From the Topology of Natural Images to Topological CNNs

Alex Sheng

University of Michigan

March 18, 2026

Opening question

Why should **topology** have anything to do with **images**?

Opening question

Why should **topology** have anything to do with **images**?

Maybe there is a common “shape" to all natural images.

Opening question

Why should **topology** have anything to do with **images**?

Maybe there is a common “shape" to all natural images.

Idea: study 3×3 local pixel patches of images with high contrast.

Local pixel patches

- Assuming we are working with grayscale images, then each 3×3 patch can be viewed as a vector $\mathbf{x} = (x_1, \cdots, x_9) \in \mathbb{R}^9$. We have ~ 4.2 million of these.

Local pixel patches

- Assuming we are working with grayscale images, then each 3×3 patch can be viewed as a vector $\mathbf{x} = (x_1, \cdots, x_9) \in \mathbb{R}^9$. We have ~ 4.2 million of these.

- We throw away those patches with low contrast. (For this, we have to define a suitable norm that measures contrast.)

Local pixel patches

- Assuming we are working with grayscale images, then each 3×3 patch can be viewed as a vector $\mathbf{x} = (x_1, \cdots, x_9) \in \mathbb{R}^9$. We have ~ 4.2 million of these.

- We throw away those patches with low contrast. (For this, we have to define a suitable norm that measures contrast.)
- For the rest of the high contrast patches, we subtract the mean and contrast-normalize. Under a suitable coordinate change, the processed data now lies on S^7.

Findings (Lee et al., 2003)

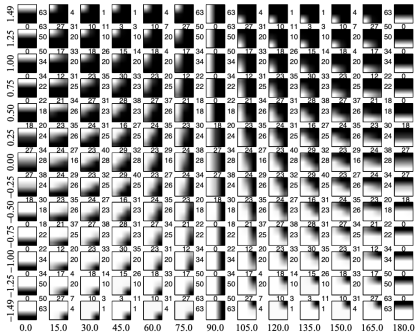
- The distribution of these $\mathbf{x} \in S^7$ is extremely sparse and highly non-uniform!

Findings (Lee et al., 2003)

- The distribution of these $\mathbf{x} \in S^7$ is extremely sparse and highly non-uniform!
- Half of the patches occupy less than 6% of the total volume of S^7.

Findings (Lee et al., 2003)

- The distribution of these $\mathbf{x} \in S^7$ is extremely sparse and highly non-uniform!
- Half of the patches occupy less than 6% of the total volume of S^7.
- They concentrate around a continuous 2D submanifold of **blurred step edges**.



Findings (Lee et al., 2003)

- The estimated density near that manifold behaves like $p(\theta) \sim \theta^{-2.5}$ for small distance θ.

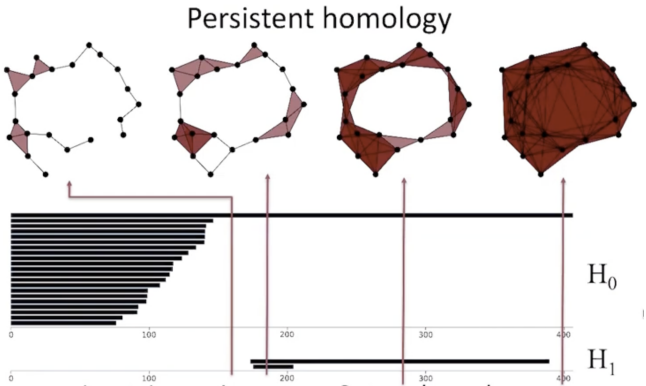
Findings (Lee et al., 2003)

- The estimated density near that manifold behaves like $p(\theta) \sim \theta^{-2.5}$ for small distance θ.
- About 50% of patches lies within 26° of the manifold, and this neighborhood occupies only about 9% of the sphere's volume.

Findings (Lee et al., 2003)

- The estimated density near that manifold behaves like $p(\theta) \sim \theta^{-2.5}$ for small distance θ.
- About 50% of patches lies within 26° of the manifold, and this neighborhood occupies only about 9% of the sphere's volume.
- Next question: *can you say more about this embedded 2D submanifold?*

A quick intro to persistent homology



- Input: Increasing spaces. Output: barcode.
- Significant features persist.
- Cubic computation time in the number of simplices.

Findings (Carlsson et al., 2008)

- Let $X = S^7$. Look at the densest regions $X(k, p)$ of those 3×3 pixel patches.

