VISION TRANSFORMERS PROVABLY LEARN SPATIAL STRUCTURE

Vladimir Zaigrajew
vladimir.zaigrajew.dokt@pw.edu.pl

Vision transformers provably learn spatial structure

[PDF] neurips.cc

S Jelassi, M Sander, Y Li

Advances in Neural Information Processing Systems, 2022 • proceedings.neurips.cc

Abstract

Vision Transformers (ViTs) have recently achieved comparable or superior performance to Convolutional neural networks (CNNs) in computer vision. This empirical breakthrough is even more remarkable since ViTs discards spatial information by mixing patch embeddings and positional encodings and do not embed any visual inductive bias (eg\spatial locality). Yet, recent work showed that while minimizing their training loss, ViTs specifically learn spatially delocalized patterns. This raises a central question: how do ViTs

SHOW MORE

☆ Save Cite Cited by 86 Related articles All 6 versions

Samy Jelassi

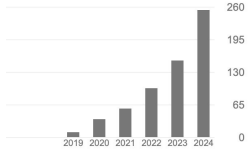
Harvard University

Verified email at fas.harvard.edu

Deep Learning

Cited by

	All	Since 2019
Citations	613	613
h-index	13	13
i10-index	14	14



Michael E. Sander

Other names

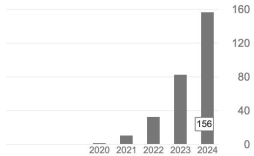
Google DeepMind

Verified email at google.com - Homepage

Machine Learning Applied Mathematics

Cited by

	All	Since 2019
Citations	284	284
h-index	7	7
i10-index	6	6



Meta Review of Paper9794 by Area Chair yr4v

NeurIPS 2022 Conference Paper9794 Area Chair yr4v

21 Aug 2022 at 21:21 NeurIPS 2022 Conference Paper9794 Meta Review Readers: Everyone Show Revisions

Recommendation: Accept

Confidence: Less certain

Metareview:

This paper provides a theoretical analysis of the empirical finding that Vision Transformers learn position embeddings that recapitulate the spatial structure of the training data, even though this spatial structure is no longer explicitly represented after the image is split into patches. The reviewers are generally satisfied by the soundness of the theory, but there is some disagreement regarding the significance of the contribution. The AC believes this paper asks an interesting theoretical question, even if (as is often true) it can only be answered in a simplified setting, and the answer is nontrivial. The AC thus recommends acceptance.

Award: No

Yuanzhi Li

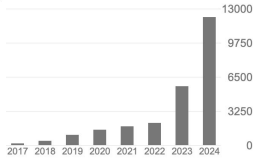
Assistant Professor at CMU

Verified email at andrew.cmu.edu

Machine Learning

Cited by

	All	Since 2019
Citations	25525	24610
h-index	52	50
i10-index	93	92



TLDR

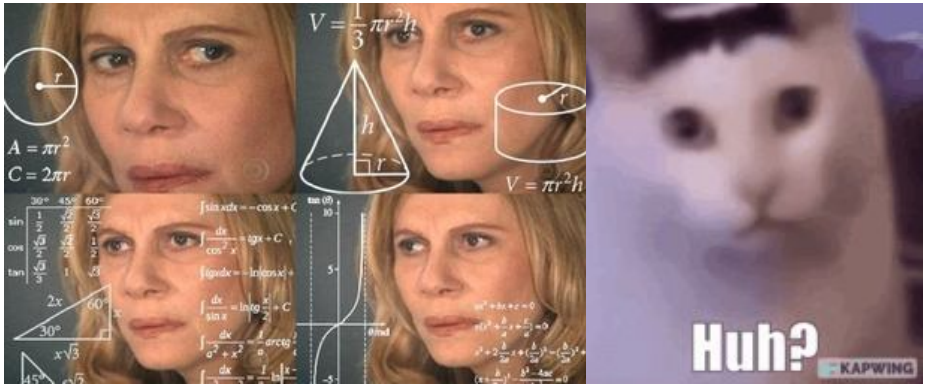
Metareview:
This paper provides a theoretical analysis of the empirical finding that Vision Transformers learn position embeddings that recapitulate the spatial structure of the training data, even though this spatial structure is no longer explicitly represented after the image is split into patches. The reviewers are generally satisfied by the soundness of the theory, but there is some disagreement regarding the significance of the contribution. The AC believes this paper asks an interesting theoretical question, even if (as is often true) it can only be answered in a simplified setting, and the answer is nontrivial. The AC thus recommends acceptance.

Summary:
This work aims to understand how Transformers learn inductive biases as convolution when trained on large datasets with gradient descent. They provide detailed analysis and find that ViT relies on the positional attention mechanism that disentangles patches and positional encodings. And it also provides experimentS on the Gaussian data that show that ViTs do not learn the correct inductive bias under the assumption that "characterizing the distributions under which ViTs recover the structure of the function is an important question." Lastly, it also provides experiments on CIFAR and ImageNet.

Summary:
Vision Transformers

The paper focuses on trying to explain how Vision Transformers learn the spatial locality biases on images. To do so, authors mathematically analyze the learning of convolution-like structures on ViT and test their idea with some synthetic datasets based on CIFAR.

My TLDR



DETAILS: Generally speaking, this paper is VERY difficult to follow. My background is computer science but I found it hard to read this paper that is FULL of definitions, assumptions, definitions, theorems but, most importantly, the paper introduces many many letters in each section.

The math seems reasonable but I have to admit that around line 200 I was already a bit lost and I could not follow the paper so well. I might have missed something. This raises also a problem for me: I did not understand how ViT learns the spatial inductive bias after reading 9 pages only about that. Moreover, I am not even sure that the tests done with 1-layer ViTs are generalizable to more layers.

Moreover, I think the paper misses one main point: explaining WHY it is important to learn how ViT learns the inductive biases. I know it sounds a stupid comment, but without this explanation a reader is not intrigued by the paper and the heavy math inside it. Then, once we know that “ViT do not just group nearby pixel together” why is it useful? I mean, I am very naive but self-attention is based on content, so I am not surprised that it does not group nearby pixel together.

Main Question Authors want to answer:

from a theoretical perspective, how do ViTs manage to learn these local connectivity patterns by simply minimizing their training loss using gradient descent from random initialization?

Main Question Authors want to answer:
from a theoretical perspective, how do ViTs manage to learn these local connectivity patterns by simply minimizing their training loss using gradient descent from random initialization?

Contributions:

- Define the concept of performing patch association, which refer to the ability of learning spatial connectivity patterns on a dataset.
- Introduce a structured classification dataset and a **simplified** ViT model that can answer the main question.
- Prove that a one-layer single-head ViT model trained with gradient descent on our synthetic dataset performs patch association and generalizes.
- After pre-training simplified model can be fine-tuned to transfer to a downstream dataset that shares the same structure as the source dataset (and may have different features).
- Paper show that simplified ViT is competitive with the vanilla ViT on the ImageNet, CIFAR-10/100 and SVHNs datasets

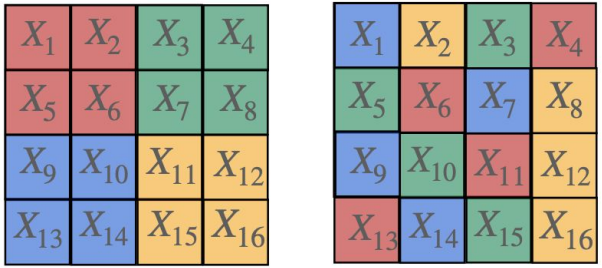
First Definitions

Definition 2.1 (Data distribution with spatial structure). *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^{d\times D}\times\{-1,1\}$ where each patch $\boldsymbol{X}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_D)\in\mathbb{R}^{d\times D}$ has label $y\in\{-1,1\}$. We say that $\mathcal{D}$ is spatially structured if*

- *there exists a partition of $[D]$ into L disjoint subsets i.e. $[D]=\bigcup_{\ell=1}^{L}\mathcal{S}_\ell$ with $\mathcal{S}_\ell\subsetneq D$ and $|\mathcal{S}_\ell|=C$.*
- *there exists a labeling function f^* satisfying $\mathbb{P}[yf^*(\boldsymbol{X})>0]=1-d^{-\omega(1)}$ and,*

$$f^*(\boldsymbol{X}):=\sum_{\ell\in[L]}\phi((\boldsymbol{X}_i)_{i\in\mathcal{S}_\ell}),\quad \text{where } \phi\colon\mathbb{R}^{d\times C}\to\mathbb{R} \text{ is an arbitrary function.}$$

Definition 2.2 (Patch association for ViTs). *Let $\mathcal{D}$ be as in Definition 2.1. Let $\mathcal{M}\colon\mathbb{R}^{d\times D}\to\{-1,1\}$ be a transformer and $\boldsymbol{P}^{(\mathcal{M})}$ its positional encodings matrix. We say that $\mathcal{M}$ performs patch association on $\mathcal{D}$ if for all $\ell\in[L]$ and $i\in\mathcal{S}_\ell$, we have $\mathrm{Top}_C\,\{\langle\boldsymbol{p}_i^{(\mathcal{M})},\boldsymbol{p}_j^{(\mathcal{M})}\rangle\}_{j=1}^{D}=\mathcal{S}_\ell$.*



A Question they ask: would ViTs really learn those Sℓ after training to match the labeling function f?
The Answer: Without further assumptions on the data distribution, we next show that the answer is **NO**.

