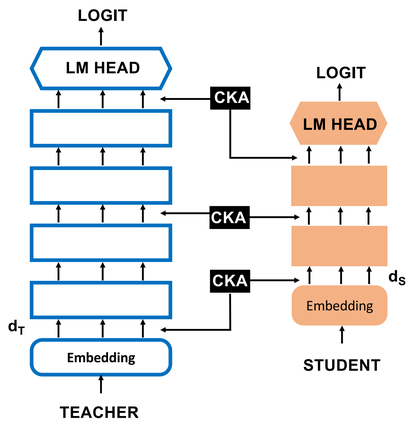
Improving Language Model Distillation through Hidden State Matching

Sayantan Dasgupta & Trevor Cohn
University of Melbourne

Hidden State Matching

Goal: To match the alternating hidden states between the teacher(T) and the student(S) with different dimensions

$$d_T \neq d_S$$



Possible losses between the hidden states

1. Linear Loss:

$$\mathcal{L}_H = \text{MSE}(H_T, AH_S)$$

2. CKA Loss:

$$\mathcal{L}_H = 1 - \frac{\|\Sigma_{TS}\|_F}{\sqrt{\|\Sigma_{TT}\|_F}\sqrt{\|\Sigma_{SS}\|_F}}$$

$$\Sigma_{SS} = \frac{1}{N-1}\tilde{H}_S^\top \tilde{H}_S$$

$$\Sigma_{TT} = \frac{1}{N-1}\tilde{H}_T^\top \tilde{H}_T$$

$$\Sigma_{TS} = \frac{1}{N-1}\tilde{H}_T^\top \tilde{H}_S$$

$$\tilde{H}_{S(T)} = H_{S(T)} - \frac{1}{N}\sum_{i=1}^{N} h_{S(T)_i}$$

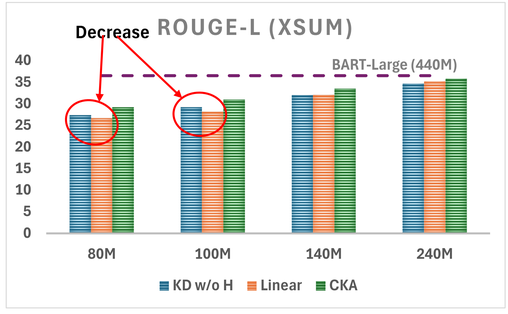
Final loss for KD:

$$\mathcal{L}_{CLM}(S) + \mathcal{L}_{KLD}(S,T) + \mathcal{L}_H(S,T)$$

Consistent across Versatile Tasks

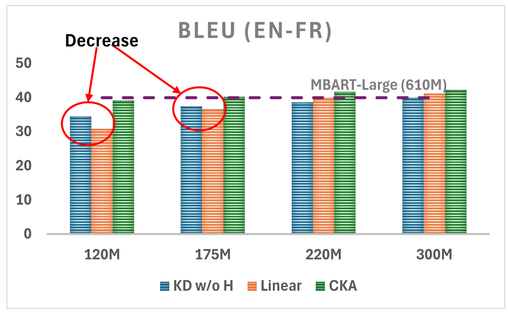
Summarization:

- Distilled BART-Large (440M) fine-tuned on XSUM and CNN-Dailymail dataset for abstractive summarization
- Student size varies from 80M (5.5×) to 240M (1.8×)
- For smaller students (80M and 100M), Linear loss degrades the performance
- Distillation performed from scratch on task-specific datasets



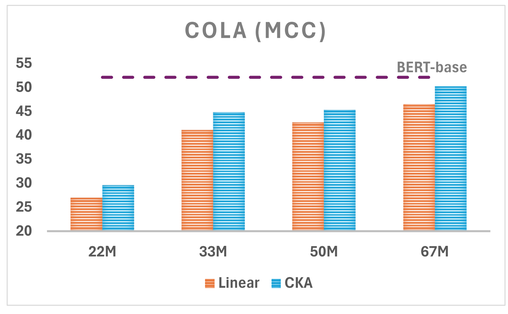
Machine Translation:

- Distilled MBART-Large (610M) fine-tuned on EN-RO dataset from WMT'16 and EN-FR dataset from IWSLT'17
- Student size varies from 120M (5×) to 300M (2×)
- For smaller students (120M and 300M), Linear loss degrades the performance
- All students undergo pretrained distillation on MC4 before using the task-specific datasets



Classification:

- Distilled BERT-base (110M) on C4
- Students vary in size from 22M (5×) to 67M (1.6×)
- Fine-tuned the distilled students on GLUE tasks
- CKA produces a better average GLUE score
- The Mathew Correlation Coefficient on COLA is shown on the second plot, which is the most difficult of the GLUE tasks



High Compression Ratio (∼ 20×)

- Remains stable for the distillation of Flan-T5 3B with high compression ratios up to ∼ 20×
- Distilled students produce high BLEU scores for English to Spanish translation on WMT'13
- Linear loss fails to converge for Flan-T5 distillation

BLEU scores on EN → ES (Flan T5 3B: 28.0)

Flan T5-3B →	**145M (20×)**	**250M (12×)**	**425M (7×)**	**780M (4×)**
KD w/o H	25.2	27.3	29.4	30.6
CKA	27.2	29.3	30.8	31.8

Project Github: