EAI lab 5

Knowledge Distillation

數據所 RE6121011 徐仁瓏

1. Comparison of student models w/ & w/o KD

Model	Knowledge Distillation	Test Loss ↓	Test Accuracy(%) ↑
Teacher	-	0.503	88.300
Student	-	0.475	87.780
Student	Response-based	0.303	89.290
Student	Feature-based	0.213	88.330

2. How you choose the Temperature and alpha in response-based KD

為了選擇最佳的參數組合，我進行了一系列實驗，以尋找較佳的配置：

首先，在 **Temperature** 的選擇上，理論上較高的 Temperature 會使教師模型的輸出機率更加平滑，從而讓學生模型能學習到更多細緻的信息，而不僅限於正確類別的標籤。實驗結果顯示，當 **Alpha** 固定為 0.5 時，**Temperature = 4** 的表現最佳。

其次，在 **Alpha** 的選擇上，Alpha 用於平衡原始硬標籤損失與蒸餾損失的比重。較高的 Alpha 更偏重於教師模型的軟標籤。在固定 **Temperature = 4** 的情況下，我比較了 **Alpha = 0.4** 和 **Alpha = 0.6**，結果發現 **Alpha = 0.4** 的效果更佳。

最終選定參數配置為 **Temperature = 4** 和 **Alpha = 0.4**，此時 Test Accuracy 達到 **89.29%**，相較於未使用知識蒸餾（KD）方法的基準提升了 **1.51%**。

Model	Temperature	Alpha	Test Loss ↓	Test Accuracy(%) ↑
Student	3	0.5	0.263	88.320
Student	4	0.5	0.283	88.820
Student	5	0.5	0.312	87.870
Student	4	0.4	0.303	**89.290**
Student	4	0.6	0.260	87.950

3. How you extract features from the choosing intermediate layers

為了從 ResNet 中提取特徵，我針對其四個 Block 的結構進行實驗。具體來說，我保存了老師模型和學生模型通過這四個 Block 後的特徵，目的是讓學生模型學習到老師模型的特徵分佈，進而提升學生模型的表現。

在參數設置方面，我首先針對 **Alpha** 進行實驗。Alpha 用於平衡 MMD 損失和 Cross Entropy 損失的權重。實驗結果顯示，當 **Alpha = 0.5** 時，模型性能相對較佳，但仍未能超越未使用 KD 方法的學生模型。為此，我進一步嘗試在不同特徵之間分配不同的權重，讓學生模型更注重學習老師模型靠近輸出的後層特徵。這是因為後層特徵與輸出結果的相關性更高。

使用這種配置後，模型的 Test Accuracy 達到 **88.33%**，成功超越了未使用 KD 方法的學生模型。

Model	Weight	Alpha	Test Loss ↓	Test Accuracy(%) ↑
Student	[0.25, 0.25, 0.25, 0.25]	0.4	0.210	86.980
Student	[0.25, 0.25, 0.25, 0.25]	0.5	0.235	87.720
Student	[0.25, 0.25, 0.25, 0.25]	0.6	0.277	87.460
Student	[0.1, 0.2, 0.3, 0.4]	0.5	0.213	**88.330**

4. How you design the loss of your response-based and feature-based KD

- **Response-based KD**

我設計的 response-based 知識蒸餾損失函數結合了教師網絡和學生網絡之間的軟目標和硬目標。軟目標使用 **KL 散度 (Kullback-Leibler Divergence)** 作為損失，通過溫度 T 平滑教師的輸出，讓學生學習教師在不同類別上的不確定性。硬目標則使用 **交叉熵損失 (CrossEntropy Loss)** 來確保學生學到正確的分類能力。

兩者結合的權重由參數 alpha 控制，alpha 越高，學生越依賴教師的軟目標知識，1 - alpha 則控制硬目標的比例。這樣的設計能幫助學生網絡在分類性能和知識蒸餾之間取得平衡，提升其泛化能力。

- **Feature-based KD**

在設計 feature-based 知識蒸餾時，我採用了**最大均值差異 (MMD)** 來匹配教師和學生網絡之間的中間層特徵分佈。特徵蒸餾的目標是讓學生網絡學習到教師網絡在不同層次上提取到的

深層特徵表示。為此，我計算了教師和學生在每一層特徵圖之間的 MMD，這有助於減少學生特徵與教師特徵之間的分佈差異。

為了計算 MMD 損失，我使用了**高斯核 (Gaussian Kernel)** 來衡量學生和教師特徵之間的相似性。為了平衡不同層次特徵對總損失的影響，我引入了 layer_weights，讓每一層的特徵根據其對最終結果的重要性賦予不同的權重。此外，我還引入了參數 alpha 來控制 MMD 損失和交叉熵損失的比例，從而確保學生模型在分類任務和特徵匹配之間取得良好的平衡。

這樣的設計使得學生模型不僅能學到教師網絡的輸出分類結果，還能學習到教師網絡在中間層次上的特徵表示，從而提升學生模型的泛化能力和最終性能。

5. Problems you met & how you solved them

在進行 Feature-based KD 蒸餾時，學生模型的表現一直無法超越未使用 KD 方法的學生模型。這可能是因為學生模型在初始訓練時已經達到了較好的效果，因此使用 KD 的提升空間有限。為了解決這個問題，我嘗試在各個層之間分配不同的權重，讓淺層特徵的權重較小，而深層特徵的權重較大。這是基於深層特徵與最終輸出結果相關性更高的假設，目的是讓學生模型在深層特徵的學習上更接近老師模型。透過這種調整，最終使得使用 Feature-based KD 的學生模型表現超越了未使用 KD 的學生模型。

此外，在蒸餾過程中，我一開始不確定在使用 MMD 損失後是否還需要加入 Cross Entropy 損失。為此，我先進行了僅使用 MMD 損失的實驗，結果顯示學生模型的效果非常差。後來我加入了 Cross Entropy 損失，並通過 **alpha** 參數平衡 MMD 和 Cross Entropy 損失的權重，這才使學生模型的表現得到了顯著提升。

對於損失函數輸出結果的損失感到困惑，從實驗結果中通常準確度越高的結果，損失應該會越低，但是實驗結果並不如預期。我認為可能和不同損失函數之間是無法比較損失這件事情有關，損失函數是一個相對而非絕對的指標，因此設定在同一個損失函數底下的損失之間的比較才會具有意義，在不同損失函數之間的損失可能是無法去做比較的。