ViTs do not always learn patch association

Consider the case where all the patches Xj are i.i.d. standard Gaussian and `f` is a one-hidden layer CNN with cubic activation. The label y of any X is then given by sing(f(X))

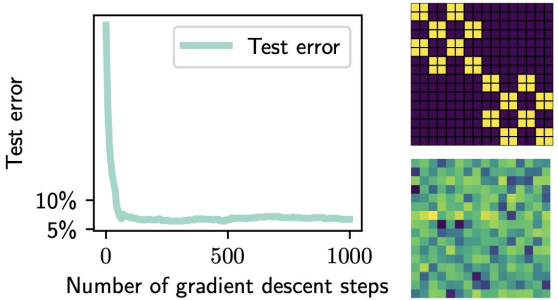


Figure 2: Left: Test error of the ViT on the convolution structured dataset. Upper Right: Grid displaying the input patches. Yellow squares represent spatially localized sets $\mathcal{S}_\ell$. Those sets are taken into account when computing the convolutional function f^*. Lower Right: Learnt $\boldsymbol{P}^\top \boldsymbol{P}$ looks random compared to upper one.

This is not surprising, since the data distribution D is Gaussian, and thus lacks spatial structure.

ViTs do not always learn patch association

Consider the case where all the patches Xj are i.i.d. standard Gaussian and `f` is a one-hidden layer CNN with cubic activation. The label y of any X is then given by sing(f(X))

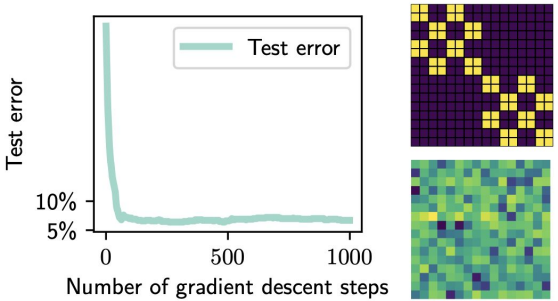


Figure 2: Left: Test error of the ViT on the convolution structured dataset. Upper Right: Grid displaying the input patches. Yellow squares represent spatially localized sets $\mathcal{S}_\ell$. Those sets are taken into account when computing the convolutional function f^*. Lower Right: Learnt $\boldsymbol{P}^\top \boldsymbol{P}$ looks random compared to upper one.

This is not surprising, since the data distribution D is Gaussian, and thus lacks spatial structure.

WE NEED MORE ASSUMPTIONS ON D

Synthetic Dataset

Assumption 1 (Data distribution with specific spatial structure)**.** *Let $\mathcal{D}$ be a distribution as in Definition 2.1 and $\boldsymbol{w}^* \in \mathbb{R}^d$ be an underlying feature. We suppose that each data-point $\boldsymbol{X}$ is defined as follow*

- *Uniformly sample an index $\ell(\boldsymbol{X})$ from $[L]$ and for $j \in \mathcal{S}_{\ell(\boldsymbol{X})}$, $\boldsymbol{X}_j = y\boldsymbol{w}^* + \boldsymbol{\xi}_j$, where $y\boldsymbol{w}^*$ is the informative feature and $\boldsymbol{\xi}_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2(\mathbf{I}_D - \boldsymbol{w}^*\boldsymbol{w}^{*\top}))$ (**signal set**).*
- *For $\ell \in [L]\backslash\{\ell(\boldsymbol{X})\}$ and $j \in \mathcal{S}_\ell$, $\boldsymbol{X}_j = \delta_j \boldsymbol{w}^* + \boldsymbol{\xi}_j$, where $\delta_j = 1$ with probability $q/2$, -1 with same probability and 0 otherwise, and $\boldsymbol{\xi}_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2(\mathbf{I}_D - \boldsymbol{w}^*\boldsymbol{w}^{*\top}))$ (**random sets**).*

Our dataset can be viewed as an extreme simplification of real-world image datasets where there is a set of adjacent patches that contain a useful feature (e.g. the nose of a dog) and many patches that have uninformative or spurious features e.g. the background of the image.

