ICLR

Project Page

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

Content by: Chen Zhang.

February 5, 2026

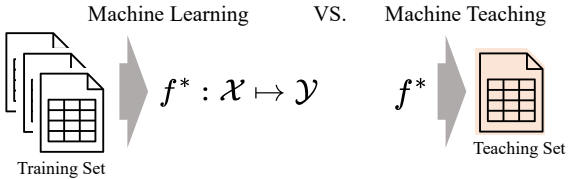
Overview

Nonparametric Teaching

What is Nonparametric Teaching?

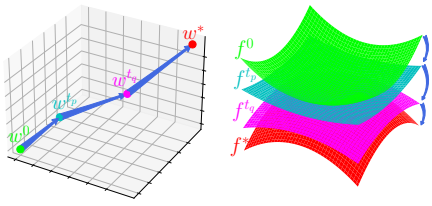
Nonparametric Teaching builds on the idea of *machine teaching* [7]–involving designing a training set (dubbed the teaching set) to help the learner rapidly converge to the target functions–but relaxes the assumption of target functions being parametric [1, 2], allowing for the teaching of nonparametric (viz. non-closed-form) target functions, with a focus on function space.

Machine teaching can be considered as an inverse problem of machine learning, where machine learning aims to learn a model from a dataset, while machine teaching aims to find a minimal dataset from the target model.



“Parametric” VS. “Nonparametric”

The parametric case [1, 2] assumes that f can be represented by a set of parameters $\boldsymbol{w}$, *e.g.*, $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ with input $\boldsymbol{x}$[1].



(a) Parametric IMT (b) Nonparametric IMT

Parametric assumption results in difficulty when the target models are defined to be implicit mappings, such as mappings from sequences to properties learned by attention mechanisms. Such a limitation is addressed by **Nonparametric Teaching** [4, 5, 6, 3], which generalizes model space from a finite dimensional one to an infinite dimensional one.

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

Attention Neural Teaching (AtteNT)

Sequence Property Learning

Many modern learning tasks involve sequences as inputs, such as sentences, image patches, or video frames. Each sequence is associated with a property, *e.g.*, the next-token probability, sentiment label, or class category.

Learning these tasks amounts to learning an implicit mapping from sequences to properties, which is typically achieved by attention learners.

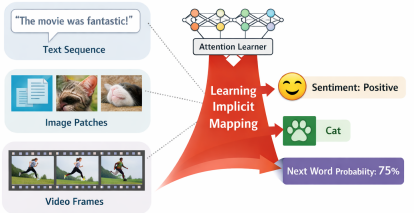


Figure: Implicit mapping from sequences to properties.

Attention Learner

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.

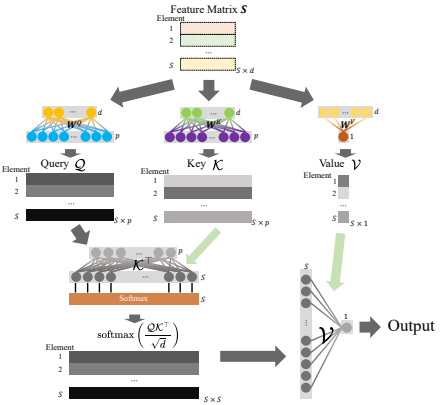


Figure: Workflow of a self-attention learner.

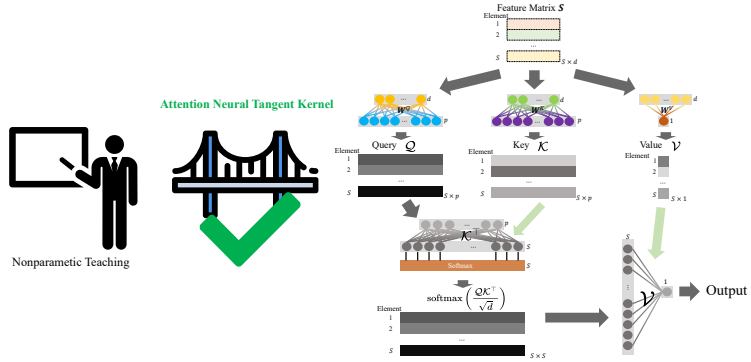
Motivation

The motivation comes from two folds:

