | **83.71/83.50** | **64.11** | **88.53** | **0.48** |

Table 4: Distillation performance (%) and RCCC (%) of MC3KD with TAKD.

### 6.2 Implementation Details

Following prior works (Sun et al., 2019; Zhou and Xu, 2022; Guo et al., 2023), we evaluate MC3KD in a task-specific setting, where the teacher model is fine-tuned on downstream tasks, and the student model is trained on these tasks during distillation. We fine-tune BERT-Base (denoted as $BERT_{12}$) as teacher model and a 24-layer Transformer model ($BERT_{24}$) as the larger teacher for each task. Pre-trained weights are sourced from the official BERT repository on HuggingFace. For student models, we use 3-layer and 6-layer Transformer architectures ($BERT_3$ and $BERT_6$), respectively. Hyperparameter details are displayed in Appendix A.

### 6.3 Experimental Results

We apply our MC3KD to distill the same student model using teacher models of varying sizes, with results shown in Table 2. As observed, the larger teacher model, $BERT_{24}$, outperforms the smaller $BERT_{12}$ in its own performance. Moreover, under MC3KD, larger teachers consistently yield better distillation results, suggesting that MC3KD can enhance the effectiveness of larger BERT teachers. Table 3 shows that MC3KD outperforms both KD and PKD in distillation performance, while MC3KD really enlarges linear differences among student BERTs' intermediate layers. Table 7 in appendix B.2 presents more comparative data.

We also compare MC3KD with TAKD (Mirzadeh et al., 2020), a classic method designed to enhance the teaching ability of larger teacher models. TAKD introduces an intermediate assistant model, whose size falls between the teacher and student, to facilitate knowledge transfer. In our experiments, we use $BERT_{24}$ as the teacher, $BERT_6$ as the assistant, and $BERT_3$ as the student. TAKD improves upon PKD by increasing the diversity of the student model, leading to moderate performance gains. However, as shown in Table 4, TAKD fails to fully leverage the substantial linear differences across the intermediate layers of larger teacher models, resulting in a smaller RCCC reduction and worse performance compared to MC3KD.

| Method | CoLA (Matt) | CoLA (RCCC) | RTE (Acc) | RTE (RCCC) | MRPC (F1/Acc) | MRPC (RCCC) | STS-B (Pear/Spear) | STS-B (RCCC) | SST-2 (Acc) | SST-2 (RCCC) |
|---|---|---|---|---|---|---|---|---|---|---|
| only selection | 16.25 | 32.35 | 60.29 | 80.25 | 88.42/82.60 | 85.45 | 82.63/81.92 | 75.60 | 88.07 | 0.56 |
| only MC3 | 18.14 | 24.87 | 59.57 | 80.35 | 88.93/83.82 | 82.22 | 81.90/81.50 | 79.68 | 88.53 | **0.45** |
| MC3KD | **20.05** | **15.32** | **61.73** | **78.41** | **92.28/88.97** | **70.75** | **83.71/83.50** | **64.11** | **88.53** | 0.48 |

| Method | QNLI (Acc) | QNLI (RCCC) | QQP (F1/Acc) | QQP (RCCC) | MNLI-m (Acc) | MNLI-m (RCCC) | MNLI-mm (Acc) | MNLI-mm (RCCC) |
|---|---|---|---|---|---|---|---|---|
| only selection | 84.64 | 68.87 | 84.83/88.32 | 0.44 | 76.66 | 42.78 | 76.85 | 54.66 |
| only MC3 | 84.84 | 65.47 | 84.81/88.26 | 0.45 | 76.54 | 46.71 | 76.83 | 61.09 |
| MC3KD | **85.10** | **62.28** | **85.09/88.48** | **0.39** | **76.85** | **35.85** | **77.35** | **50.54** |

Table 5: Ablation Study: Distillation performance (%) and RCCC (%) of different partial MC3KD and full MC3KD.

| Teacher-Student (Method) | RTE (Acc) | RTE (RCCC) | SST-2 (Acc) | SST-2 (RCCC) | MRPC (F1/Acc) | MRPC (RCCC) | CoLA (Matt) | CoLA (RCCC) | STS-B (Pear/Spear) | STS-B (RCCC) |
|---|---|---|---|---|---|---|---|---|---|---|
| $XLNet_6$ (teacher) (Yang et al., 2019) | 66.07 | - | 91.86 | - | 97.32/96.32 | - | 34.86 | - | 84.91/85.04 | - |
| $XLNet_{12}$ (teacher) (Yang et al., 2019) | 66.43 | - | 94.50 | - | 96.34/94.85 | - | 42.90 | - | 85.32/85.31 | - |
| $XLNet_6$-$XLNet_3$ (PKD) | 53.07 | 93.72 | 89.56 | 94.94 | 88.46/82.35 | 94.01 | 15.72 | 90.37 | 82.06/82.69 | 93.77 |
| $XLNet_6$-$XLNet_3$ (MC3KD) | 57.04 | 93.34 | 90.14 | 93.76 | 89.40/84.07 | 93.73 | 16.74 | 88.97 | 82.15/83.05 | 93.36 |
| $XLNet_{12}$-$XLNet_3$ (MC3KD) | **58.12** | **92.74** | **90.71** | **83.67** | **90.57/86.27** | **93.00** | **17.82** | **85.28** | **82.30/82.84** | **93.15** |
| $Electra_6$ (teacher) (Clark et al., 2020) | 56.68 | - | 88.88 | - | 93.52/90.93 | - | 33.66 | - | 76.70/77.26 | - |
| $Electra_{12}$ (teacher) (Clark et al., 2020) | 79.78 | - | 95.30 | - | 99.11/98.78 | - | 67.63 | - | 88.48/88.63 | - |
| $Electra_6$-$Electra_3$ (PKD) | 53.79 | 99.53 | 87.16 | 3.73 | 83.22/74.51 | 98.39 | 12.90 | 86.76 | 75.01/75.49 | 8.67 |
| $Electra_6$-$Electra_3$ (MC3KD) | 55.23 | 97.50 | 87.73 | 2.02 | 83.78/75.74 | 97.34 | 13.08 | 31.61 | 75.85/76.06 | 5.40 |
| $Electra_{12}$-$Electra_3$ (MC3KD) | **56.68** | **96.61** | **87.96** | **0.58** | **84.21/77.21** | **96.57** | **14.65** | **23.48** | **75.97/76.53** | **2.03** |
| $Albert_6$ (teacher) (Lan, 2019) | 67.51 | - | 91.63 | - | 98.76/98.28 | - | 47.46 | - | 87.30/87.28 | - |
| $Albert_{12}$ (teacher) (Lan, 2019) | 77.26 | - | 92.32 | - | 98.59/98.04 | - | 58.12 | - | 89.01/88.85 | - |
| $Albert_6$-$Albert_3$ (PKD) | 63.18 | 92.92 | 89.22 | 77.88 | 96.00/94.36 | 91.68 | 31.19 | 26.79 | 84.60/84.63 | 86.36 |
| $Albert_6$-$Albert_3$ (MC3KD) | 64.26 | 92.09 | 89.45 | 54.91 | 97.18/96.08 | 90.88 | 33.13 | 25.24 | 84.88/85.09 | 81.82 |
| $Albert_{12}$-$Albert_3$ (MC3KD) | **65.70** | **90.34** | **89.79** | **14.26** | **98.21/97.55** | **88.22** | **35.89** | **8.10** | **85.23/85.61** | **80.12** |
| $Deberta_6$ (teacher) (He et al., 2021) | 68.23 | - | 93.58 | - | 98.40/97.79 | - | 53.35 | - | 87.14/87.10 | - |
| $Deberta_{12}$ (teacher) (He et al., 2021) | 57.76 | - | 94.95 | - | 96.93/95.83 | - | 55.76 | - | 86.93/86.88 | - |
| $Deberta_6$-$Deberta_3$ (PKD) | 55.60 | 17.21 | 91.63 | 25.71 | 87.88/81.62 | 15.45 | 32.80 | 11.33 | 78.52/79.44 | 20.81 |
| $Deberta_6$-$Deberta_3$ (MC3KD) | 57.04 | 17.06 | **91.86** | **25.29** | 88.71/82.84 | 15.38 | 34.77 | 11.32 | 79.48/80.00 | 20.25 |
| $Deberta_{12}$-$Deberta_3$ (MC3KD) | **58.84** | **16.67** | 91.74 | 25.62 | **90.42/86.03** | **15.33** | **34.89** | **11.17** | **80.24/80.68** | **19.86** |

Table 6: Distillation performance (%) and RCCC (%) of BERT-relevant and other models.

### 6.4 Ablation Study

Our method comprises two key components: teacher layer selection in Section 5.1 and maximizing canonical correlation coefficients (MC3) in Section 5.2. We also conduct ablation experiments comparing partial MC3KD variants (denoted as 'only selection' and 'only MC3') which each only contains one single component in MC3KD with the full MC3KD containing both components. In 'only selection', we choose proper intermediate layers as teacher layer using the criteria we propose in Section 5.1, while vanilla KD loss is applied as distillation loss. In 'only MC3', teacher layers are selected by PKD-skip strategy and distillation loss is based on MC3 we propose in Section 5.2.

As shown in Table 5, only the complete MC3KD yields optimal distillation performance and maximizes the linear differences among the student model's intermediate layers, underscoring the indispensability of both components.

### 6.5 Generalizability and Transferability

We also evaluate MC3KD on additional models. Some are BERT variants (Clark et al., 2020; Lan, 2019; He et al., 2021) and others are unrelated to BERT (Yang et al., 2019). As shown in Table 6, MC3KD not only enhances the effectiveness of BERT compression but also enables larger teacher models to achieve superior distillation performance. The experimental results further demonstrate that better distillation performance is also reflected in greater linear diversity among the intermediate layers of student models.