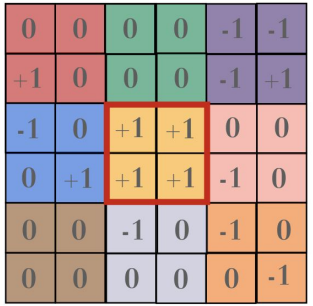


Figure 3: Visualization of a data-point $\boldsymbol{X}$ in $\mathcal{D}$ when the $\mathcal{S}_\ell$'s are spatially localized. Each square depicts a patch $\boldsymbol{X}_j$ and squares of the same color belong to the same set $\mathcal{S}_\ell$. "0" indicates that the patch does not have a feature, "1" stands for feature $1 \cdot \boldsymbol{w}^*$ and "-1" for feature $-1 \cdot \boldsymbol{w}^*$. The large red square depicts the signal set $\ell(\boldsymbol{X})$. Although there are more "-1"'s than "+1"'s, the label of $\boldsymbol{X}$ is $+1$ since there are only "+1"'s inside the signal set.

Simplified ViT

Theorem 3.2. *Let $\mathcal{D}$ be defined as in Assumption 1. There exists a (one-layer) transformer $\mathcal{M}$ so that $\mathbb{P}[f^*(\boldsymbol{X})\mathcal{M}(\boldsymbol{X}) \leqslant 0] = d^{-\omega(1)}$ but for all $\ell \in [L]$, $i \in \mathcal{S}_\ell$, $\mathrm{Top}_C\ \{\langle \boldsymbol{p}_i^{(\mathcal{M})}, \boldsymbol{p}_j^{(\mathcal{M})} \rangle\}_{j=1}^D \cap \mathcal{S}_\ell = \varnothing$.*

Original

Definition 3.1 (Self-attention [Bahdanau et al., 2014, Vaswani et al., 2017])**.** *The attention mechanism [Bahdanau et al., 2014, Vaswani et al., 2017] in the single-head case is defined as follow. Let $\boldsymbol{X} \in \mathbb{R}^{d\times D}$ a data point and $\boldsymbol{P} \in \mathbb{R}^{d\times D}$ its positional encoding. The self-attention mechanism computes*

1. *the sum of patches and positional encodings i.e. $\boldsymbol{X} = \boldsymbol{X} + \boldsymbol{P}$.*
2. *the attention matrix $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^\top$ where $\boldsymbol{Q} = \boldsymbol{X}^\top \boldsymbol{W_Q}$, $\boldsymbol{K} = \boldsymbol{X}^\top \boldsymbol{W_K}$, $\boldsymbol{W_Q}, \boldsymbol{W_K} \in \mathbb{R}^{d\times d}$.*
3. *the score matrix $\boldsymbol{S} \in \mathbb{R}^{D\times D}$ with coefficients $S_{i,j} = \exp(A_{i,j}/\sqrt{d}) / \sum_{r=1}^D \exp(A_{i,r}/\sqrt{d})$.*
4. *the matrix $\boldsymbol{V} = \boldsymbol{X}^\top \boldsymbol{W_V}$, where $\boldsymbol{W_V} \in \mathbb{R}^{d\times d}$.*

It finally outputs $\mathrm{SA}((\boldsymbol{X};\boldsymbol{P})) = \boldsymbol{S}\boldsymbol{V} \in \mathbb{R}^{d\times D}$.

Proposed

Definition 3.2 (Positional attention)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{d\times D}$ and $\boldsymbol{P} \in \mathbb{R}^{d\times D}$ the positional encoding. The positional attention mechanism takes as input the pair $(\boldsymbol{X};\boldsymbol{P})$ and computes:*

1. *the attention matrix $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^\top$ where $\boldsymbol{Q} = \boldsymbol{P}^\top \boldsymbol{W_Q}$, $\boldsymbol{K} = \boldsymbol{P}^\top \boldsymbol{W_K}$ and $\boldsymbol{W_Q}, \boldsymbol{W_K} \in \mathbb{R}^{d\times d}$.*
2. *the score matrix $\boldsymbol{S} \in \mathbb{R}^{D\times D}$ with coefficients $S_{i,j} = \exp(A_{i,j}/\sqrt{d}) / \sum_{r=1}^D \exp(A_{i,r}/\sqrt{d})$.*
3. *the matrix $\boldsymbol{V} = \boldsymbol{X}^\top \boldsymbol{W_V}$, where $\boldsymbol{W_V} \in \mathbb{R}^{d\times d}$.*

It outputs $\mathrm{PA}((\boldsymbol{X};\boldsymbol{P})) = \boldsymbol{S}\boldsymbol{V}$.

