ICL-VS-FT

Tomer Bar Natan[1]

Tel-Aviv University, Tel-Aviv, Israel
{tomerb5@mail}.tau.ac.il

Abstract. Large pretrained language models have shown surprising in-context learning (ICL)

Keywords: NLP · LLM · ICL · Linearization

1 Literature Review

In-context learning (ICL) is a machine learning approach where a model fine-tunes its knowledge and adapts its behavior based on specific contextual information or examples, allowing it to perform better on tasks related to that context. It enables models to leverage domain-specific or task-specific knowledge without extensive retraining, making them more versatile and adaptable. In their work, [2] explores the remarkable ability of language models, particularly GPT-3, to learn and perform tasks with minimal examples, demonstrating their potential as versatile few-shot learners. The authors showcase the models' impressive performance across a wide range of tasks and emphasize their capacity to generalize from limited data, highlighting the transformative impact of these models on various natural language processing applications.

In recent research, there has been a growing interest in understanding the relationship between two key concepts: in-context learning (ICL) and gradient descent (GD)-based fine-tuning, particularly in the context of transformer models ([5,**?**]). This research seeks to uncover how ICL, which involves adapting and learning in specific contexts, can be effectively integrated with the iterative optimization process of GD, especially when fine-tuning transformer models. However, the majority of the examination was on models that had relaxed constraints and featured linear attention mechanisms:

$$LinearAttn(K, V, q) = KV^q \tag{1}$$

The paper [5], develops an explicit weight values for a linear self-attention layer, achieving an update equivalent to a single iteration of gradient descent (GD) aimed at minimizing mean squared error. Moreover, the authors demonstrate how multiple self-attention layers can progressively execute curvature adjustments, leading to enhancements over standard gradient descent. They proposed the following:

Given a 1-head linear attention layer and the tokens $e_j = (x_j, y_j)$, for $j = 1, ..., N$, one can construct key, query and value matrices W_K, W_Q, W_V as well as the projection matrix P such that a Transformer step on every token e_j is identical to the gradient-induced dynamics $e_j \rightarrow (x_j, y_j) + (0, -\delta W x_j) = (x_j, y_j) + PVK^T q_j$ such that $e_j = (x_j, y_j - \delta y_j)$. For the test data token (x_{N+1}, y_{N+1}) the dynamics are identical.

By doing so, they demonstrate the capability of linear attention to execute gradient descent on the deep representations constructed by the transformer.

Another paper ([**?**]) expand the findings from linear attention to conventional attention mechanisms, substantiating their claims with empirical data. Inspired by [1] and [4], the idea in this is paper to explain language models as meta-optimizers.

Consider W_0 and ΔW, both belonging to $\mathbb{R}^{d_{out} \times d_{in}}$, where W_0 represents the initial parameter matrix, and ΔW signifies the updating matrix. Additionally, let x be a member of $\mathbb{R}^{d_{in}}$, serving as the input representation. A linear layer, subject to optimization via gradient descent, can be articulated as follows:

$$\mathcal{F}(x) = (W_0 + \Delta W)x \tag{2}$$

In the context of the back-propagation algorithm, the determination of ΔW entails the aggregation of outer products derived from historical input representations $x'_i \in \mathbb{R}^{d_{in}}$ and their corresponding error signals $e_i \in \mathbb{R}^{d_{out}}$:

$$\Delta W = \sum_i e_i \otimes x'_i \tag{3}$$

Notably, e_i is the result of scaling historical output gradients by $-\gamma$, the negative learning rate.

By equations (2) and (3), we can derive the dual manifestation of linear layers, optimized through gradient descent, as follows:

$$\begin{aligned} \mathcal{F}(x) &= (W_0 + \Delta W)x \\ &= W_0 x + \Delta W x \\ &= W_0 x + \sum_i (e_i \otimes x'_i)x \\ &= W_0 x + \sum_i e_i (x_i^{'T} x) \\ &= W_0 x + \text{LinearAttn}(E, X', x) \end{aligned} \tag{4}$$

Here, E denotes historical output error signal values, X' corresponds to historical inputs employed as keys, and x serves as the current input, operating as the query.

Their experiments convincingly reveal that a model fine-tuned through gradient steps and a model prompted with in-context examples appear to perform analogous functions, exhibiting similar behaviors on inputs. Additionally, they observe significant similarities in the internal behaviors of these two models.

2 Method

In the following sections we describe the evaluation metrics used to compare the behavior of ICL and finetuning. **TODO**: Adress that we use Dai's metrics.

Prediction Recall

From the perspective of model prediction, models with similar behavior should have aligned predictions. We measure the recall of correct ICL predictions to correct finetuning predictions as suggested by [3]. Given a set of test examples, we count the subsets of examples correctly predicted by each model: $C_{\text{ZSL}}, C_{\text{ICL}}, C_{\text{FT}}$. To compare the update each method induces to the model's prediction we subtract correct predictions made in the ZSL setting. Finally we compute the **Rec2FTP** score as: $\frac{|(C_{\text{ICL}} \cap C_{\text{FT}}) \setminus C_{\text{ZSL}}|}{|C_{\text{FT}} \setminus C_{\text{ZSL}}|}$. A higher Rec2FTP score suggests that ICL covers more correct behavior of finetuning from the perspective of the model prediction.

Attention Output Direction

Considering the hidden state representation space of an attention layer of a model, we compare the updates to the attention output representation (**SimAOU**)[3].

References

1. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I.: Theoretical foundation of potential functions method in pattern recognition (2019), https://api.semanticscholar.org/CorpusID:92987925
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
3. Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., Wei, F.: Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers (2023)
4. Irie, K., Csordás, R., Schmidhuber, J.: The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention (02 2022)
5. Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., Vladymyrov, M.: Transformers learn in-context by gradient descent. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 35151–35174. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/von-oswald23a.html