Teaching Tiny Minds: Exploring Methods to Enhance Knowledge Distillation for Small Language Models

Hong Meng Yam
Stanford University
hongmeng@stanford.edu

Nathan Paek
Stanford University
nathanjp@stanford.edu

Abstract

In this paper, we build off of the success of the previous BabyLM challenge winner's model, BabyLlama, to explore various methods of enhancing knowledge distillation for small language models. Our main focus is on investigating how small a language model can be while still maintaining competitive performance. We experiment with three main approaches: (1) DistilledGPT-44M, which uses smaller teacher models and a more compact student model compared to BabyLlama; (2) ContrastiveLlama-58M, which incorporates contrastive loss into the knowledge distillation process; and (3) MaskedAdversarialLlama-58M, incorporates adversarial loss into the knowledge distillation process. Using the 10M-word dataset from the BabyLM challenge's strict-small track, we evaluate our models on the BLiMP, EWoK, and GLUE benchmarks. Our results show that effective knowledge distillation can still be achieved with significantly smaller teacher and student models. In particular, our model DistilledGPT-44M is able to achieve better performance than one of last year's winning entries, LTG-BERT, while achieving similar performance but cutting training time by around 70% and parameters by around 25% compared to the other winning entry, BabyLlama.

1 Introduction

Since 2017, transformers have been everywhere in NLP (Vaswani, 2017). Their non-autoregressive nature allows for high parallelization, leading to unprecedented scalability. In recent years, a number of models with trillions of parameters have emerged, such as Google's Switch Transformer (1.6 trillion) and Huawei's PanGu-Σ (1.1 trillion). Models like these or models in the billions demand enormous computational resources and vast swathes of training data. They consume substantial energy, raising concerns about their environmental impact (Bender et al., 2021). The costs of continued scaling are also increasingly prohibitive, highlighting the need for more sample-efficient model architectures.

The 2024 BabyLM challenge (Choshen et al., 2024), by limiting the amount of training data available, in some ways aims to address the computational concerns around large models. However, the scope of the contest focuses more on limiting the amount of training data rather than limiting parameter size or compute; participants have the freedom to use models as large as they want. But last year, the winners of BabyLM (Timiryasov and Tastet, 2023) demonstrated with their model BabyLlama that a small model can outperform models close to an order of magnitude larger on NLP tasks. Model parameter efficiency does not necessarily mean worse results; in fact, in some cases it means better results.

BabyLlama used knowledge distillation and ensemble learning to distill knowledge from two teacher models - GPT2-705M and Llama-360M - to a smaller Llama-58M student model (Hinton et al., 2015). As a model compression technique, knowledge distillation (KD) has several advantages: it only requires access to the teacher model's output logits (not its weights), and it is also model agnostic.

Building on BabyLlama's success, we aim to demonstrate that even smaller teachers and students can achieve competitive performance, further pushing the boundaries of parameter efficiency. We explore the impact of using teachers with fewer parameters and distilling knowledge into even smaller student models. We also explore incorporating different losses into the distillation training, such as contrastive loss and adversarial loss.

Our results suggest that effective knowledge distillation can be achieved with significantly smaller teacher and student models, demonstrating competitive performance even with reduced parameters. We find that our DistilledGPT-44M model, despite

The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, pages 302–307

having much fewer parameters, achieves results comparable to the original BabyLlama-58M on key benchmarks. Our experiments with contrastive and adversarial learning techniques in the distillation process, which albeit less promising, reveal interesting trade-offs between different aspects of model performance.

2 Dataset

We use the provided 10M dataset from the strict-small track and build on the BabyLlama repository https://github.com/timinar/BabyLlama. Following their preprocessing steps, we apply regex-based cleaning and train a Byte-Pair Encoding tokenizer on the training set. The train and dev sets are split into 128-token chunks, with the model being presented a new random permutation of these chunks in each epoch. Validation loss is computed at the end of each epoch using a fixed, randomly sampled subset of the dev set.

3 Evaluation

Evaluation of model performance was done using the BabyLM evaluation suite (Choshen et al., 2024). This consists of the following benchmarks:

- **BLiMP:** BLiMP (Benchmark of Linguistic Minimal Pairs for English) evaluates language models on their ability to identify grammatical acceptability. It presents pairs of sentences that differ by one linguistic element, testing the model's understanding of 12 areas of English morphology, syntax, and semantics, such as anaphor agreement and filler-gap constructions. It measures how well models assign higher probability to the grammatically correct sentence in each pair.