Findings (Carlsson et al., 2008)

- Let $X = S^7$. Look at the densest regions $X(k, p)$ of those 3×3 pixel patches.
  - For $x \in X$ and $k > 0$, let $\rho_k(x)$ denote the distance from x to its k-th nearest neighbor.

Findings (Carlsson et al., 2008)

- Let $X = S^7$. Look at the densest regions $X(k, p)$ of those 3×3 pixel patches.
  - For $x \in X$ and $k > 0$, let $\rho_k(x)$ denote the distance from x to its k-th nearest neighbor.
  - The function ρ_k is inversely proportional to the density.

Findings (Carlsson et al., 2008)

- Let $X = S^7$. Look at the densest regions $X(k, p)$ of those 3×3 pixel patches.
  - For $x \in X$ and $k > 0$, let $\rho_k(x)$ denote the distance from x to its k-th nearest neighbor.
  - The function ρ_k is inversely proportional to the density.
  - Choose a cutoff parameter p, i.e., a percentage of densest points that we shall use.

Findings (Carlsson et al., 2008)

- Let $X = S^7$. Look at the densest regions $X(k, p)$ of those 3×3 pixel patches.
  - For $x \in X$ and $k > 0$, let $\rho_k(x)$ denote the distance from x to its k-th nearest neighbor.
  - The function ρ_k is inversely proportional to the density.
  - Choose a cutoff parameter p, i.e., a percentage of densest points that we shall use.
- $X(300, 30)$ has first Betti number $b_1 = 1$; $X(15, 30)$ has $b_1 = 5$.

Findings (Carlsson et al., 2008)

- Let $X = S^7$. Look at the densest regions $X(k, p)$ of those 3×3 pixel patches.
  - For $x \in X$ and $k > 0$, let $\rho_k(x)$ denote the distance from x to its k-th nearest neighbor.
  - The function ρ_k is inversely proportional to the density.
  - Choose a cutoff parameter p, i.e., a percentage of densest points that we shall use.
- $X(300, 30)$ has first Betti number $b_1 = 1$; $X(15, 30)$ has $b_1 = 5$.
- $X(100, 10)$ has $b_2 = 1$, i.e., the dense subset is filling out a 2-manifold.

Findings (Carlsson et al., 2008)

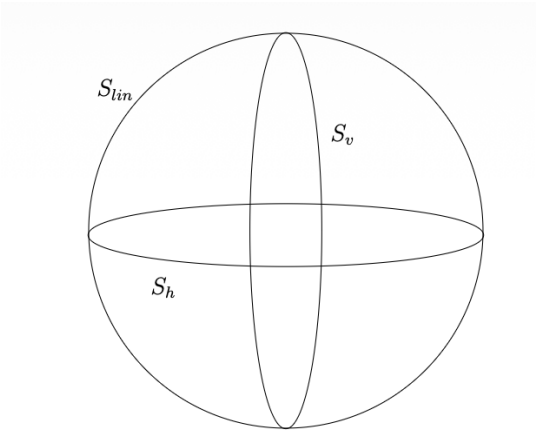
- Let $X = S^7$. Look at the densest regions $X(k, p)$ of those 3×3 pixel patches.
  - For $x \in X$ and $k > 0$, let $\rho_k(x)$ denote the distance from x to its k-th nearest neighbor.
  - The function ρ_k is inversely proportional to the density.
  - Choose a cutoff parameter p, i.e., a percentage of densest points that we shall use.
- $X(300, 30)$ has first Betti number $b_1 = 1$; $X(15, 30)$ has $b_1 = 5$.
- $X(100, 10)$ has $b_2 = 1$, i.e., the dense subset is filling out a 2-manifold.
- But Betti numbers alone does not determine the manifold uniquely...

The three-circle model

- One candidate of $X(15, 30)$, according to its Betti numbers, is a “three-circle" model:

The three-circle model

- One candidate of $X(15, 30)$, according to its Betti numbers, is a “three-circle" model:



Space of polynomials

Key observation

Given a polynomial $p(x, y) \in \mathbb{R}[x, y]$, evaluating it on $\{-1, 0, 1\}^2$ gives an element in $\mathbb{R}^9$, which we can regard as a 3×3 pixel patch.

Space of polynomials

Key observation

Given a polynomial $p(x, y) \in \mathbb{R}[x, y]$, evaluating it on $\{-1, 0, 1\}^2$ gives an element in $\mathbb{R}^9$, which we can regard as a 3×3 pixel patch.