These results show that our method generalizes to other models and provides a solution for capacity mismatch in classification tasks with few classes, demonstrating strong generalizability and transferability. We hope our approach can provide insights for related research, particularly in the compression of other language models and feature-based knowledge distillation methods.

## 7 Conclusion

In this paper, we address the challenge of larger BERT models failing to achieve better distillation performance. Our analysis reveals a strong correlation between enhanced generalization and increased linear differences among intermediate layers. We find existing methods like PKD fail to effectively amplify these differences, limiting the distillation potential of larger models. To overcome this, we propose MC3KD, which maximizes Canonical Correlation Coefficients (CCC) between intermediate layers. Experimental results demonstrate that MC3KD successfully increases CCC and enables larger BERT models to achieve superior distillation performance.

## Limitations

Experiments show that MC3KD's training time is approximately 1.6 times that of the original method. A promising direction for future work is to develop more efficient alternatives.

Another recently popular metric for analyzing linear relationships is the Concordance Correlation Coefficient (Lawrence and Lin, 1989). Our work focuses on providing new insights into improving knowledge distillation for large BERT teachers. We leave exploring the application of this metric to increase linear differences among student model layers for future work.

## Acknowledgements

This work is supported by National Science and Technology Major Project (GrantNo. 2022ZD0114805). Professor De-Chuan Zhan is the corresponding author.

## References

TW Anderson. 1985. An introduction to multivariate statistical analysis. *Biometrics*, 41(3):815.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *International Conference on Learning Representations (ICLR)*.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zixiang Ding, Guoqing Jiang, Shuai Zhang, Lin Guo, and Wei Lin. 2024. How to trade off the quantity and capacity of teacher ensemble: Learning categorical distribution to stochastically employ a teacher for distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17915–17923.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. 2024. Revisit the essence of distilling knowledge through calibration. In *Forty-first International Conference on Machine Learning, ICML 2024*.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Jinyang Guo, Jiaheng Liu, Zining Wang, Yuqing Ma, Ruihao Gong, Ke Xu, and Xianglong Liu. 2023. Adaptive contrastive knowledge distillation for bert compression. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8941–8953.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Investigating learning dynamics of bert fine-tuning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 87–92.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Trevor Hastie. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *International Conference on Learning Representations (ICLR)*.

Yihui He, Xiangyu Zhang, and Jian Sun. 2017. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

H Hotelling. 1936. Relations between two sets of variates. *Biometrika*.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Berkay Köprü and Engin Erzin. 2020. Multimodal continuous emotion recognition using deep multitask learning with correlation loss. *arXiv preprint arXiv:2011.00876*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Wonbin Kweon, SeongKu Kang, and Hwanjo Yu. 2021. Bidirectional distillation for top-k recommender system. In *Proceedings of the Web Conference 2021*, pages 3861–3871.

Zhenzhong Lan. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Xin-Chun Li, Wen-Shu Fan, Shaoming Song, Yinchuan Li, Shao Yunfeng, De-Chuan Zhan, et al. 2022. Asymmetric temperature scaling makes larger networks teach well again. *Advances in neural information processing systems*, 35:3830–3842.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. 2019. A tensorized transformer for language modeling. *Advances in neural information processing systems*, 32.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. 2018. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. 2024. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15731–15740.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

Siqi Sun, Zhe Gan, Yu Cheng, Yuwei Fang, Shuohang Wang, and Jingjing Liu. 2020. Contrastive distillation on intermediate representations for language model compression. *arXiv preprint arXiv:2009.14167*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. 2022. Efficient knowledge distillation from model checkpoints. *Advances in Neural Information Processing Systems*, 35:607–619.

A Warstadt. 2019. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Yang Yang, Zhao-Yang Fu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. 2021a. Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport. *IEEE Trans. Knowl. Data Eng.*, 33(2):696–709.

Yang Yang, Jinyi Guo, Guangyu Li, Lanyu Li, Wenjie Li, and Jian Yang. 2024a. Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning. *Frontiers Comput. Sci.*, 18(3):181335.

Yang Yang, Nan Jiang, Yi Xu, and De-Chuan Zhan. 2024b. Robust semi-supervised learning by wisely leveraging open-set data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):8334–8347.

Yang Yang, Zhen-Qiang Sun, Hengshu Zhu, Yanjie Fu, Yuanchun Zhou, Hui Xiong, and Jian Yang. 2023a. Learning adaptive embedding considering incremental class. *IEEE Trans. Knowl. Data Eng.*, 35(3):2736–2749.

Yang Yang, Wenjuan Xi, Luping Zhou, and Jinhui Tang. 2024c. Rebalanced vision-language retrieval considering structure-aware distillation. *IEEE Trans. Image Process.*, 33:6881–6892.

Yang Yang, De-Chuan Zhan, Yi-Feng Wu, Zhi-Bin Liu, Hui Xiong, and Yuan Jiang. 2021b. Semi-supervised multi-modal clustering and classification with incomplete modalities. *IEEE Trans. Knowl. Data Eng.*, 33(2):682–695.

