PalenkaLM: Deeply Supervised Baby-Llama

Jan Stipl, Kasym Manatayev, Leah Le
CMSC 395 Natural Language Processing, University of Richmond, VA 23173

Overview

Our approach builds upon a **Baby-Llama** baselines and further improves it using Deeply Supervised Knowledge Distillation (DSKD). In addition to vanilla KD, we add an average MSE loss between student's hidden layers' output and teachers' hidden layers' outputs to the loss function. We achieved outperformance on several tasks comparing to previous BabyLM winners.

Baselines / Background

Baselines: GPT2-44M, GPT2-705M, GPT2-small-97M, Llama-60M, Llama-360M

Distilled: DistilledGPT-44M

- Teachers: GPT2-44M, Llama-60M
- Student: GPT2-44M

BabyLlama-1-58M

- Teachers: GPT2-705M, Llama-360M
- Student: Llama-60M

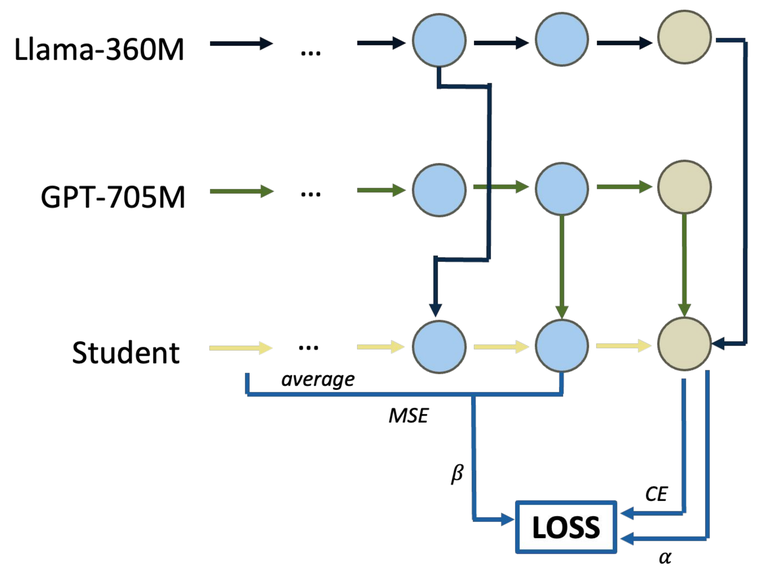
We created **PalenkaLlama1-58M**

- Teachers: GPT2-705M, Llama-360M
- Student: Llama-60M

Loss function

$$\text{Cross-Entropy}_{\text{(student)}} + \alpha * \text{loss\_logits}_{\text{(teachers' last layers, student's last layer)}} + \beta * \text{hidden\_loss}_{\text{(teachers' hidden layers projected to student's hidden layer)}}$$

Methodology / Architecture



Grid Search (Hyperparameter Tuning)

Logit Loss Weight (α)	Hidden Loss Weight (β)	Validation Loss
0.1	0.1	6.24
0.1	0.3	7.43
0.1	**0.5**	**8.62**
…	…	…
0.9	0.9	27.99

Results

Model Name	BLiMP (supplement)	BLiMP (filtered)	EWoK (filtered)
BabyLlama1-58M-strict	0.581 ± 0.0054	0.676 ± 0.0016	0.501 ± 0.0057
DistilledGPT-44M-strict	0.588 ± 0.0053	0.658 ± 0.0016	0.500 ± 0.0057
GPT2-44M-strict	0.591 ± 0.0056	0.633 ± 0.0017	0.501 ± 0.0057
GPT2-705M-strict	0.574 ± 0.0058	0.657 ± 0.0017	0.501 ± 0.0057
GPT2-small-97M-strict	0.563 ± 0.0058	0.662 ± 0.0017	**0.506 ± 0.0057**
Llama-360M-strict	**0.610 ± 0.0055**	0.654 ± 0.0017	0.499 ± 0.0057
Llama-60M-strict	0.567 ± 0.0057	0.637 ± 0.0017	0.499 ± 0.0057
PalenkaLlama1-58M-strict -L0.1-H0.1	0.596 ± 0.0054	0.694 ± 0.0016	0.500 ± 0.0057
PalenkaLlama1-58M-strict -L0.1-H0.3	0.599 ± 0.0053	0.693 ± 0.0016	0.499 ± 0.0057
PalenkaLlama1-58M-strict -L0.1-H0.5	0.606 ± 0.0050	**0.694 ± 0.0016**	0.503 ± 0.0057

References

- *Knowledge Distillation with Deep Supervision*
- *Distilling the Knowledge in a Neural Network*
- *Baby Llama: Knowledge Distillation from an Ensemble of Teachers Trained on a Small Dataset with No Performance Penalty*
- *Teaching Tiny Minds: Exploring Methods to Enhance Knowledge Distillation for Small Language Models*
- *Language models are unsupervised multitask learners*