- If $p(x, y) = ax + by$, then they produce *linear gradients*:
  - $p(x, y) = x$: left column dark, middle gray, right column bright;
  - $p(x, y) = y$: bottom dark, middle gray, top bright;
  - $p(x, y) = x + y$: a diagonal ramp.

Space of polynomials

Key observation

Given a polynomial $p(x, y) \in \mathbb{R}[x, y]$, evaluating it on $\{-1, 0, 1\}^2$ gives an element in $\mathbb{R}^9$, which we can regard as a 3×3 pixel patch.

- If $p(x, y) = ax + by$, then they produce *linear gradients*:
  - $p(x, y) = x$: left column dark, middle gray, right column bright;
  - $p(x, y) = y$: bottom dark, middle gray, top bright;
  - $p(x, y) = x + y$: a diagonal ramp.
- If $p(x, y) = cx^2 + bx + a$, then they produce *quadratic variations*:
  - $p(x, y) = x^2$: column changes bright–dark–bright;
  - $p(x, y) = x^2 + \lambda x$ tilts that vertical bar towards one side.

A mathematical “proof"

- Consider the space T of degree-2 polynomials

$$p(x, y) = c(ax + by)^2 + d(ax + by),$$

where $(a, b) \in S^1$ and $(c, d) \in S^1$. Then $T \simeq S^1 \times S^1$.

A mathematical “proof"

- Consider the space T of degree-2 polynomials

$$p(x, y) = c(ax + by)^2 + d(ax + by),$$

  where $(a, b) \in S^1$ and $(c, d) \in S^1$. Then $T \simeq S^1 \times S^1$.
- Evaluating on $\{-1, 0, 1\}^2$ gives a map $T \to \mathbb{R}^9$; subtracting the mean and contrast-normalizing, we get map $T \to S^7$.

A mathematical “proof"

- Consider the space T of degree-2 polynomials

$$p(x, y) = c(ax + by)^2 + d(ax + by),$$

  where $(a, b) \in S^1$ and $(c, d) \in S^1$. Then $T \simeq S^1 \times S^1$.
- Evaluating on $\{-1, 0, 1\}^2$ gives a map $T \to \mathbb{R}^9$; subtracting the mean and contrast-normalizing, we get map $T \to S^7$.
- This polynomial parametrization comes with a symmetry:

$$(a, b, c, d) \sim (a, -b, c, -d).$$

  In angle coordinates (θ, ϕ) of $S^1 \times S^1$, this is becomes $(\theta, \phi) \sim (\theta + \pi, 2\pi - \phi)$.

The Klein bottle

- With these identifications,

$$S^1 \times S^1/((\theta, \phi) \sim (\theta + \pi, 2\pi - \phi))$$

is homeomorphic to K, the Klein bottle. So in fact $T \simeq K$.

The Klein bottle

- With these identifications,

$$S^1 \times S^1/((\theta, \phi) \sim (\theta + \pi, 2\pi - \phi))$$

  is homeomorphic to K, the Klein bottle. So in fact $T \simeq K$.
- They proved that the map $K \to S^7$ is injective, i.e., under this polynomial parametrization, the Klein bottle is embedded as a 2-manifold inside S^7.

The Klein bottle

- With these identifications,

$$S^1 \times S^1/((\theta, \phi) \sim (\theta + \pi, 2\pi - \phi))$$

  is homeomorphic to K, the Klein bottle. So in fact $T \simeq K$.
- They proved that the map $K \to S^7$ is injective, i.e., under this polynomial parametrization, the Klein bottle is embedded as a 2-manifold inside S^7.
- This is a particularly convincing model for that 2-submanifold:
  - The Klein bottle has $\mathbb{Z}/2\mathbb{Z}$-homology equal to the persistent homology.

The Klein bottle

- With these identifications,

$$S^1 \times S^1/((\theta, \phi) \sim (\theta + \pi, 2\pi - \phi))$$

  is homeomorphic to K, the Klein bottle. So in fact $T \simeq K$.
- They proved that the map $K \to S^7$ is injective, i.e., under this polynomial parametrization, the Klein bottle is embedded as a 2-manifold inside S^7.
- This is a particularly convincing model for that 2-submanifold:
  - The Klein bottle has $\mathbb{Z}/2\mathbb{Z}$-homology equal to the persistent homology.
  - The 1-skeleton of a Klein bottle is exactly the “three-circle" figure.