Yang Yang, Da-Wei Zhou, De-Chuan Zhan, Hui Xiong, Yuan Jiang, and Jian Yang. 2023b. Cost-effective incremental deep model: Matching model capacity with the least sampling. *IEEE Trans. Knowl. Data Eng.*, 35(4):3575–3588.

Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. 2024d. Vitkd: Feature-based knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1379–1388.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Shuoxi Zhang, Hanpeng Liu, Yuyi Wang, Kun He, Jun Lin, and Yang Zeng. 2025. Class discriminative knowledge distillation. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Ran Zhou, Yang Liu, Wei Xia, Yu Guo, Zhongwei Huang, Haitao Gan, and Aaron Fenster. 2024. Jocorank: Joint correlation learning with ranking similarity regularization for imbalanced fetal brain age regression. *Computers in Biology and Medicine*, 171:108111.

Wangchunshu Zhou and Canwen Xu. 2022. Bert learns to teach: Knowledge distillation with meta learning. In *Annual Meeting of the Association for Computational Linguistics*.

Zhi-Hua Zhou. 2016. Learnware: on the future of machine learning. *Frontiers Comput. Sci.*, 10(4):589–590.

## A Setting of Hyperparameters

We set the number of hidden units in the final softmax layer to 768, the batch size to 32, and the number of training epochs to 4 across all experiments. We adopt AdamW (Loshchilov, 2017) as optimizer. Following prior works (Sun et al., 2019; Guo et al., 2023), we conduct a hyperparameter search over the student learning rate from $\{2e^{-5}, 3e^{-5}, 5e^{-5}\}$, hyperparameter $\alpha$ from $\{0.2, 0.5, 0.6, 0.7\}$ and hyperparameter $\beta$ from $\{0.1, 1.0, 10.0\}$. The temperature parameter is fixed at $T = 20$ since it primarily influences the output distribution, while our method focuses on the linear relationships between intermediate layers and remains largely unaffected by temperature. Thus, tuning the temperature is unnecessary. The remaining hyperparameters are identical to those used in pre-training the teacher network.

## B Supplementary Experimental Results

### B.1 Variation of CCC across epochs

To further illustrate the relationship between distillation performance and the linear differences among the student's intermediate layers, we track the student's generalization performance and SVCCA matrix (represented by RCCC in figures) at the end of each epoch during distillation and visualize the results as line charts. In each subplot of Figure 4, using $\text{BERT}_{24}$ as the teacher and $\text{BERT}_6$ as the student, we observe that as distillation progresses, the student model's performance consistently improves while the RCCC value decreases. This indicates that across different datasets, lower RCCC values correlate with improved distillation performance. Figure 5 further confirms that this trend holds consistently across different teacher-student pairs on the given QQP dataset. These findings reinforce our conclusion: increasing the linear differences among the student model's intermediate layers enhances its generalization ability, thereby improving distillation performance.

### B.2 More Comparative Data of MC3KD

Table 7 serves as a supplement to Table 3, it compares the distillation performance and RCCC values of MC3KD against PKD and KD across additional teacher–student pairs.

We perform comparisons with another classic method DynaKD (Ding et al., 2024). Table 8 shows that our method not only achieves a lower RCCC, but also outperforms DynaKD in distillation performance.

We have also tested our method on GPT style models (Radford et al., 2019) on Wikitext-2 test dataset. In Table 9, "PPL" stands for perplexity, quantifying the model's uncertainty in predicting the next token within long text sequences—lower perplexity indicates stronger performance. Table 9 demonstrates that, vertically, MC3KD consistently outperforms KD, and, horizontally, it mitigates performance degradation when distilling from increasingly larger GPT teachers.

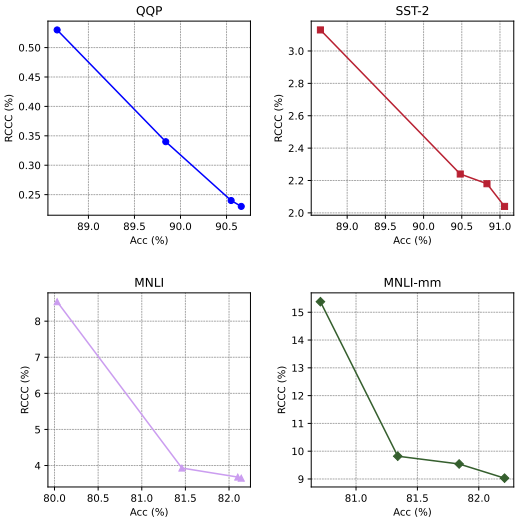


Figure 4: The relationship between KD performance and the linear differences among the student's intermediate layers across different datasets, given a specific teacher and student model.

