WHAT DO VISION TRANSFORMERS LEARN? A VISUAL EXPLORATION

Presented by Krishna Chaitanya Paladugu
23rd Feb, 2023

UNIVERSITY OF
GEORGIA

Outline

- Abstract
- Introduction
- VIT Feature Visualization
- Last-Layer Token Mixing
- Comparison of VITs and CNNs
- VITs with Language Model Supervision
- Discussion

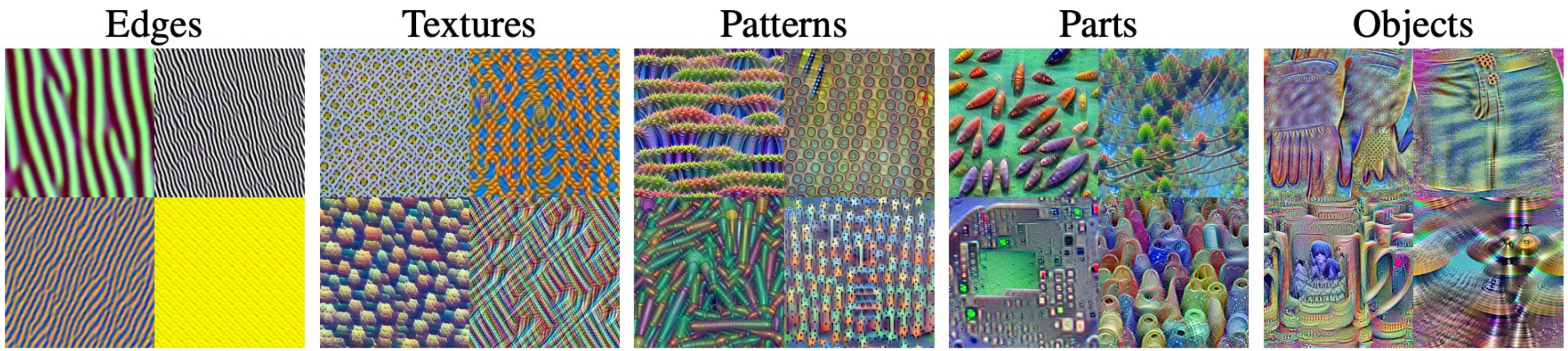
Abstract

- Vision transformers (ViTs) are quickly becoming the de-facto architecture for computer vision, yet we understand very little about why they work and what they learn.
- They observe that neurons in ViTs trained with language model supervision (e.g., CLIP) are activated by semantic concepts rather than visual features.
- They also explore the underlying differences between ViTs and CNNs, and find that transformers detect image background features, just like their convolutional counterparts, but their predictions depend far less on high-frequency information.
- On the other hand, both architecture types behave similarly in the way features progress from abstract patterns in early layers to concrete objects in late layers. In addition, they show that ViTs maintain spatial information in all layers except the final layer.
- Finally, they conduct large-scale visualizations on a wide range of ViT variants, including DeiT, CoaT, ConViT, PiT, Swin, and Twin, to validate the effectiveness of their method.

Introduction

- Recent years have seen the rapid proliferation of vision transformers (ViTs) across a diverse range of tasks from image classification to semantic segmentation to object detection

- They show that if properly applied to the correct representations, feature visualizations can indeed succeed on VITs.



The progression for visualized features of ViT B-32. Features from early layers capture general edges and textures.

Moving into deeper layers, features evolve to capture more specialized image components and finally concrete objects.

Introduction

- By dissecting and visualizing the internal representations in the transformer architecture, they found that patch tokens preserve spatial information throughout all layers except the last attention block. The last layer of ViTs learns a token-mixing operation akin to average pooling, such that the classification head exhibits comparable accuracy when ingesting a random token instead of the CLS token.

They find that language supervision in CLIP (Radford et al., 2021) results in neurons that respond to complex abstract concepts. This includes even a “death neuron” that responds to the abstract concept of morbidity.

Introduction

- When performing activation maximizing visualizations, we notice that ViTs consistently generate higher quality image backgrounds than CNNs. Thus, they try masking out image foregrounds during inference, and find that ViTs consistently outperform CNNs when exposed only to image backgrounds.