Simplification 3.1. *In the positional attention mechanism, we set $d = D$, $\boldsymbol{W_K} = \boldsymbol{I}_D$ and $\boldsymbol{W_Q} = \boldsymbol{I}_D$ which implies $\boldsymbol{A} = \boldsymbol{P}^\top \boldsymbol{P}$. We set $\boldsymbol{W_V} = [\boldsymbol{v}, \dots, \boldsymbol{v}] \in \mathbb{R}^{d\times D}$ where $\boldsymbol{v} \in \mathbb{R}^d$. Finally, we set $\boldsymbol{A}$ and $\boldsymbol{v}$ as trainable parameters. Besides, without loss of generality, we train all $A_{i,j}$ for $i \neq j$ and leave the diagonals of $\boldsymbol{A}$ fixed.*

Simplified ViT

Original

Definition 3.1 (Self-attention [Bahdanau et al., 2014, Vaswani et al., 2017]). *The attention mechanism [Bahdanau et al., 2014, Vaswani et al., 2017] in the single-head case is defined as follow. Let* $\boldsymbol{X} \in \mathbb{R}^{d\times D}$ *a data point and* $\boldsymbol{P} \in \mathbb{R}^{d\times D}$ *its positional encoding. The self-attention mechanism computes*

1. *the sum of patches and positional encodings i.e.* $\boldsymbol{\mathcal{X}} = \boldsymbol{X} + \boldsymbol{P}$.
2. *the attention matrix* $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^\top$ *where* $\boldsymbol{Q} = \boldsymbol{\mathcal{X}}^\top \boldsymbol{W_Q}$, $\boldsymbol{K} = \boldsymbol{\mathcal{X}}^\top \boldsymbol{W_K}$, $\boldsymbol{W_Q}, \boldsymbol{W_K} \in \mathbb{R}^{d\times d}$.
3. *the score matrix* $\boldsymbol{S} \in \mathbb{R}^{D\times D}$ *with coefficients* $S_{i,j} = \exp(A_{i,j}/\sqrt{d})/\sum_{r=1}^{D}\exp(A_{i,r}/\sqrt{d})$.
4. *the matrix* $\boldsymbol{V} = \boldsymbol{\mathcal{X}}^\top \boldsymbol{W_V}$, *where* $\boldsymbol{W_V} \in \mathbb{R}^{d\times d}$.

It finally outputs $\mathrm{SA}((\boldsymbol{X};\boldsymbol{P})) = \boldsymbol{S}\boldsymbol{V} \in \mathbb{R}^{d\times D}$.

Proposed

Definition 3.2 (Positional attention). *Let* $\boldsymbol{X} \in \mathbb{R}^{d\times D}$ *and* $\boldsymbol{P} \in \mathbb{R}^{d\times D}$ *the positional encoding. The positional attention mechanism takes as input the pair* $(\boldsymbol{X};\boldsymbol{P})$ *and computes:*

1. *the attention matrix* $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^\top$ *where* $\boldsymbol{Q} = \boldsymbol{P}^\top \boldsymbol{W_Q}$, $\boldsymbol{K} = \boldsymbol{P}^\top \boldsymbol{W_K}$ *and* $\boldsymbol{W_Q}, \boldsymbol{W_K} \in \mathbb{R}^{d\times d}$.
2. *the score matrix* $\boldsymbol{S} \in \mathbb{R}^{D\times D}$ *with coefficients* $S_{i,j} = \exp(A_{i,j}/\sqrt{d})/\sum_{r=1}^{D}\exp(A_{i,r}/\sqrt{d})$.
3. *the matrix* $\boldsymbol{V} = \boldsymbol{X}^\top \boldsymbol{W_V}$, *where* $\boldsymbol{W_V} \in \mathbb{R}^{d\times d}$.

It outputs $\mathrm{PA}((\boldsymbol{X};\boldsymbol{P})) = \boldsymbol{S}\boldsymbol{V}$.