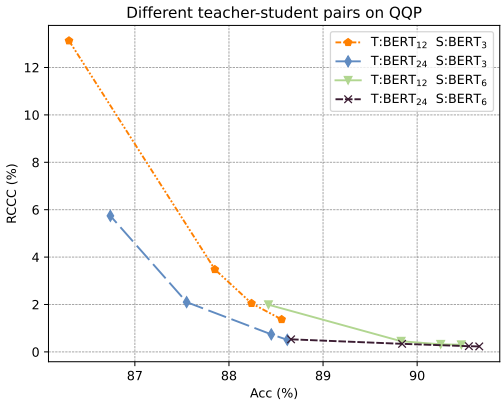


Figure 5: The relationship between KD performance and the linear differences among the student's intermediate layers for different teacher-student pairs on a given dataset. In the legend, T represents the teacher model, and S represents the student model.

| Method | CoLA (Matt) | CoLA (RCCC) | RTE (Acc) | RTE (RCCC) | MRPC (F1/Acc) | MRPC (RCCC) | STS-B (Pear/Spear) | STS-B (RCCC) |
|---|---|---|---|---|---|---|---|---|
| | | | Teacher: $BERT_{12}$ | | Student: $BERT_3$ | | | |
| KD | 16.29 | 21.51 | 57.40 | 82.15 | 86.71/79.41 | 86.10 | 80.99/80.85 | 83.27 |
| PKD | 16.72 | 19.06 | 59.21 | 80.43 | 90.46/85.78 | 84.99 | 82.48/82.11 | 79.78 |
| MC3KD | **16.97** | **17.40** | **60.29** | **79.71** | **91.57/87.50** | **80.82** | **83.21/82.72** | **68.81** |
| | | | Teacher: $BERT_{24}$ | | Student: $BERT_3$ | | | |
| KD | 14.89 | 38.78 | 57.76 | 81.67 | 88.05/82.11 | 85.73 | 82.38/82.30 | 70.31 |
| PKD | 15.28 | 34.37 | 58.84 | 81.26 | 86.00/79.41 | 86.54 | 81.43/80.69 | 84.29 |
| MC3KD | **20.05** | **15.32** | **61.73** | **78.41** | **92.28/88.97** | **70.75** | **83.71/83.50** | **64.11** |
| | | | Teacher: $BERT_{12}$ | | Student: $BERT_6$ | | | |
| KD | 42.97 | 7.84 | 65.70 | 70.60 | 97.02/95.83 | 75.68 | 88.18/87.85 | 78.48 |
| PKD | 42.50 | 8.98 | 65.34 | 70.85 | 98.05/97.30 | 74.23 | 87.91/87.61 | 79.46 |
| MC3KD | **44.12** | **6.47** | **66.43** | **70.25** | **98.58/98.04** | **73.81** | **88.26/87.97** | **77.84** |
| | | | Teacher: $BERT_{24}$ | | Student: $BERT_6$ | | | |
| KD | 42.35 | 9.27 | 64.98 | 70.74 | 94.55/92.40 | 75.76 | 88.01/87.76 | 78.83 |
| PKD | 42.12 | 10.04 | 64.62 | 71.25 | 97.15/96.08 | 74.97 | 88.07/87.85 | 79.24 |
| MC3KD | **44.81** | **6.14** | **67.15** | **69.71** | **99.64/99.51** | **69.87** | **88.37/87.98** | **73.01** |

Table 7: The relationship between distillation performance (%) and RCCC (%) on various teacher-student pairs.

| Method | CoLA (Matt) | CoLA (RCCC) | RTE (Acc) | RTE (RCCC) | MRPC (F1/Acc) | MRPC (RCCC) | STS-B (Pear/Spear) | STS-B (RCCC) | SST-2 (Acc) | SST-2 (RCCC) |
|---|---|---|---|---|---|---|---|---|---|---|
| DynaKD | 42.39 | 10.02 | 66.43 | 70.07 | 97.01/95.83 | 74.87 | 88.12/87.84 | 78.86 | 90.94 | 1.33 |
| MC3KD | **44.81** | **6.14** | **67.15** | **69.71** | **99.64/99.51** | **69.87** | **88.37/87.98** | **73.01** | **91.40** | **0.46** |

| Method | QNLI (Acc) | QNLI (RCCC) | QQP (F1/Acc) | QQP (RCCC) | MNLI-m (Acc) | MNLI-m (RCCC) | MNLI-mm (Acc) | MNLI-mm (RCCC) |
|---|---|---|---|---|---|---|---|---|
| DynaKD | 88.36 | 73.87 | **87.24/90.49** | **0.33** | 81.80 | 13.97 | 82.04 | 13.76 |
| MC3KD | **88.94** | **71.52** | 87.17/90.43 | 0.53 | **82.14** | **3.65** | **82.21** | **9.03** |

Table 8: Distillation performance (%) and RCCC (%) of MC3KD and DynaKD.

| Method | Teacher | Student | PPL | Teacher | Student | PPL |
|---|---|---|---|---|---|---|
| KD | GPT2-medium | GPT2-small | 28.50 | GPT2-large | GPT2-small | 29.24 |
| MC3KD | GPT2-medium | GPT2-small | 23.71 | GPT2-large | GPT2-small | 23.57 |

Table 9: Distillation performance (%) on sequence generation tasks.