- **Efficiency bottleneck of attention learners.** Training attention-based models is notoriously expensive due to quadratic attention computation and large-scale parameterization. This issue becomes particularly severe for LLMs and Vision Transformers, where redundant or low-informative sequences dominate the training process.
- **Lack of principled data selection mechanisms.** Despite their success, attention learners are typically trained using uniformly sampled data, without explicitly leveraging which sequences are most informative for accelerating convergence.
- **Limited scope of nonparametric teaching.** Existing nonparametric teaching frameworks are mainly developed for graph-structured or shallow models, and their applicability to attention-based learners remains largely unexplored.

From Parameter Space to Function Space

We show that the evolution of an attention learner driven by parameter gradient descent can be equivalently expressed as functional gradient descent in nonparametric teaching.



Attention Neural Tangent Kernel (ANTK)

The Attention Neural Tangent Kernel (ANTK) characterizes the evolution of an attention learner under parameter gradient descent.

It captures how attention induces importance-adaptive updates during training.

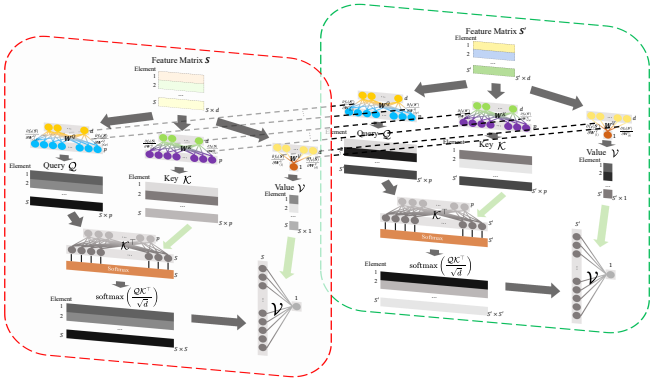


Figure: Illustration of dynamic ANTK.

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 selects sequences whose predicted properties exhibit the largest discrepancy from their targets.

By feeding only these informative sequences to the learner, AtteNT amplifies gradient magnitude and accelerates convergence of attention learners. This process can be interpreted as a form of functional gradient amplification under the attention neural tangent kernel.

Experiments and Results

We conduct extensive experiments on NLP and CV tasks to validate the effectiveness of AtteNT.

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**

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**

Cont.

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

Contribution Summary

Contributions Summary

Main Contributions:

- We propose **Atte**ntion **N**eural **T**eaching (AtteNT), a novel paradigm that interprets attention learner training through nonparametric teaching.
- We establish a theoretical bridge between parameter gradient descent of attention learners and functional gradient descent via ANTK.
- We demonstrate substantial training acceleration for LLMs (13.01%) and ViTs (20.58%) without sacrificing performance.

Thank you for listening!

References I

[1] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In ICML, 2017.

[2] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James Rehg, and Le Song. Towards black-box iterative machine teaching. In ICML, 2018.

[3] Chen Zhang, Weixin Bu, Zeyi Ren, Zhengwu Liu, Yik-Chung Wu, and Ngai Wong. Nonparametric teaching for graph property learners. In ICML, 2025.

[4] Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric iterative machine teaching. In ICML, 2023.

[5] Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric teaching for multiple learners. In NeurIPS, 2023.

[6] Chen Zhang, Steven Tin Sui Luo, Jason Chun Lok Li, Yik-Chung Wu, and Ngai Wong. Nonparametric teaching of implicit neural representations. In ICML, 2024.

[7] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. arXiv preprint arXiv:1801.05927, 2018.