Simplification 3.1. *In the positional attention mechanism, we set* $d = D$, $\boldsymbol{W_K} = \boldsymbol{I}_D$ *and* $\boldsymbol{W_Q} = \boldsymbol{I}_D$ *which implies* $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$. *We set* $\boldsymbol{W_V} = [\boldsymbol{v},\dots,\boldsymbol{v}] \in \mathbb{R}^{d\times D}$ *where* $\boldsymbol{v} \in \mathbb{R}^d$. *Finally, we set* $\boldsymbol{A}$ *and* $\boldsymbol{v}$ *as trainable parameters. Besides, without loss of generality, we train all* $A_{i,j}$ *for* $i \neq j$ *and leave the diagonals of* $\boldsymbol{A}$ *fixed.*

Under Simplification 3.1, our simplified ViT model is then a two attention layer with a single head:

$$F(\boldsymbol{X}) = \sum_{i=1}^{D}\sigma\Big(D\sum_{j=1}^{D}S_{i,j}\langle \boldsymbol{v}, \boldsymbol{X}_j\rangle\Big) \quad \text{with} \quad S_{i,j} = \exp(A_{i,j}/\sqrt{d})/\sum_{r=1}^{D}\exp(A_{i,r}/\sqrt{d}), \qquad \text{(T)}$$

How model learns patch association when optimizing only to minimize loss function

Lemma 4.1 and **Lemma 4.2** imply that instead of optimizing over A and v, we can instead consider the scalar variables $\alpha^{(t)}, \gamma^{(t)}$ and $\rho^{(t)}$

Event I: At this point, the model is nothing else than a generalized linear model that would not generalize because there are much more noisy tokens than signal ones

Main insights:

- because of the initialization and the data structure, we have patch association for any time t.
- simplified ViT model uses patch association to minimize the population loss (Event III). Without patch association, the model would only be a generalized linear model that does not minimize the loss.

Event I (Initial Phase):

- α(t) is small initially
- A stays constant
- α(t) increases until reaching specific threshold
- Only updates v, not attention weights

Event II (Middle Phase):

- α(t) becomes large
- Attention weights start updating
- γ(t) (within-set attention) increases more than ρ(t) (between-set attention)
- Leads to patch association

Event III (Final Phase):

- With patch association established
- α(t) increases again
- Population risk converges to o(1)
- Model fits labeling function

Idealized and realistic learning problems. Given a dataset $\mathcal{Z} = \{(\boldsymbol{X}[i], y[i])\}_{i=1}^{N}$ sampled from $\mathcal{D}$, we solve the empirical risk minimization problem for the logistic loss defined by:

$$\min_{\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{v}}} \quad \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{-y[i] F(\boldsymbol{X}[i])}\right) := \widehat{\mathcal{L}}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{v}}). \tag{E}$$

Instead of directly analyzing (E), we introduce a proxy where we minimize the population risk

$$\min_{\boldsymbol{A}, \boldsymbol{v}} \quad \mathbb{E}_{\mathcal{D}}\left[\log\left(1 + e^{-y F(\boldsymbol{X})}\right)\right] := \mathcal{L}(\boldsymbol{A}, \boldsymbol{v}). \tag{P}$$

We refer to (E) as the *realistic* problem while (P) as the *idealized* problem.

How model learns patch association when optimizing only to minimize loss function

Lemma 4.1 and Lemma 4.2 imply that instead of optimizing over A and v, we can instead consider the scalar variables: $\alpha^{(t)}, \gamma^{(t)}$ and $\rho^{(t)}$

Event I: At this point, the model is nothing else than a generalized linear model that would not generalize because there are much more noisy tokens than signal ones

Main insights:

- because of the initialization and the data structure, we have patch association for any time t.
- simplified ViT model uses patch association to minimize the population loss (Event III). Without patch association, the model would only be a generalized linear model that does not minimize the loss.

Event I (Initial Phase):

- α(t) is small initially
- A stays constant
- α(t) increases until reaching specific threshold
- Only updates v, not attention weights

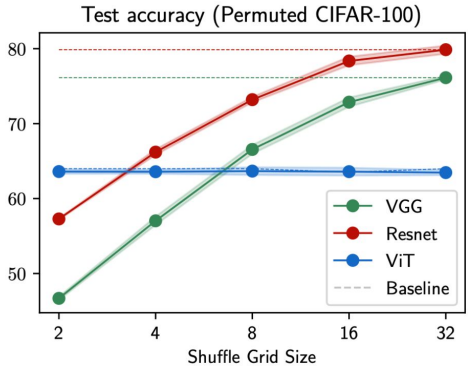
Event II (Middle Phase):