The Klein bottle

- With these identifications,

$$S^1 \times S^1/((\theta, \phi) \sim (\theta + \pi, 2\pi - \phi))$$

  is homeomorphic to K, the Klein bottle. So in fact $T \simeq K$.
- They proved that the map $K \to S^7$ is injective, i.e., under this polynomial parametrization, the Klein bottle is embedded as a 2-manifold inside S^7.
- This is a particularly convincing model for that 2-submanifold:
  - The Klein bottle has $\mathbb{Z}/2\mathbb{Z}$-homology equal to the persistent homology.
  - The 1-skeleton of a Klein bottle is exactly the “three-circle" figure.
  - In other words, the Klein bottle model fits the experimental observations.

Carlsson et al., 2008

Main takeaway

If we look at high contrast 3×3 pixel patches through a adjustable density filter $X(k, p)$, then when the density resolution becomes finer (meaning k and p decreases), we see a primary circle (linear gradient), then two secondary circles (quadratic variation), and finally, a 2-dimensional hole.

The Klein bottle is a particularly convincing and mathematically sound model for the 2-submanifold of dense patches, sitting inside S^7.

Aside: biological observations

- Singh et al. (2008) studied multi-neuron recordings in the primary visual cortex (V1) of macaques.

Aside: biological observations

- Singh et al. (2008) studied multi-neuron recordings in the primary visual cortex (V1) of macaques.
- They also found that the population activity of neurons is not random in the high-dimensional space.

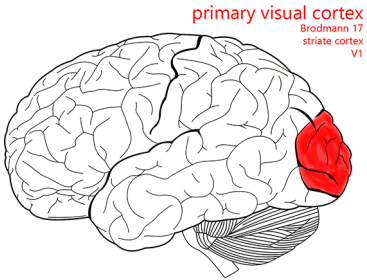
Aside: biological observations

- Singh et al. (2008) studied multi-neuron recordings in the primary visual cortex (V1) of macaques.
- They also found that the population activity of neurons is not random in the high-dimensional space.
- Rather, it appears to concentrates on a structured low-dimensional manifold, most strongly resembling a sphere with a strong circle component.

Aside: biological observations

- Singh et al. (2008) studied multi-neuron recordings in the primary visual cortex (V1) of macaques.
- They also found that the population activity of neurons is not random in the high-dimensional space.
- Rather, it appears to concentrates on a structured low-dimensional manifold, most strongly resembling a sphere with a strong circle component.
- They also used persistent homology in their analysis.

A cute picture of macaque



CNNs

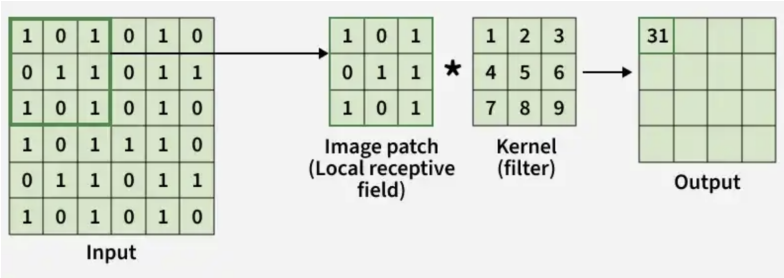
- Both images and visual neuron activities suggest that there is an underlying geometry.

CNNs

- Both images and visual neuron activities suggest that there is an underlying geometry.
- We now turn our attention to convolutional neural networks.

CNNs

- Both images and visual neuron activities suggest that there is an underlying geometry.
- We now turn our attention to convolutional neural networks.
- In a convolutional layer, there are filters (kernels) that scan through an input from the previous layer. The weights on these filters are to be learned.



Carlsson-Gabrielsson (2018)

- Using 3×3 filters, we can look at the space of all such filters in a *trained* CNN. Like before, each filter gives a vector $\mathbf{x} \in \mathbb{R}^9$, with which we perform a density filtration.

Carlsson-Gabrielsson (2018)

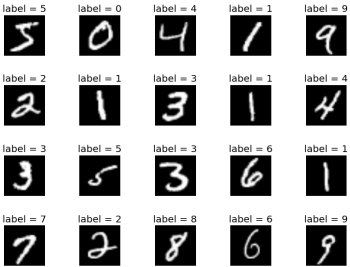
- Using 3×3 filters, we can look at the space of all such filters in a *trained* CNN. Like before, each filter gives a vector $\mathbf{x} \in \mathbb{R}^9$, with which we perform a density filtration.
- Compute the persistent homology of the thus formed point cloud in $\mathbb{R}^9$.

