Nonparametric Teaching of Attention Learners

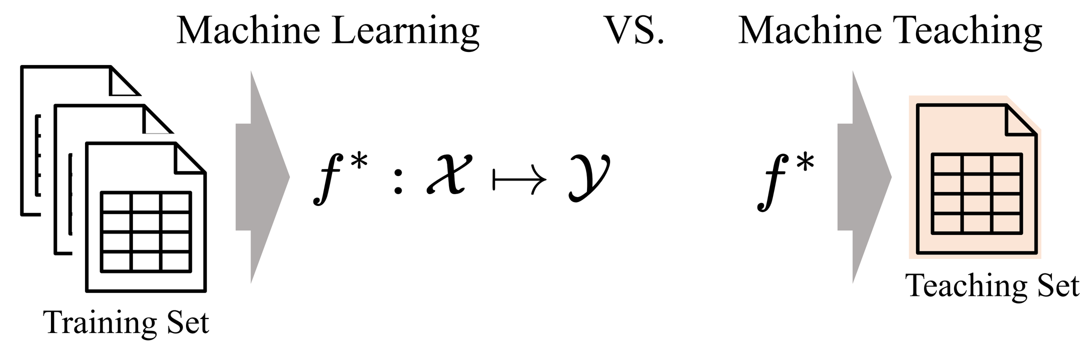
Chen Zhang[1]*, Jianghui Wang[2]*, Bingyang Cheng[1], Zhongtao Chen[1], Wendong Xu[1], Cong Wang[3], Marco Canini[2], Francesco Orabona[2], Yik-Chung Wu[1], Ngai Wong[1]

[1]The University of Hong Kong [2]King Abdullah University of Science and Technology [3]Independent Researcher

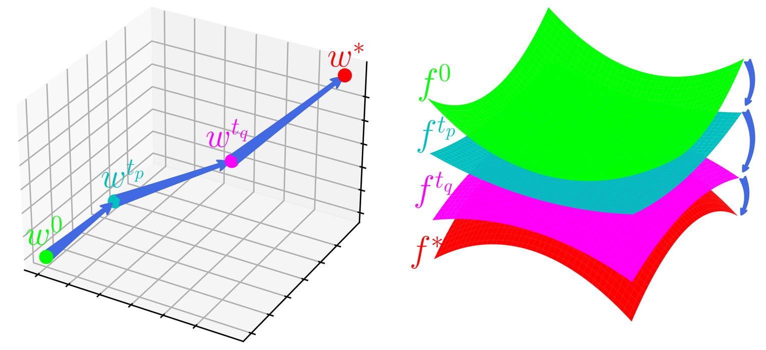
Nonparametric Teaching

Nonparametric teaching (NT) (Zhang et al., 2023b;a; 2024a) presents a theoretical framework to facilitate efficient example selection when the target function is nonparametric, i.e., implicitly defined.

It builds on the idea of *machine teaching* (Zhu, 2015; Zhu et al., 2018), which involves designing a training set (dubbed the teaching set) to help the learner rapidly converge to the target functions.



NT (Zhang et al., 2023b;a; 2024a) relaxes the assumption of target functions† f being parametric (Liu et al., 2017; 2018), which is f can be represented by a set of parameters $\boldsymbol{w}$, *e.g.*, $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ with input $\boldsymbol{x}$, to encompass the teaching of nonparametric target functions.



(a) Parametric IMT (b) Nonparametric IMT

†The loss $\mathcal{L}$ can be general for different tasks, *e.g.*, square loss for regression and hinge loss for classification.

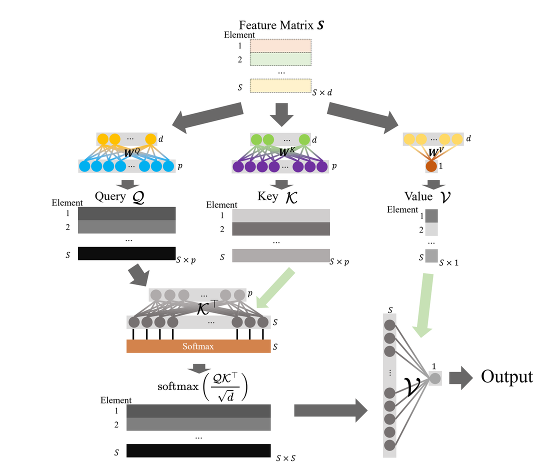
Attention Learners

Attention learners, such as transformers, are designed to learn implicit mappings from sequences to their properties by adaptively assigning importance to each element in the sequence.

