

A Theoretical Understanding of Shallow Vision Transformers: Learning, Generalization, and Sample complexity

Hongkang Li ¹, Meng Wang ¹, Sijia Liu^{2,3}, Pin-Yu Chen³,

¹Rensselaer Polytechnic Institute

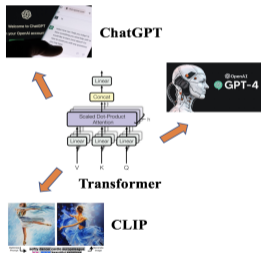
²Michigan State University

³IBM Research

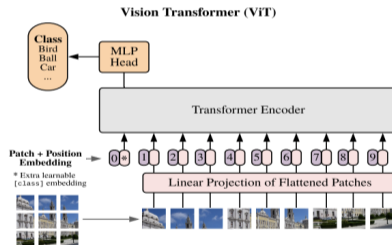
International Conference on Learning Representations (ICLR 2023)
May, 2023

Transformers and Vision Transformers (ViTs)

- Transformers achieved great empirical success in numerous areas.
- Transformer-based models gradually become prevalent in vision tasks.



Transformer-based foundation models



Vision Transformer [Dosovitskiy et al.21]

Under what conditions does a Vision Transformer achieve satisfactory generalization?

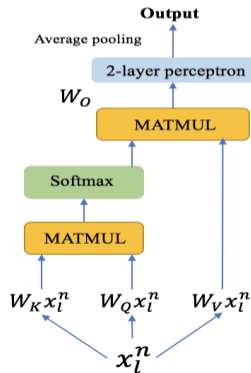
Problem formulation and the ViT model

We study a binary classification problem with the dataset $\{\mathbf{X}^n, y^n\}_{n=1}^N$ with L patches, i.e., tokens in each \mathbf{X}^n . Each token is a noisy version of a pattern. There are M patterns in total, where two are discriminative patterns that can determine the label.

- Labeling function: majority voting of discriminative tokens.
- Learner network: a shallow ViT with a single-head self-attention layer and a two-layer perceptron.

$$F(\mathbf{X}^n) = \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \mathbf{a}_{(l)}^\top \text{Relu}(\mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n)) \quad (1)$$

- Loss function: Hinge loss function. Training with SGD.



Main theoretical results

Data model: We sample a set of tokens for each data. Define

- label-relevant tokens: tokens with the pattern that corresponds to the **exact** label of the data.
- confusion tokens: tokens with the pattern that corresponds to the **other** label of the data.

Theorem 1

Given a sufficient large model and α_* , $\alpha_{\#}$ with

$$m \gtrsim M^2 \log N, \quad (2)$$

$$\alpha_* \geq \alpha_{\#}/c, \quad \alpha_*, \alpha_{\#} : \text{average fraction of label-relevant, confusion tokens} \quad (3)$$

for some $c \in (0, 1/(2e))$, and large enough sizes of mini-batch and the set of sampled tokens for each data, **zero generalization error** is achieved with a sample complexity N and a number of iterations T :

$$N \geq \Omega(\alpha_*^{-2}), \quad T = \Theta(\alpha_*^{-1} \eta^{-1}), \quad \eta : \text{step size} \quad (4)$$

Main insights

- **Requirements for the data:** the fraction of label-relevant tokens is much more than that of confusion tokens in each data.
- **Sample complexity N :** linear in α_*^{-2} .
- **Required number of iterations T :** linear in α_*^{-1} .
- **Technical novelty:** A new theoretical framework to analyze the nonconvex interactions in shallow ViTs, which contain a trainable self-attention layer.

Comparison between ViT and CNN

Proposition 1

With an approximately the same size of the model, the sample complexity of using CNN to achieve zero generalization error is $\Omega(\alpha_*^{-4})$, which is an increase by a factor of α_*^{-2} compared to ViT.

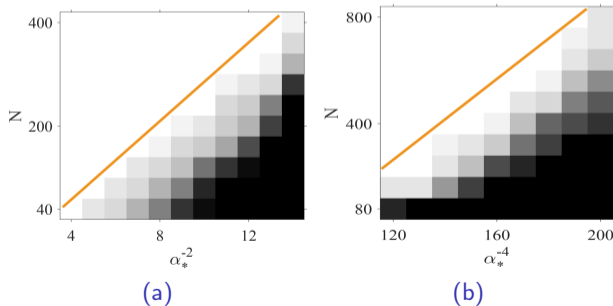


Figure 1: The impact of α_* on the sample complexity for (a) ViT and (b) CNN.

Sparse attention map and token sparsification

Proposition 2

- The summation of attention weights correlated with label-relevant tokens converges to $1 - \eta^C$ at a sublinear rate of $O(1/t)$ for $C > 0$ when t is large.
- Removing label-irrelevant tokens or tokens with large noise can improve the generalization.

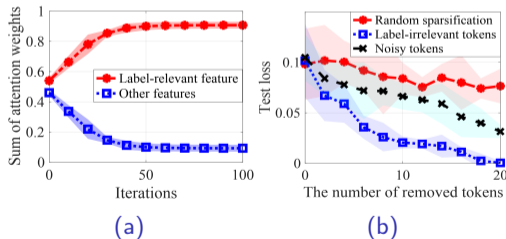






Figure 2: (a) Concentration of attention weights (b) Impact of token sparsification on testing loss.

-  Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I.
Attention is All you Need.
In Advances in Neural Information Processing Systems 2017.
-  Radford A., Kim J., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. (2021).
Learning transferable visual models from natural language supervision.
Proceedings of the 38th International Conference on Machine Learning (ICML 2021).
-  OpenAI (2023)
GPT-4 Technical Report
-  Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N.
An image is worth 16x16 words: Transformers for image recognition at scale.
International Conference on Learning Representations (ICLR 2021)