Carlsson-Gabrielsson (2018)

- Using 3×3 filters, we can look at the space of all such filters in a *trained* CNN. Like before, each filter gives a vector $\mathbf{x} \in \mathbb{R}^9$, with which we perform a density filtration.
- Compute the persistent homology of the thus formed point cloud in $\mathbb{R}^9$.
- There is another technique in TDA, called the Mapper algorithm, which allows us to visualize the graph-shaped summary of a point cloud.

MNIST result

- On MNIST, the densest regions of the space of first convolutional layer filters extracted from 100 separately trained CNNs form a **primary circle** of “edge detectors".
- The second layer filters is less straighforward.



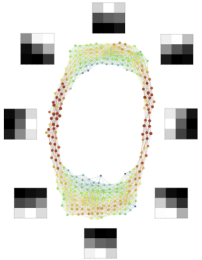


Figure 3: MNIST layer 1

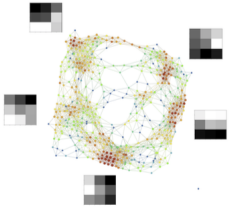
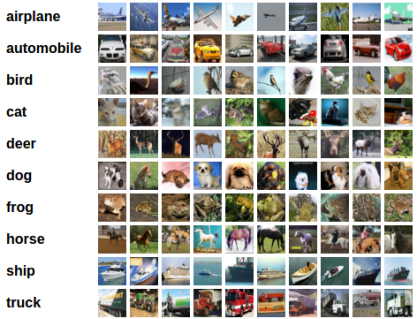


Figure 5: MNIST layer 2

CIFAR-10 results

- On CIFAR-10, first layer, they observed a strong primary circle together with two secondary circles, reminiscent of the three-circles model.
- The persistent homology computation confirms that.



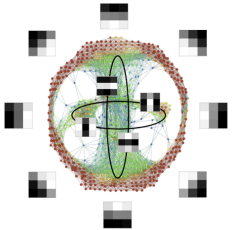


Figure 7: First layer, CIFAR-10, separate colors

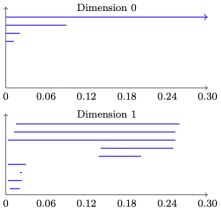


Figure 8: Persistence barcode, Figure 7

VGG16 results

- VGG16 is a well-known pretrained CNN, trained on Imagenet, a large image database. It has 13 convolutional layers.
- First two layers give exactly a primary circle. As we go deeper, more complex patterns start to evolve, abstracting finer motifs about an image.

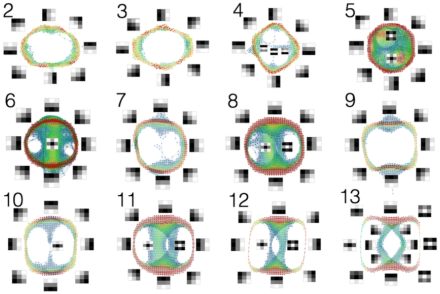


Figure 10: VGG16

Training dynamics

- The topology of the space of filters evolves during learning.
- Simple topological structure emerges during training, changes over time, and can migrate across layers

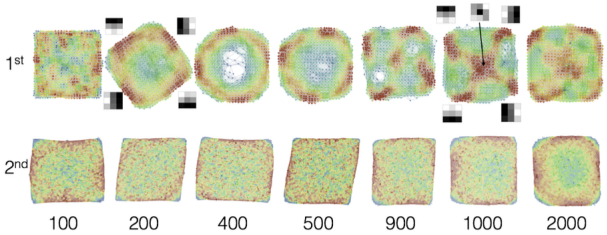


Figure 9: CIFAR-10 learning

Topology and the ability to generalize

- A network trained on SVHN generalizes much better to MNIST than vice versa.
- Meanwhile, the primary circle learned on SVHN is “stronger" than the one learned on MNIST, as measured by persistence, i.e., lifetime of the corresponding homology class.
- Moreover, they show a a direct correlation between the persistence of the strongest 1-dimensional homology class in the first-layer filters and the network’s test accuracy.

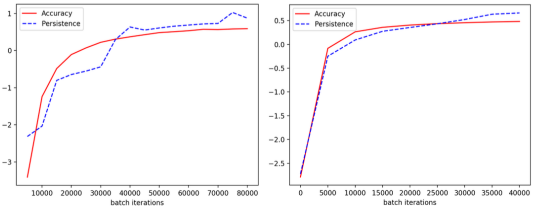


Fig. 12: Test Acc. and Persistence. Left: MNIST, right: SVHN

Topology as inductive bias

- When a network's weight space organizes into a simple, strong topological model, that simplicity may be evidence that the network has captured a more general hypothesis about the task rather than memorized quirks of one dataset.