- Additionally, convolutional neural networks are known to rely heavily on high-frequency texture information in images. In contrast, we find that ViTs perform well even when high-frequency content is removed from their inputs.

The contributions are:

- They observe applying standard methods of feature visualization to the relatively high-dimensional features of the position-wise feedforward layer results in successful and informative visualizations.

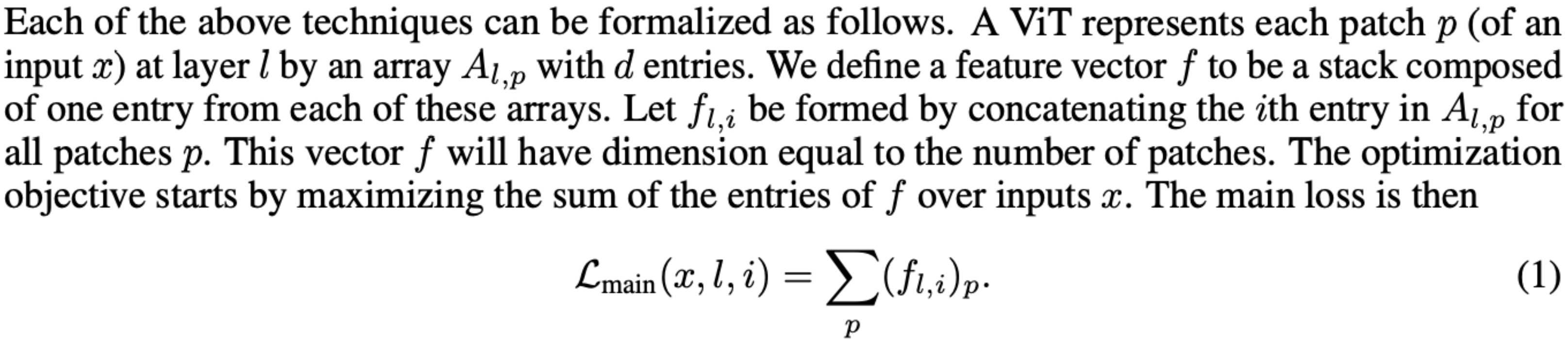
- They show that patch-wise image activation patterns for ViT features essentially behave like saliency maps, highlighting the regions of the image a given feature attends to.

- They compare the behavior of ViTs and CNNs, finding that ViTs make better use of background information and rely less on high-frequency, textural attributes.

- They investigate the effect of natural language supervision with CLIP on the types of features extracted by ViTs.

VIT Feature Visualization

- Like many visualization techniques, They take gradient steps to maximize feature activations starting from random noise
- To improve the quality of the images, they penalize total variation and also employ the Jitter augmentation, the ColorShift augmentation, and augmentation ensembling.
- Finally, find that Gaussian smoothing facilitates better visualization in our experiments as is common in feature visualization. Each of these techniques can be formalized as follows:



Each of the above techniques can be formalized as follows. A ViT represents each patch p (of an input x) at layer l by an array $A_{l,p}$ with d entries. We define a feature vector f to be a stack composed of one entry from each of these arrays. Let $f_{l,i}$ be formed by concatenating the ith entry in $A_{l,p}$ for all patches p. This vector f will have dimension equal to the number of patches. The optimization objective starts by maximizing the sum of the entries of f over inputs x. The main loss is then

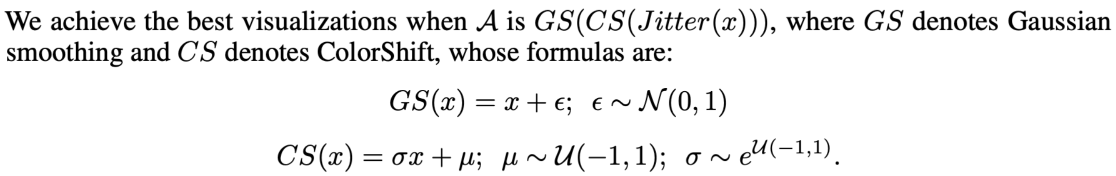
$$\mathcal{L}_{\text{main}}(x, l, i) = \sum_p (f_{l,i})_p. \tag{1}$$