Given an input sequence $\boldsymbol{S} \in \mathbb{R}^{S \times d}$, a self-attention layer computes

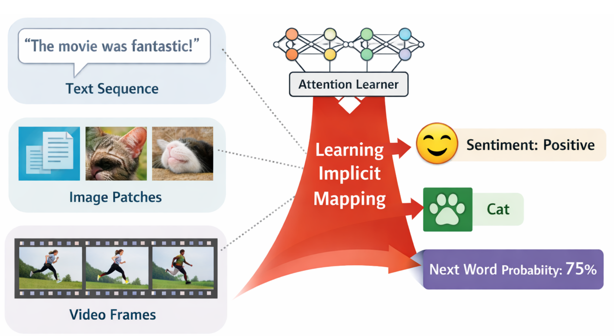
$$f_\theta(\boldsymbol{S}) = \mathrm{softmax}\left(\frac{\mathcal{Q}(\boldsymbol{S})\mathcal{K}(\boldsymbol{S})^\top}{\sqrt{d}}\right)\mathcal{V}(\boldsymbol{S}),$$

where the attention weights determine how much each element contributes to the output.



Self-attention (attention learner).

This adaptive weighting enables long-range dependency modeling and makes attention learners highly expressive for NLP and vision.



Implicit mapping f^* between $\boldsymbol{S}$ and its property $f^*(\boldsymbol{S})$.

Yet, training these models typically requires many gradient steps over large datasets, and most samples contribute redundant or low-information updates. *Can we accelerate attention learner training via example selection, with theoretical guarantees?*

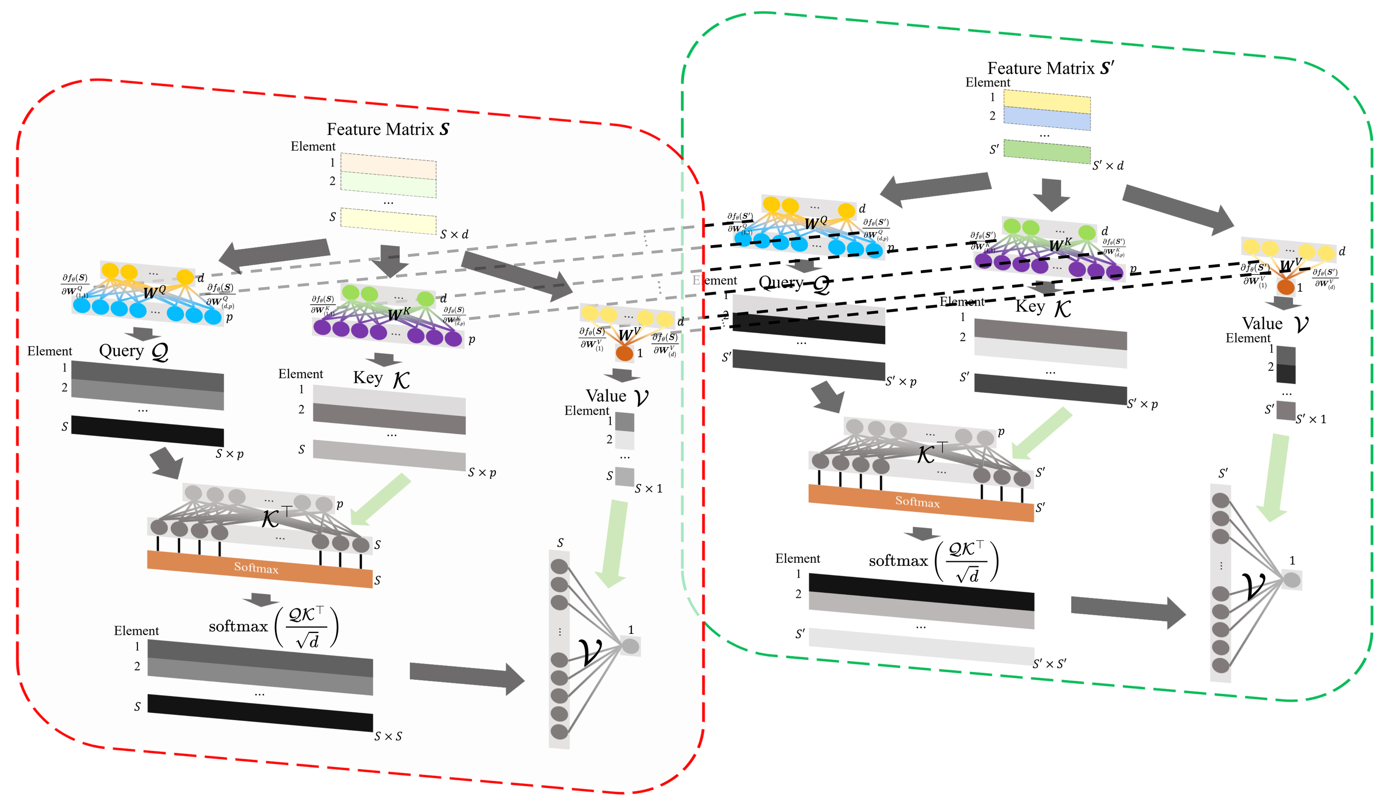
Content by: Chen Zhang.