## C Supplementary Figures

The following figures provide supplementary information to support the main text.

Figure 6 provides a more detailed illustration of Figure 2 across additional tasks, showing that the larger the teacher model, the more pronounced the linear differences among its intermediate layers.

Figure 7 is a large-scale visualization that clearly reflects the specific values of linear differences across the teacher model's intermediate layers. It echoes the discussion in Section 5.1, highlighting that the PKD method may miss intermediate layers well-suited for teaching.

Figure 8 shows that the below-diagonal elements in the SVCCA matrix—indicating linear diversity between adjacent layers—tend to have smaller values as model size grows. This suggests larger models are more likely to contain intermediate layers that capture critical information, reinforcing the potential of larger teacher models for better distillation, as discussed in Section 4.

Figure 9 depicts the observation in Section 3.2: if one element in an SVCCA matrix is greater than its counterpart in another, this pattern typically holds across most corresponding elements. Thus, comparing SVCCA matrices can be effectively reduced to comparing their RCCC values.

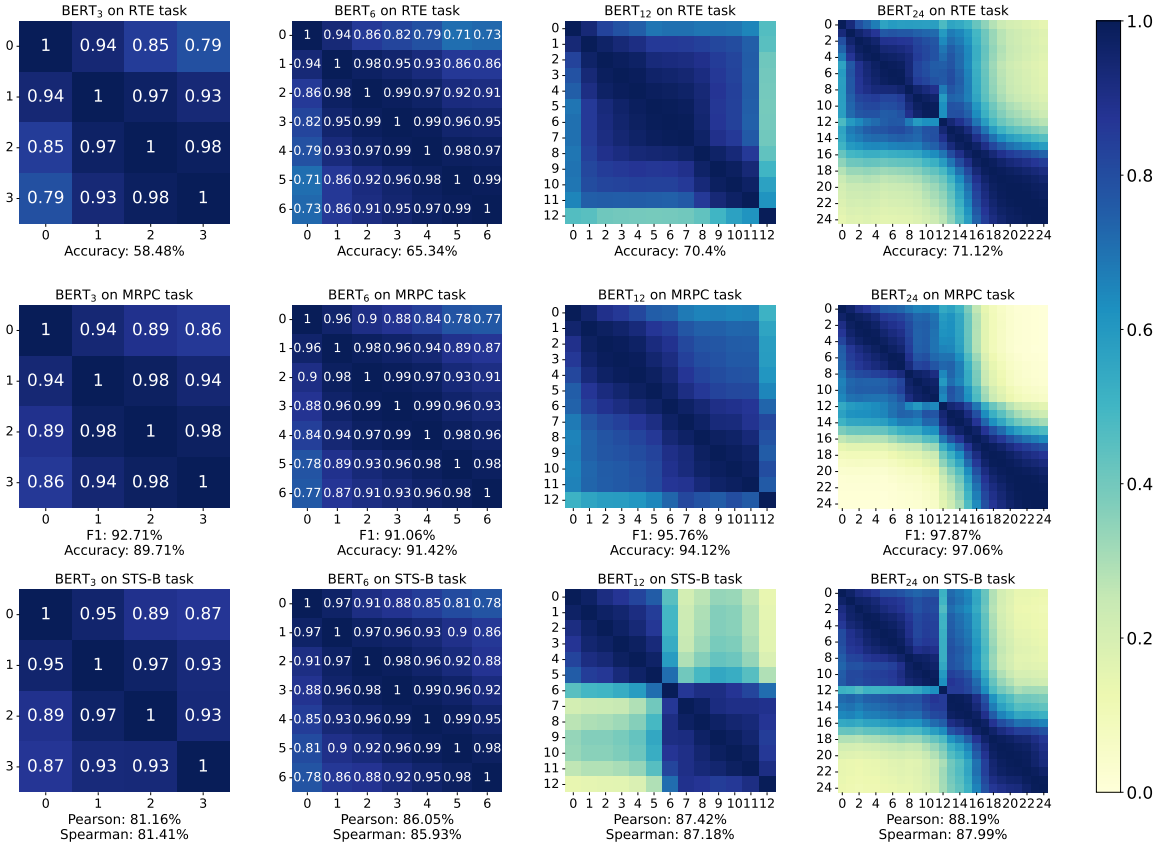


Figure 6: Heatmaps of SVCCA for BERT pre-trained models of varying sizes on several tasks.