VIT Feature Visualization

- The feature vector f is defined as a stack of entries from the arrays that represent each patch of an input x at layer l. The optimization objective starts by maximizing the sum of the entries of f over inputs x.
- To improve the quality of the images, the objective includes total variation regularization by adding the term λT V (x) to the objective. T V represents the total variation, and λ is the hyperparameter controlling the strength of its regularization effect.

$$x^* = \arg\max_x \sum_k \mathcal{L}_{\text{main}}(a_k(x), l, i) + \lambda TV(a_k(x)).$$

VIT Feature Visualization



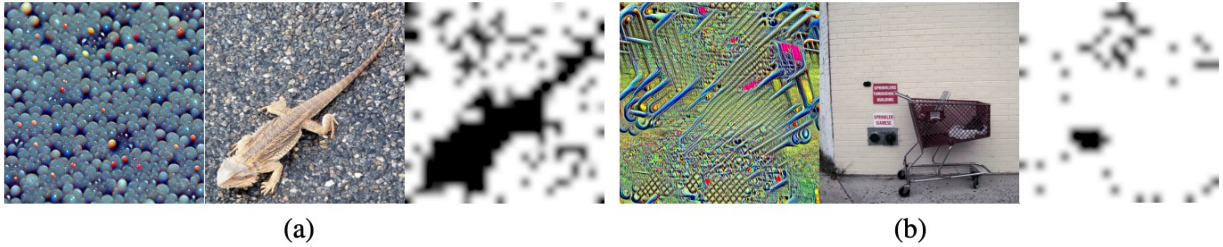
We achieve the best visualizations when $\mathcal{A}$ is $GS(CS(Jitter(x)))$, where GS denotes Gaussian smoothing and CS denotes ColorShift, whose formulas are:

$$GS(x) = x + \epsilon; \;\; \epsilon \sim \mathcal{N}(0,1)$$

$$CS(x) = \sigma x + \mu; \;\; \mu \sim \mathcal{U}(-1,1); \;\; \sigma \sim e^{\mathcal{U}(-1,1)}.$$

- To better understand the content of a visualized feature, they pair every visualization with images from the ImageNet validation/train set that most strongly activate the relevant feature.
- Moreover, they plot the feature’s activation pattern by passing the most activating images through the network and showing the resulting pattern of feature activations

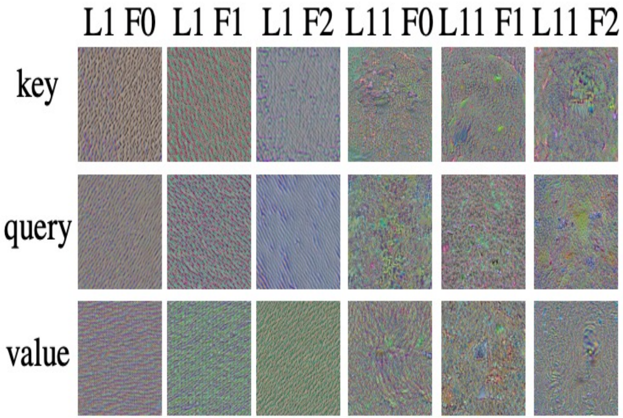
VIT Feature Visualization



- From the leftmost panel, they hypothesize that this feature corresponds to gravel.
- The most activating image from the validation set (middle) contains a lizard on a pebbly gravel road. Interestingly, the gravel background lights up in the activation pattern (right), while the lizard does not.
- The activation pattern in this example behaves like a saliency map