- **EWoK:** EWoK (Elements of World Knowledge) evaluates language models on their ability to build and apply internal world models. It tests models' understanding of concepts and contexts by presenting them with minimal pairs of scenarios where models must determine the plausibility of context-target combinations. The framework spans 11 knowledge domains.

- **GLUE:** GLUE (General Language Understanding Evaluation) evaluates language models on a variety of natural language understanding tasks. It covers tasks such as sentiment analysis, text similarity, question answering, and textual entailment. LORA finetuning is used for GLUE in this case, though due to computational constraints, this was only evaluated for DistilledGPT-44M, as it was the only one that showed substantial improvement for BLiMP and EWoK. We get the macroaverage by averaging scores across 9 subtasks - all the subtasks in the BabyLM evaluation suite except for CoLA as CoLA only reports matthews correlation scores.

4 Experiments

4.1 Baselines

We first trained GPT2 models in various sizes (18M, 44M, 97M, 705M) and Llama models in various sizes (20M, 60M, 360M) as a baseline and as future teacher models, using same hyperparameters used in the code of BabyLlama (Timiryasov and Tastet, 2023). Model parameters can be found in Table 1.

4.2 DistilledGPT-44M

For our first experiment, we explore the effect of using smaller teacher models and a more compact student model compared to the original BabyLlama configuration.

We use GPT2-44M and Llama-60M as teacher models, both of which are substantially smaller than the GPT2-705M and Llama-360M teachers used in the original BabyLlama. For the student model, we opt for GPT2-44M instead of the Llama-58M used in BabyLlama. This configuration is a significant reduction in the total number of parameters across both teachers and student.

The knowledge distillation process follows the same general approach as BabyLlama. We first train both teacher models (GPT2-44M and Llama-60M) on our dataset. We then train the GPT2-44M student model using a combination of cross-entropy student loss with true labels, and distillation loss between the student's output and each teacher's output. Model architecture for GPT2-44M follows the baseline 44M model.

4.3 ContrastiveLlama-58M

For our second experiment, we bring contrastive loss into the knowledge distillation process. Contrastive learning tries to bring the representations of similar samples closer together while pushing dissimilar samples apart in the embedding space (Chen et al., 2020). For our task, we use contrastive

Hyperparameter	GPT2-18M-all	GPT2-44M-all	GPT2-97M-all	GPT2-705M-all	Llama-20M-all	llama-60M-all	Llama-360M-G10
Hidden dimension size	320	768	768	1536	384	768	1024
Number of layers	2	2	12	24	2	2	24
Number of attention heads	4	8	12	16	4	8	8
Residual dropout	0.0	0.0	0.0	0.1	N/A	N/A	N/A
Attention dropout	0.0	0.0	0.0	0.1	N/A	N/A	N/A
Embedding dropout	0.0	0.0	0.0	0.1	N/A	N/A	N/A
Learning rate	7e-4	7e-4	7e-4	2.5e-4	3e-4	3e-4	3e-4
Batch size	128	128	128	128	128	128	128
Number of epochs	6	6	6	4	4	4	4
Gradient accumulation steps	2	2	2	16	1	1	8
Warmup steps	300	300	300	300	300	300	300
Mixed precision training (fp16)	True	True	True	True	True	True	True

Table 1: Model hyperparameters for baseline models

loss to encourage the student to produce similar hidden representations to the teacher for the same input while distinguishing between representations of different inputs.

We use GPT2-705M and Llama-360M as teacher models and Llama-58M as the student model. The contrastive loss is computed using the N-pair loss formulation, which considers one positive pair and multiple negative pairs in each training iteration. We set N to 32. For each training sample, we generate 31 negative samples by randomly selecting other samples from the same batch. The positive pair consists of the hidden representations of the teacher and student for the same input, while negative pairs are formed by pairing the teacher's representation with the student's representations for different inputs.