Topology as inductive bias

- When a network's weight space organizes into a simple, strong topological model, that simplicity may be evidence that the network has captured a more general hypothesis about the task rather than memorized quirks of one dataset.
- This motivates us to treat topology as a useful **inductive bias** in CNN design!

Topology as inductive bias

- When a network's weight space organizes into a simple, strong topological model, that simplicity may be evidence that the network has captured a more general hypothesis about the task rather than memorized quirks of one dataset.
- This motivates us to treat topology as a useful **inductive bias** in CNN design!
- Next-up question: *can we systematically build natural image topology into CNNs?*

Topological CNNs

- The idea is to construct filters using the circle or Klein bottle geometry.

Topological CNNs

- The idea is to construct filters using the circle or Klein bottle geometry.
- For $M = S^1$, define a family polynomials in two variables:

$$F_{S^1}(\theta)(x, y) = x \cos\theta + y \sin\theta, \quad \theta \in S^1.$$

Topological CNNs

- The idea is to construct filters using the circle or Klein bottle geometry.
- For $M = S^1$, define a family polynomials in two variables:

$$F_{S^1}(\theta)(x, y) = x\cos\theta + y\sin\theta, \quad \theta \in S^1.$$

- For $M = K$, define the family to be

$$F_K(\theta_1, \theta_2)(x, y) = \sin(\theta_2)(x\cos\theta_1 + y\sin\theta_1) + \cos(\theta_2)\, Q(x\cos\theta_1 + y\sin\theta_1)$$

for $(\theta_1, \theta_2) \in K \simeq S^1 \times S^2/\sim$, where $Q(t) = 2t^2 - 1$.

Topological CNNs

- The idea is to construct filters using the circle or Klein bottle geometry.
- For $M = S^1$, define a family polynomials in two variables:

$$F_{S^1}(\theta)(x, y) = x\cos\theta + y\sin\theta, \quad \theta \in S^1.$$

- For $M = K$, define the family to be

$$F_K(\theta_1, \theta_2)(x, y) = \sin(\theta_2)(x\cos\theta_1 + y\sin\theta_1) + \cos(\theta_2)\, Q(x\cos\theta_1 + y\sin\theta_1)$$

  for $(\theta_1, \theta_2) \in K \simeq S^1 \times S^2/\sim$, where $Q(t) = 2t^2 - 1$.
- These are algebraic models for the family of local patches that lie on S^1 and K.

Constructing filters

- Now, we can construct filters of size $k \times k$ from this parametrization, as follow.

Constructing filters

- Now, we can construct filters of size $k \times k$ from this parametrization, as follow.
- Divide $[-1, 1]^2$ into k^2 grids $\mathrm{grid}(n, m)$.

Constructing filters

- Now, we can construct filters of size $k \times k$ from this parametrization, as follow.
- Divide $[-1, 1]^2$ into k^2 grids $\mathrm{grid}(n, m)$.
- For a point $\kappa \in M$, the filter associated to it is obtained by grid-averaging:

$$\mathrm{Filter}_\kappa(n, m) = \int_{\mathrm{grid}(n,m)} F_M(\kappa)(x, y)\, dx\, dy.$$

So every point on S^1 or K becomes one concrete filter.

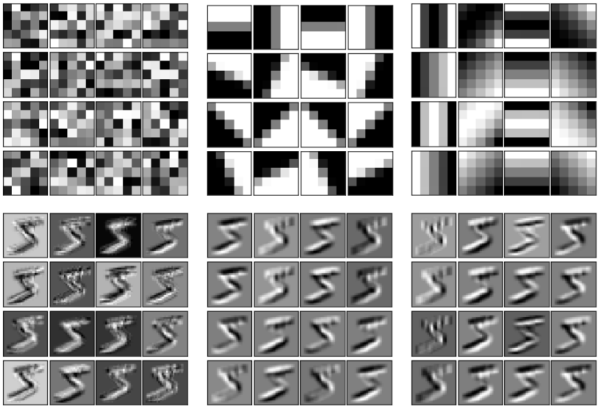
Constructing filters

- Now, we can construct filters of size $k \times k$ from this parametrization, as follow.
- Divide $[-1, 1]^2$ into k^2 grids $\mathrm{grid}(n, m)$.
- For a point $\kappa \in M$, the filter associated to it is obtained by grid-averaging:

$$\mathrm{Filter}_\kappa(n, m) = \int_{\mathrm{grid}(n,m)} F_M(\kappa)(x, y)\, dx\, dy.$$

  So every point on S^1 or K becomes one concrete filter.
- In implementation, we pick a discretization of S^1 or K, and index the output channels by these points. These filters, called *circle filters* or *Klein bottle filters*, are frozen during training. Instead of letting the network to learn these weights, we initialize the network with these weights!