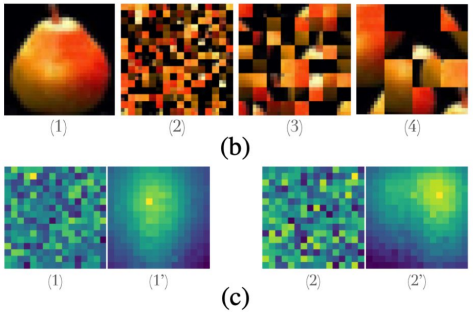
- α(t) becomes large
- Attention weights start updating
- γ(t) (within-set attention) increases more than ρ(t) (between-set attention)
- Leads to patch association

Event III (Final Phase):

- With patch association established
- α(t) increases again
- Population risk converges to o(1)
- Model fits labeling function

Numerical Experiments (1/10 pages)





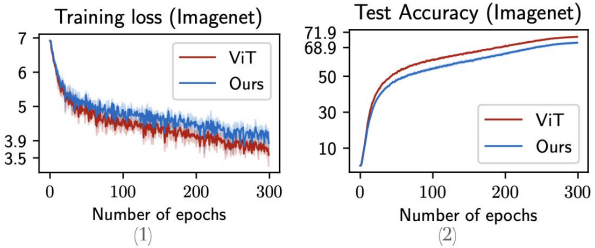


Figure 6: Training loss (1) and test accuracy (2) obtained using a ViT-tiny-patch16-224 on Imagenet. ViT using positional attention (Ours) gets 68.9% test accuracy while vanilla ViT (ViT) gets 71.9%.

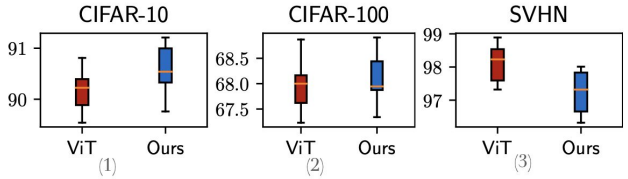
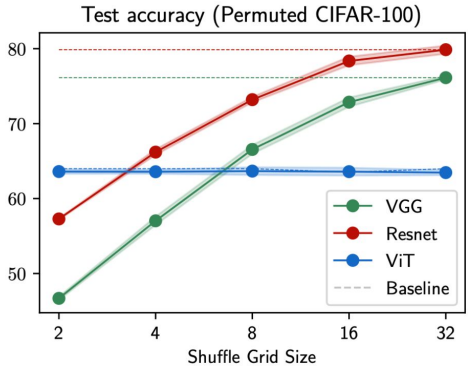
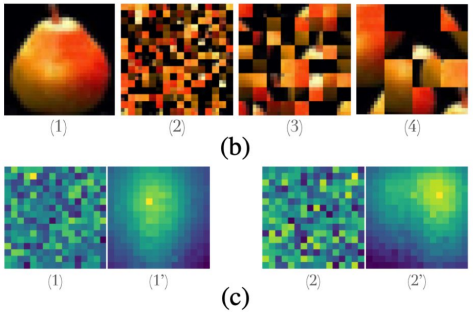


Figure 7: Test accuracy obtained with a ViT using vanilla attention (ViT) and positional attention (Ours) on CIFAR-10 (1), CIFAR-100 (2) and SVHN (3). Our model competes with the vanilla ViT. Patch size 4 and average over 10 seeds for this experiment.

Numerical Experiments (1/10 pages)

For the ViT, we set the patch size to **2**





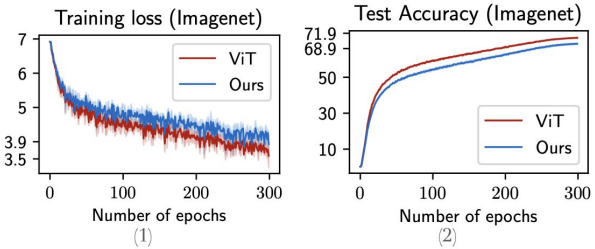


Figure 6: Training loss (1) and test accuracy (2) obtained using a ViT-tiny-patch16-224 on Imagenet. ViT using positional attention (Ours) gets 68.9% test accuracy while vanilla ViT (ViT) gets 71.9%.

Only **71** words were used in this section.

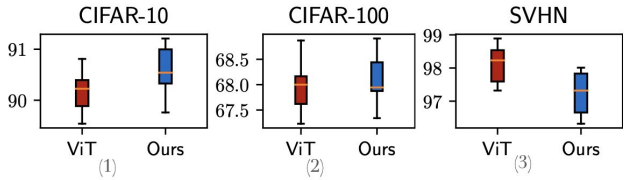


Figure 7: Test accuracy obtained with a ViT using vanilla attention (ViT) and positional attention (Ours) on CIFAR-10 (1), CIFAR-100 (2) and SVHN (3). Our model competes with the vanilla ViT. Patch size 4 and average over 10 seeds for this experiment.