We subdivide the overall loss into 39% cross-entropy student loss with true labels, 39% distillation loss, and 22% N-pair contrastive loss computed on the hidden representations of the teacher and student models. This relative weights of loss were chosen through a preliminary linear search for optimal weights by training on a very small subset of data. Model architecture for student model follows that of BabyLlama-58M model.

4.4 MaskedAdversarialLlama-58M

Our next experiment incorporates adversarial learning into the distillation process by implementing the MATE-KD (Masked Adversarial TExt, a Companion to Knowledge Distillation) algorithm (Rashid et al., 2021). MATE-KD enhances traditional knowledge distillation by introducing an adversarial text generator.

The MATE-KD process consists of two main steps:

- Maximization step: A pre-trained masked language model (MLM) generator is trained to perturb the input text by maximizing the divergence between teacher and student logits. This generator learns to create challenging examples that highlight the differences between the teacher and student models.
- Minimization step: The student model is then trained using knowledge distillation on both the original and perturbed training samples, encouraging it to match the teacher's performance on both standard and adversarial inputs.

For our implementation, we use ELECTRA-56M as the generator, pretraining it on our dataset. Our teacher models are GPT2-44M and Llama-60M, both pretrained on our dataset. The student model remains Llama-58M, consistent with our previous experiments. We equally weight cross-entropy student loss with true labels, knowledge distillation loss, and adversarial distillation loss on perturbed samples in our loss function. Model architecture for student model follows that of BabyLlama-58M model.

5 Results

Our results for these 3 experiments can be found in Table 2.

5.1 DistilledGPT-44M

Our DistilledGPT-44M results are encouraging, as they demonstrate that our significantly smaller model configuration can still achieve competitive performance.

From table 2, we can see that DistillGPT-44M manages to outperform both its parent models, GPT2-44M (which scored 58.2 on BLiMP Supplement and 65.6 on BLiMP Filtered) and Llama-60M (which scored 56.7 on BLiMP Supplement and 63.5 on BLiMP Filtered). This shows that

Child Model	Parent Model 1	Parent Model 2	BLiMP Supplement	BLiMP Filtered	EWoK
BabyLlama-58M	GPT2-705M	Llama-360M	59.5	69.8	50.7
ContrastiveLlama-58M	GPT2-705M	Llama-360M	59.3	68.5	50.0
MaskedAdverserialLlama-58M	GPT2-44M	Llama-60M	56.8	65.9	49.6
DistilledGPT-44M	GPT2-44M	Llama-60M	58.8	66.8	50.0

Table 2: Summary of BLiMP filtered, BLiMP supplement and EWOK results for various methods tried for improving knowledge distillation

	BLiMP Supplement	BLiMP Filtered	EWoK
GPT2-18M	55.9	63.7	49.7
GPT2-44M*	58.2	65.6	50.4
GPT2-97M	58.0	66.0	50.6
GPT2-705M	56.7	66.1	50.6
Llama-20M	56.6	62.8	50.2
Llama-60M*	56.7	63.5	49.6
Llama-360M	55.1	68.2	50.5
LTG-BERT	60.8	60.6	48.9
BabyLlama-58M	59.5	69.8	50.7
DistilledGPT-44M	*58.8*	*66.8*	*50.0*

Table 3: Summary of BLiMP Filtered, BLiMP Supplement, and EWoK performance compared to various benchmarks. Our model is in italics, and * represents its teacher models

DistillGPT-44M is able to draw insights from both parents.

This shows that beyond the normal paradigm of a much larger parent model training a student model, we can use collaborative multi-teacher knowledge distillation to create a model that outperforms both parent models.

We also ran finetuned DistilledGPT-44M on GLUE and compared it against BabyLlama-58M baseline results released by BabyLM organizers, and showed that it comparably (Table 3). DistilledGPT-44M excels in tasks requiring nuanced contextual understanding, such as RTE (natural language inference) and WSC (Winograd Schema Challenge), suggesting strong capability in reasoning tasks. While BabyLlama-58M outperforms DistilledGPT-44M on similarity-focused tasks like QQP and sentiment analysis in SST-2, DistilledGPT-44M's competitive scores highlight its efficient handling of complex, context-dependent tasks, even with a smaller parameter set.