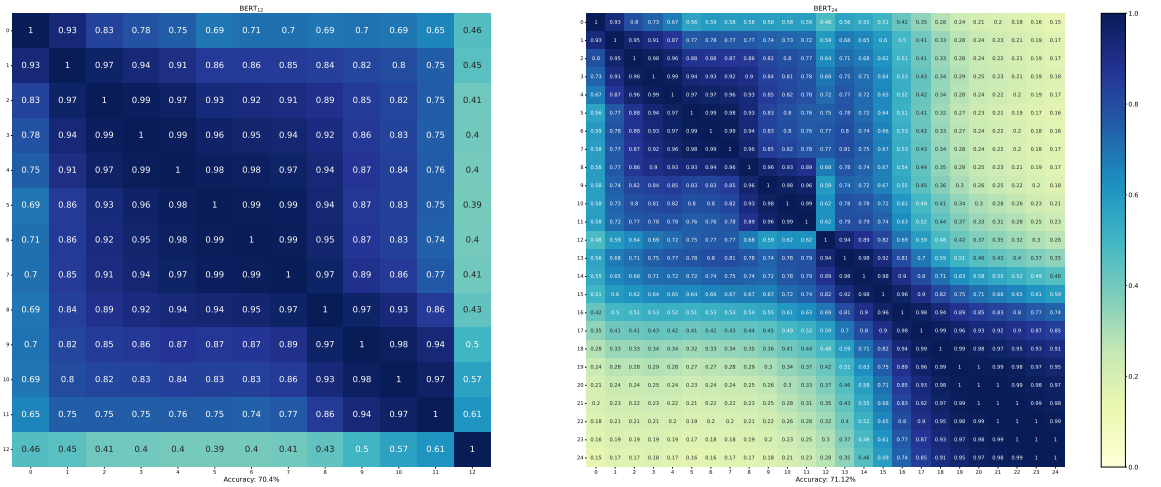

Figure 7: Heatmaps of RTE

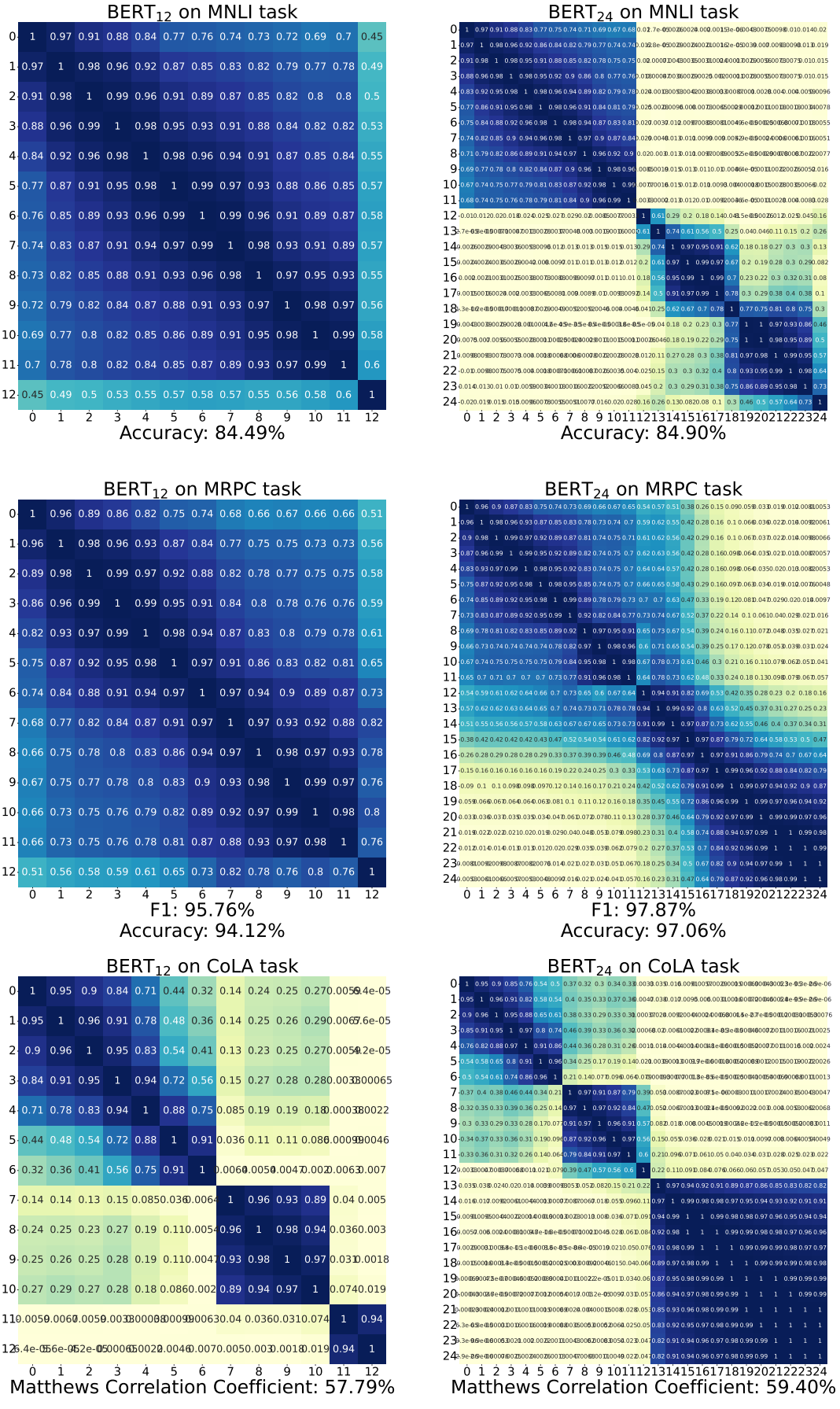


Figure 8: Zoomed-in view of the SVCCA matrix elements for the large BERT model.

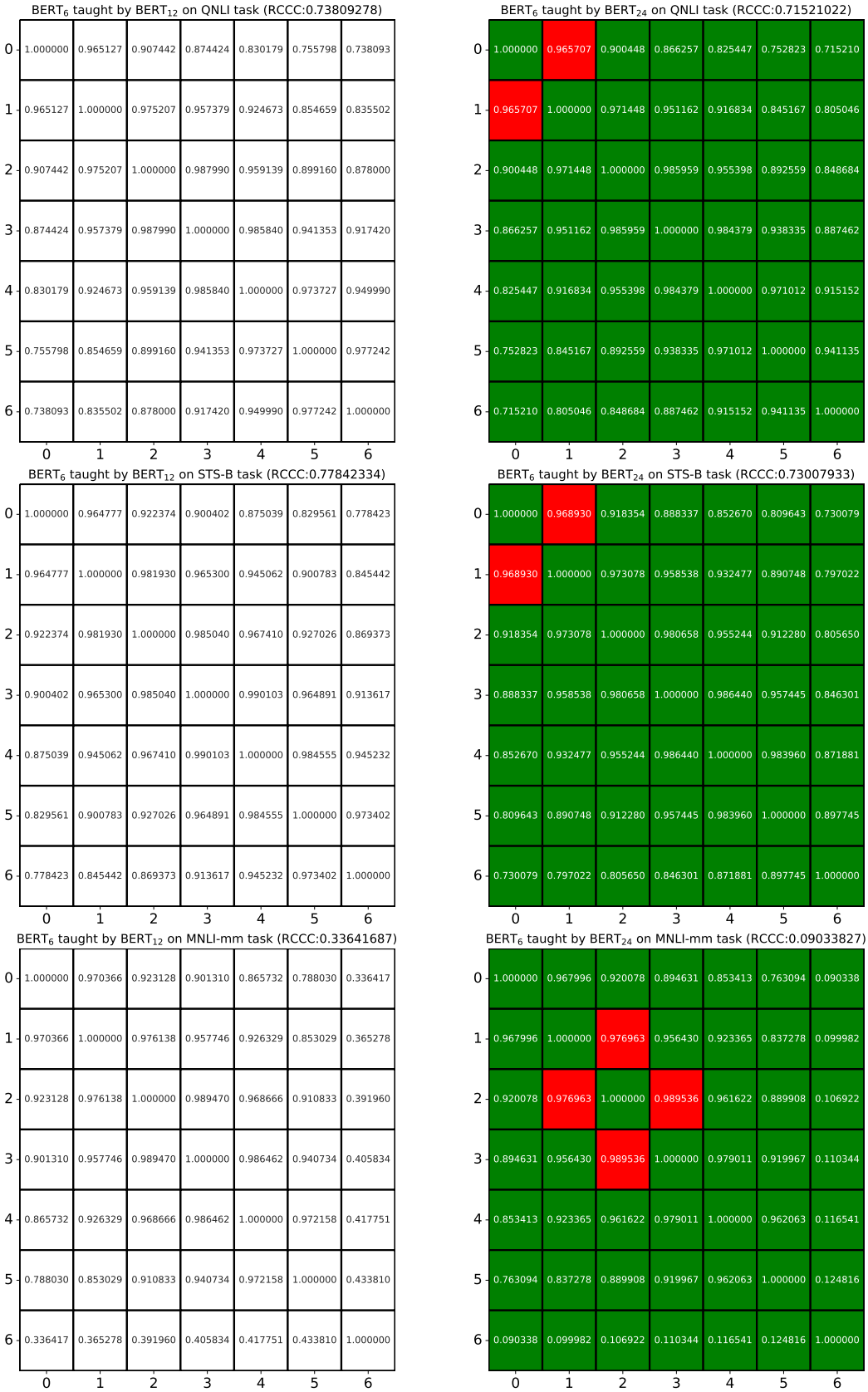


Figure 9: Comparison of SVCCA matrices for student models distilled by different teacher models. Each row shows the benchmark on the left, with the corresponding element on the right colored red if it is larger than the left matrix, otherwise green. In each row, if the RCCC element in the upper-right corner of the right matrix is smaller than the corresponding element in the left matrix, then almost all other elements are smaller than left as well. As a result, the comparison of RCCC element sizes reflects the overall size comparison of the SVCCA matrices.