Interpretability



(a) Normal (b) CF (c) KF

COL and KOL layers

- We can also inform trainable convolutional layers of topology.

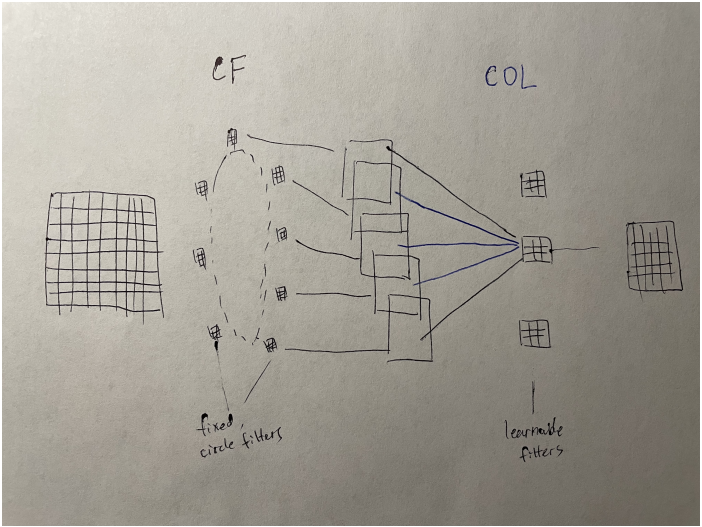
COL and KOL layers

- We can also inform trainable convolutional layers of topology.
- Instead of allowing contributions from all channels in the previous layer, we impose a tunable threshold hyperparameter δ that ignores contribution from channels that are further than δ away (in the appropriate metric) channels.

COL and KOL layers

- We can also inform trainable convolutional layers of topology.
- Instead of allowing contributions from all channels in the previous layer, we impose a tunable threshold hyperparameter δ that ignores contribution from channels that are further than δ away (in the appropriate metric) channels.
- These are called *circle one layer* (COL) and *Klein one layer* (KOL), respectively.

An illustration of CF+COL



Philosophy behind COL and KOL

- In other words, we are aggregating only “topologically nearby" feature channels.

Philosophy behind COL and KOL

- In other words, we are aggregating only “topologically nearby" feature channels.
- In some sense, we are introducing an additional notion of *locality*:
  - A standard convolutional layer says: nearby pixels should interact, distant pixels should not.
  - A COL/KOL says: in addition to that, near *feature types* should interact, distant feature types should not.

Philosophy behind COL and KOL

- In other words, we are aggregating only “topologically nearby" feature channels.
- In some sense, we are introducing an additional notion of *locality*:
  - A standard convolutional layer says: nearby pixels should interact, distant pixels should not.
  - A COL/KOL says: in addition to that, near *feature types* should interact, distant feature types should not.
- In so doing, we are accurately forcing the network to learn “important" features: edges and edge orientations (linear gradients), and quadratic variation patterns.

Robustness to Gaussian noise

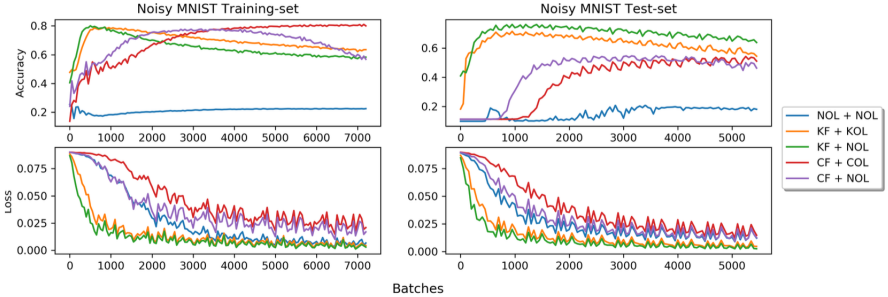


Figure 4: Two synthetic experiments on noisy MNIST data, designed to test model assumptions. The first column displays the results of the experiment where Gaussian noise is added to the training data but not the testing data. The second column displays the results of the experiment where the training data is the original MNIST data and the testing data are corrupted by Gaussian noise. The first row is testing accuracy and the second row is training loss.