Now presenting:  the AI Podcast Time

Closing Remarks

Strengths:

- Despite the numerous definitions and assumptions, the paper is well written and structured.
- Each mathematical definition is accompanied by clear examples for better understanding.
- The mathematical analysis effectively demonstrates how ViTs learn spatial associations while only optimizing the logistic loss.
- The authors present a well-designed synthetic dataset to answer the main question and provide insightful variations of attention mechanisms to prove their claims.

Closing Remarks

Strengths:

- Despite the numerous definitions and assumptions, the paper is well written and structured.
- Each mathematical definition is accompanied by clear examples for better understanding.
- The mathematical analysis effectively demonstrates how ViTs learn spatial associations while only optimizing the logistic loss.
- The authors present a well-designed synthetic dataset to answer the main question and provide insightful variations of attention mechanisms to prove their claims.

Weaknesses:

- The extreme simplifications raise questions about whether the findings truly apply to standard ViTs.
- While the paper theoretically explains how the model learns spatial associations, it lacks empirical evidence showing this process during the training of the simplified model.
- I agree with some reviewers' comments that: **Final results are not surprising** and **paper appear merely to prove that the phenomenon happens, rather than answering questions about how or why it happens.**

Closing Remarks

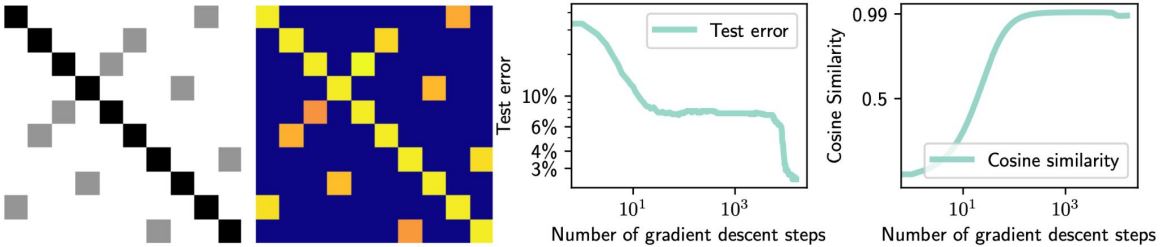


Figure 4: Illustration of Theorem 4.2. We consider the exact same setting (data generation, parameter settings...) as for the realistic case. From left to right, we first display in grey the tuples (i, j) such that $(i, j) \in \mathcal{S}_\ell$. We then plot the learned matrix $\boldsymbol{A}$ and see that coefficients with high value exactly correspond to their grey scale counterpart in the left plot. We also display test error and cosine similarity between $\boldsymbol{w}^*$ and $\boldsymbol{v}$ w.r.t the number of training steps.

Weaknesses:

- The extreme simplifications raise questions about whether the findings truly apply to standard ViTs.
- While the paper theoretically explains how the model learns spatial associations, it lacks empirical evidence showing this process during the training of the simplified model.
- I agree with some reviewers' comments that: **Final results are not surprising** and **paper appear merely to prove that the phenomenon happens, rather than answering questions about how or why it happens.**

Closing Remarks

Strengths:

- Despite the numerous definitions and assumptions, the paper is well written and structured.
- Each mathematical definition is accompanied by clear examples for better understanding.
- The mathematical analysis effectively demonstrates how ViTs learn spatial associations while only optimizing the logistic loss.
- The authors present a well-designed synthetic dataset to answer the main question and provide insightful variations of attention mechanisms to prove their claims.

Weaknesses:

- The extreme simplifications raise questions about whether the findings truly apply to standard ViTs.
- ~~While the paper theoretically explains how the model learns spatial associations, it lacks empirical evidence showing this process during the training of the simplified model.~~ Figure 4 presented suddenly without much information
- I agree with some reviewers' comments that: **Final results are not surprising** and **paper appear merely to prove that the phenomenon happens, rather than answering questions about how or why it happens.**

My Honest Opinion:
This is a valuable paper that approaches ViTs' spatial bias learning from a mathematical perspective. However, while it provides theoretical foundations, it doesn't fully answer the fundamental question or provide clear insights. This paper is just a good paper to cite, if someone is publishing something about Vits inductive bias.

VISION TRANSFORMERS PROVABLY LEARN SPATIAL STRUCTURE

Vladimir Zaigrajew
vladimir.zaigrajew.dokt@pw.edu.pl