The Bridge Between NT and Attention Learners: Attention Neural Tangent Kernel

The evolution of attention learners (*e.g.*, transformers) is typically achieved by gradient descent on parameters, while nonparametric teaching characterizes function evolution via functional gradient descent.

Our key insight is that attention induces an importance-adaptive update in parameter space, whose functional evolution can be expressed via a dynamic Attention Neural Tangent Kernel (ANTK).

$$K_{\theta^t}(\boldsymbol{S}_i, \cdot) := \left\langle \frac{\partial f_{\theta^t}(\boldsymbol{S}_i)}{\partial \theta^t}, \frac{\partial f_{\theta^t}(\cdot)}{\partial \theta^t} \right\rangle$$



AtteNT Algorithm

Algorithm 1 AtteNT Algorithm

Input: Target mapping f^* realized by a dense set of sequence-property pairs, initial ANN f_{θ^0}, the size of selected training set $m \leq N$, small constant $\epsilon > 0$ and maximal iteration number T

Set $f_{\theta^t} \leftarrow f_{\theta^0}, t = 0$

while $t \leq T$ and $\|[f_{\theta^t}(\boldsymbol{S}_i) - f^*(\boldsymbol{S}_i)]_N\|_{\mathcal{F}} \geq \epsilon$ **do**

  The teacher selects m teaching sequences:

  /* Sequences associated with the m largest $\|f_{\theta^t}(\boldsymbol{S}_i) - f^*(\boldsymbol{S}_i)\|_2$ */

  $\{\boldsymbol{S}_i\}_m{}^* = \underset{\{\boldsymbol{S}_i\}_m \subseteq \{\boldsymbol{S}_i\}_N}{\arg\max} \|[f_{\theta^t}(\boldsymbol{S}_i) - f^*(\boldsymbol{S}_i)]_m\|_{\mathcal{F}}$

  Provide $\{\boldsymbol{S}_i\}_m{}^*$ to the attention learner

  The learner updates f_{θ^t} based on received $\{\boldsymbol{S}_i\}_m{}^*$:

  // Parameter-based gradient descent

  $\theta^t \leftarrow \theta^t - \frac{\eta}{mS} \sum_{\boldsymbol{S}_i \in \{\boldsymbol{S}_i\}_m{}^*} \sum_{j=1}^{S} \nabla_\theta \mathcal{L}(f_{\theta^t}(\boldsymbol{S}_i)_{(j,:)}, f^*(\boldsymbol{S}_i)_{(j,:)})$

  Set $t \leftarrow t + 1$

end

AtteNT greedily selects sequences whose predicted properties exhibit the largest discrepancy from their targets, and uses them as the teaching set for the next update.

Intuitively, these samples carry the most actionable error signal, so the learner spends its compute budget on what it currently misunderstands most.

By training on only these informative sequences, AtteNT amplifies the effective functional gradient and accelerates convergence of attention learners, while often preserving (or even improving) downstream performance.