Rate of learning

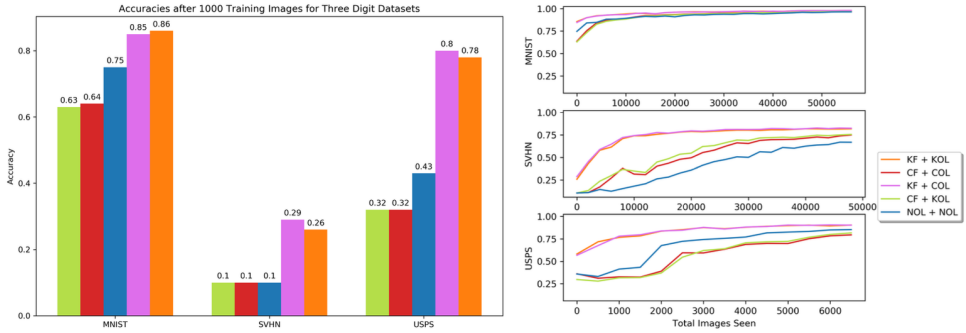
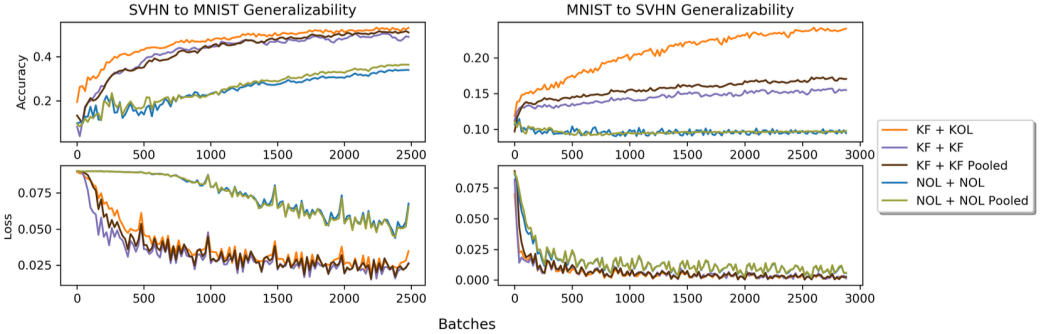
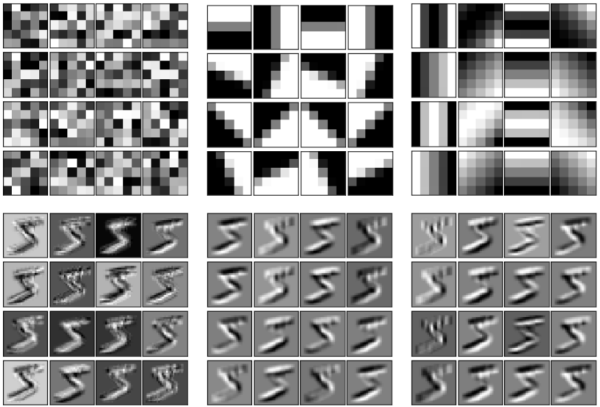


Figure 6: Left: Comparisons of testing accuracy after training on 1,000 images. Right: Full comparison of testing accuracy (y-axes) over a single epoch. MNIST was trained on 60,000, SVHN on 50,032, and USPS on 7291 images.

Generalizability



Interpretability



(a) Normal (b) CF (c) KF

Two comments

- The authors also employed an analogous methodology for video data. This is a bit more complicated, as it involves looking at the tangent bundle of K.
- I have been trying to reproduce some results myself, but have had little success so far.

Summary and takeaway

High contrast local patches of natural images concentrate on low-dimensional manifolds.

CNNs trained on images rediscover these manifolds in their learned first layer filters.

We can leverage this latent topological structure to build better CNNs.

Thank you!

Thank you for listening!

Bibliography

1. Lee et al. The Nonlinear Statistics of High-Contrast Patches in Natural Images. *International Journal of Computer Vision* 54, no. 1–3 (2003): 83–103.
2. Carlsson et al. On the Local Behavior of Spaces of Natural Images. *International Journal of Computer Vision* 76, no. 1 (2008): 1–12.
3. Singh et al. Topological Analysis of Population Activity in Visual Cortex. *Journal of Vision* 8, no. 8 (2008): 11:1–18. doi:10.1167/8.8.11.
4. Rickard Brüel Gabrielsson and Gunnar Carlsson. Topological Approaches to Deep Learning. In *Topological Data Analysis*, 119–146. Springer International Publishing, 2020.
5. Love et al. Topological Convolutional Layers for Deep Learning. *Journal of Machine Learning Research* 24 (2023): 1–35.