BabyLlama-58M demonstrates stronger generalization across a variety of sentence-pair classification tasks, excelling in QNLI, MNLI, and BoolQ. It also outperforms DistilledGPT-44M on CoLA, indicating better linguistic acceptability. However, DistilledGPT-44M's competitive performance in reasoning tasks suggests an efficient and resource-effective model, making it a viable alternative in scenarios where model size is a constraint. These results underscore DistilledGPT-44M's balance of size and performance, standing strong against the larger BabyLlama model in both accuracy and task diversity.

Additionally, the total training time was greatly reduced from the time it took to train BabyLlama-58M. When training on an A5000 GPU, we reduced the total training time from around 10 hours to around 3 hours, which is a more than 3x reduction in training time.

When running a Wilcoxon Ranked-Sum Test on DistilledGPT-44M and BabyLlama-58M for BLiMP, EWoK and GLUE tests seperately, we see that they are **statistically similar** for both BLiMP, EWoK and GLUE, showing that we are able to achieve comparable performance with greatly reduced training times.

Model	MRPC (F1)	RTE	MultiRC	QQP (F1)	QNLI	WSC	MNLI	SST-2	BoolQ	CoLA (MCC)	Macro Avg
DistilledGPT-48M	80.9	55.4	64.9	75.1	77.4	57.7	66.9	75.9	65.3	-0.01	68.8
BabyLlama-58M	82.0	49.6	60.1	83.6	82.8	38.5	72.4	86.2	65.0	2.2	68.9

Table 4: Results of DistilledGPT compared to BabyLlama-58M in GLUE Benchmark

5.2 ContrastiveLlama-58M

Our ContrastiveLlama-58M model show a slight improvement over the baseline GPT2 and Llama models of similar size, and it performs similarly to BabyLlama-58M, with no substantial difference when we perform a Wilcoxon signed rank test. Nonetheless, we currently do not see a benefit to introducing this contrastive loss giving performance remained around the same. We see a trade-off between contrastive learning and traditional knowledge distillation; in future experiments, different weighting schemes for the losses would be interesting to try.

5.3 MaskedAdversarialLlama-58M

Our MaskedAdversarialLlama-58M model shows a decrease in performance compared to both the BabyLlama-58M baseline and our other experiments. The drop is noticeable in the BLiMP Supplement task, where the score is lower than even the baseline GPT2 and the similarly-sized Llama models. This might suggest that the adversarial training might be conflicting with the student model's ability to capture certain linguistic nuances. It could be possible that the generated adversarial examples are too challenging or not representative enough of the task-specific knowledge required for these evaluations. Similarly with our contrastive experiment, trying different weighting schemes for the loss components might help in balancing the trade-off between robustness and task-specific performance in the future.

6 Limitations and Future Work

Although we showed the effectiveness of knowledge distillation with smaller models, we did not thoroughly explore the lower bounds of model size. In future experiments we could investigate even smaller student models or experiment with a wider range of teacher-student size combinations to find the optimal balance between model size and performance.

Additionally, our experiments with contrastive and adversarial learning techniques (ContrastiveLlama-58M and MaskedAdversarialLlama-58M) did not show improvements over the simpler DistilledGPT-44M model. These advanced techniques probably require further refinement or different implementation strategies to be effective: we could try different weighting schemes for loss components in contrastive and adversarial training. Additionally, for the masked adversarial model, the performance of the generator plays a critical role in generating effective perturbed inputs. Using a more powerful MLM generator, rather than the smaller ELECTRA-56M model we used, could improve the adversarial training process and create better perturbations.

7 Conclusion

Herein, we showed that knowledge distillation can be used even with two very simple parents with around the same number of parameters as the child model, to produce a child model which outperforms both parents. We present DistillGPT-44M, which outperforms both the baseline (GPT2) and one of last year's winning entry for the BabyLM challenge LTG-BERT, while maintaining comparable performance to the other winning entry BabyLlama-58M despite reducing number of parameters by around 25% and cutting training time by around 70%.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [call for papers] the 2nd babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Preprint*, arXiv:2404.06214.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Ahmad Rashid, Vasileios Lioutas, and Mehdi Rezagholizadeh. 2021. MATE-KD: Masked adversarial TExt, a companion to knowledge distillation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1062–1071, Online. Association for Computational Linguistics.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. *Preprint*, arXiv:2308.02019.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.