Main Contribution

Our key contributions are:

- We propose **Atte**ntion **N**eural **T**eaching (AtteNT), a new paradigm that reinterprets attention learner training through the theoretical lens of nonparametric teaching, enabling greedy example selection to improve learning efficiency.
- We analytically investigate the role of attention in parameter-based gradient descent, and show that the evolution of attention learners under parameter updates is consistent with functional gradient descent. In particular, the dynamic ANTK converges to an importance-adaptive canonical kernel.
- We demonstrate the effectiveness of AtteNT through extensive experiments across NLP and CV tasks: AtteNT reduces LLM fine-tuning time by 13.01% and accelerates ViT training-from-scratch by 20.58%, while preserving and often improving downstream performance.

Results (NLP): LLM Fine-tuning

Table 1: AtteNT on NLG tasks. The results are averaged over three runs, with standard deviations included. The GSM8K and MATH datasets share a math fine-tuned model, while HumanEval and MBPP use a code fine-tuned model. MT-Bench utilizes a conversation fine-tuned model. The "Avg. time" represents the average fine-tuning time for the three models.

Model	AtteNT	Avg. Time(↓)	GSM8K(↑)	MATH(↑)	HumanEval(↑)	MBPP(↑)	MT-Bench(↑)
LLaMA 2-7B	w/o	246±1m	42.96±0.12	5.06±0.16	18.35±0.31	35.65±0.25	**4.58±0.01**
	w	**213±2m**	**43.45±0.55**	**6.48±0.24**	**21.80±0.38**	**37.61±0.42**	4.49±0.02
Mistral-7B	w/o	204±2m	69.13±0.22	20.06±0.20	43.42±0.14	58.52±0.13	5.03±0.05
	w	**180±2m**	**71.26±0.23**	**23.12±0.44**	**46.55±0.25**	**61.74±0.54**	**5.32±0.04**
Gemma-7B	w/o	228±2m	75.23±0.45	30.52±0.48	53.83±0.27	65.69±0.29	5.42±0.04
	w	**201±2m**	**77.74±0.32**	**31.40±0.36**	**54.26±0.28**	**66.28±0.46**	**5.44±0.08**

Results (CV): ViT Training-from-scratch

Table 2: AtteNT across various CV downstream tasks. ImageNetS50 uses 50 categories from ImageNet for classification, evaluated by accuracy. NYUv2(S) is a semantic segmentation task with mIoU as the metric. NYUv2(D) involves depth estimation, evaluated using the δ_1 metric, which measures the percentage of pixels with an error ratio below 1.25 (Doersch & Zisserman, 2017).

Model	AtteNT	Pretraining Time(↓)	ImageNetS50(↑)	NYUv2(S)(↑)	NYUv2(D)(↑)
Multi-Modal MAE	w/o	1234m	92.2	51.9	52.1
	w	**980m**(-20.58%)	**92.3**	**52.6**	**57.2**

Ablation & Resources

Table 3: Ablation study of AtteNT pre-training configurations. Ratio controls how the fraction of selected samples increases over epochs. Interval denotes how often the subset is re-sampled. Selection specifies the sampling strategy: Random (no difficulty prior), Hard (selects only difficult samples), and Soft (Gumbel-Top-k difficulty-aware sampling). The configuration (Incremental, Incremental, Soft) in the red color row is adopted as our final AtteNT setting, as it simultaneously reduces pre-training time and improves performance on all downstream tasks.

Pre-training				Downstream		
Ratio	Interval	Selection	Training time(↓)	ImageNetS50(↑)	NYUv2(S)(↑)	NYUv2(D)(↑)
-	-	-	1234m	92.2	51.9	52.1
Cosine	Incremental	Random	966m	88.6	45.3	49.6
Cosine	Incremental	Soft	995m	92.1	52.2	58.8
Cosine	Fixed	Soft	1301m	**93.2**	53.6	61.4
Incremental	Incremental	Soft	980m	92.3	52.6	57.2
Incremental	Fixed	Soft	1319m	92.4	**53.7**	**62.1**
Cosine	Incremental	Hard	972m	91.8	49.5	57.3
Cosine	Fixed	Hard	1285m	92.1	53.0	60.8
Incremental	Incremental	Hard	**963m**	91.4	48.4	57.2
Incremental	Fixed	Hard	1302m	92.5	52.7	59.5

Project Page