Label-Aware Neural Tangent Kernel

Suyeol Yun

February 3, 2023

Label-Aware Neural Tangent Kernel

Computer Science > Machine Learning

[Submitted on 22 Oct 2020 (v1), last revised 29 Oct 2020 (this version, v2)]

Label-Aware Neural Tangent Kernel: Toward Better Generalization and Local Elasticity

Shuxiao Chen, Hangfeng He, Weijie J. Su

As a popular approach to modeling the dynamics of training overparametrized neural networks (NNs), the neural tangent kernels (NTK) are known to fall beh performance gap is in part due to the \textit{label agnostic} nature of the NTK, which renders the resulting kernel not as \textit{locally elastic} as NNs~\citep approach from the perspective of \emph{label-awareness} to reduce this gap for the NTK. Specifically, we propose two label-aware kernels that are each a su label-aware parts with increasing complexity of label dependence, using the Hoeffding decomposition. Through both theoretical and empirical evidence, we better simulate NNs in terms of generalization ability and local elasticity.

Comments: NeurIPS 2020 camera ready version, 32 pages, 2 figures, 3 tables
Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)
Cite as: arXiv:2010.11775 **[cs.LG]**
(or arXiv:2010.11775v2 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2010.11775

Submission history

From: Shuxiao Chen [view email]
[v1] Thu, 22 Oct 2020 14:54:32 UTC (601 KB)
[v2] Thu, 29 Oct 2020 17:23:44 UTC (604 KB)

Chen et al. (Neurips 2020)

Motivation

- How to explain *performance gap* between NTK and real-world NN?
- Arora (2019)

Depth	CNN-V	CNTK-V	CNTK-V-2K	CNN-GAP	CNTK-GAP	CNTK-GAP-2K
3	59.97%	64.47%	40.94%	63.81%	70.47%	49.71%
4	60.20%	65.52%	42.54%	80.93%	75.93%	51.06%
6	64.11%	66.03%	43.43%	83.75%	76.73%	51.73%
11	69.48%	65.90%	43.42%	82.92%	**77.43%**	51.92%
21	75.57%	64.09%	42.53%	83.30%	77.08%	52.22%

Table 1: Classification accuracies of CNNs and CNTKs on the CIFAR-10 dataset. CNN-V represents vanilla CNN and CNTK-V represents the kernel corresponding to CNN-V. CNN-GAP represents CNN with GAP and CNTK-GAP represents the kernel correspondong to CNN-GAP. CNTK-V-2K and CNTK-GAP-2K represent training CNTKs with only 2,000 training data.

Label-agnostic vs. Label-aware

- NN is label-aware while NTK is label-agnostic
- NTK construction is independent of the labels[1]

$$K_{NTK}(x, \tilde{x}) = \left\langle \nabla f_x \left(w^{(0)} \right), \nabla f_{\tilde{x}} \left(w^{(0)} \right) \right\rangle$$

- NN can be thought of *label-aware* because trained parameter of NN depend on labels.

[1] Remind that NTK is soley a function of a network architecture.

Label-aware NTK

- Goal is to develop a label-aware NTK.
- *Optimal* feature map $\phi^\star(x_i) = y_i$
- *Optimal* kernel $K^\star(x_i, x_j) = \langle \phi^\star(x_i), \phi^\star(x_j) \rangle = y_i y_j$
- But we don't know y_α and y_β for test cases. So design estimator of $y_\alpha y_\beta$ and augment it to NTK.

$$K_{\text{LA-NTK}}(x_i, x_j) = K_{NTK} + \lambda \mathcal{Z}(x_i, x_j, \mathcal{D})$$

where $\mathcal{Z}(x_i, x_j, \mathcal{D}) = y^\top \mathbf{M}(x_i, x_j, X)\, y$ which is a linear regression model to estimate $y_i y_j$ for given (x_i, x_j).

Performance of Label-aware NTK

	deer vs dog	cat vs deer	cat vs frog	deer vs frog	bird vs frog	Avg	Imp. (abs./rel.)
CNTK	85.15	83.55	86.95	87.55	86.35	85.91	-
LANTK-best	**86.75**	**84.85**	**87.40**	**88.30**	**87.45**	**86.95**	1.04 / 7.4%
LANTK-KR-V1	85.65	83.90	87.20	87.85	**87.45**	86.41	0.50 / 3.5%
LANTK-KR-V2	85.80	83.90	87.15	87.90	87.25	86.40	0.49 / 3.5%
LANTK-FJLT-V1	**86.75**	**84.85**	**87.40**	88.10	86.40	**86.70**	0.79 / 5.6%
LANTK-FJLT-V2	86.25	84.55	87.10	**88.30**	87.00	86.64	0.73 / 5.2%
CNN	89.00	88.50	89.00	92.50	90.15	89.83	3.92 / 27.8%

[2] "... this is a theoretically motivated work, so I am not concerned about the size of the empirical improvements."

Further thoughts

- What else remains to explain performance gap between NTK and NN?
  - Is label-agnostic asepct of NTK sufficient to explain performance gap?
- The reason of underperformance compared to CNN maybe due to the limited expressivity of $\mathcal{Z}$ used in the paper which estimates $y_i y_j$ given (x_i, x_j).
- Will there be anyway to derive this $\mathcal{Z}$ part analytically as a function of network architecture as well as we did in derivation of NTK?

References

- Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In Advances in Neural Information Processing Systems, pp. 8139–8148, 2019a.
- Shuxiao Chen, Hangfeng He, Weijie J. Su: Label-Aware Neural Tangent Kernel: Toward Better Generalization and Local Elasticity. NeurIPS 2020