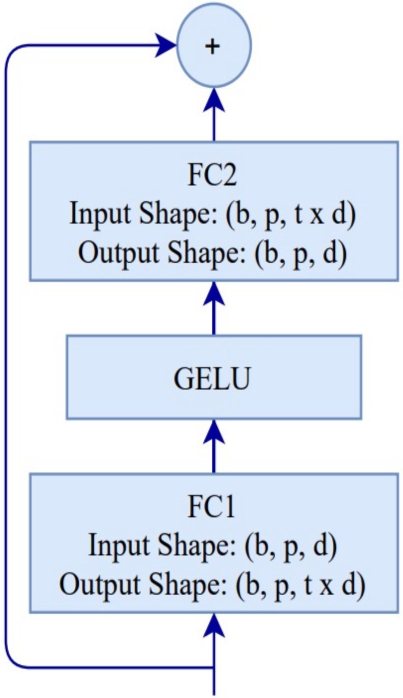
VIT Feature Visualization

- ViT-B16 is composed of 12 blocks, each consisting of multi-headed attention layers, followed by a projection layer for mixing attention heads, and finally followed by a position-wise-feed-forward layer.
- In this model, every patch is always represented by a vector of size 768 except in the feed-forward layer which has a size of 3072 (4 times larger than other layers).
- They first attempt to visualize features of the multi-headed attention layer, including visualization of the keys, queries, and values, by performing activation maximization. They find that the visualized feedforward features are significantly more interpretable than other layers.



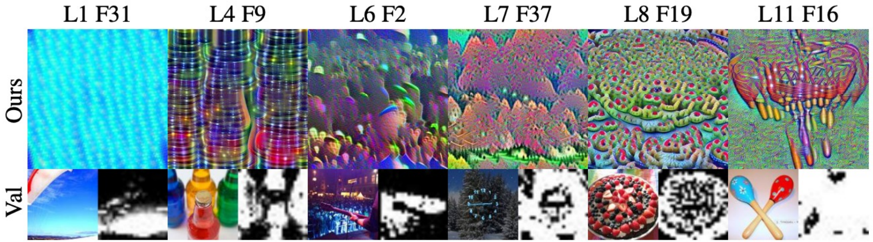
VIT Feature Visualization

- The feed-forward layer depicted in Figure takes an input of size d = 768, projects it into a t = 4 times higher dimensional space, applies the non-linearity GELU, and then projects back to d dimensional space.
- Unless otherwise stated, we always visualize the output of the GELU layers in our experiments. They hypothesize that the network exploits these high-dimensional spaces to store relatively disentangled representations.
- On the other hand, compressing the features into a lower dimensional space may result in the jumbling of features, yielding uninterpretable visualizations.



Last-Layer Token Mixing

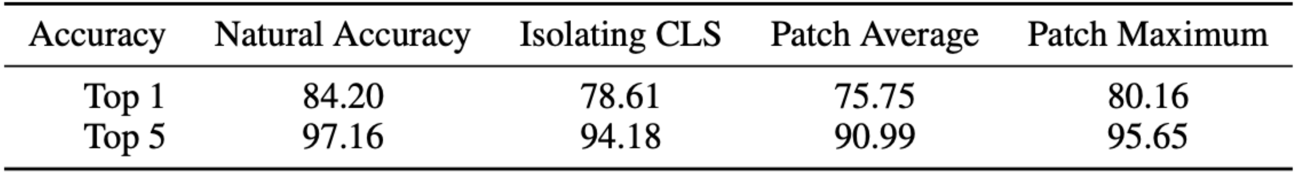
- In this section, they investigate the preservation of patch-wise spatial information observed in the visualizations of patch-wise feature activation levels which, as noted before, bear some similarity to saliency maps.



- Here the activation maps approximately segment the image with respect to some relevant aspect of the image.
- ViTs learn to preserve spatial information, despite lacking the inductive bias of CNNs.
- For classification purposes, ViTs use a fully connected layer applied only on the class token (the CLS token). It is possible that the network globalizes information in the last layer to ensure that the CLS token has access to the entire image, but because the CLS token is treated the same as every other patch by the transformer, this seems to be achieved by globalizing across all tokens.

Last-Layer Token Mixing

- After the last layer, every patch contains the same information. “Isolating CLS” denotes the experiment where attention is only performed between patches before the final attention block, while “Patch Average” and “Patch Maximum” refer to the experiment in which the classification head is placed on top of individual patches without fine-tuning. Experiments conducted on ViT-B16.

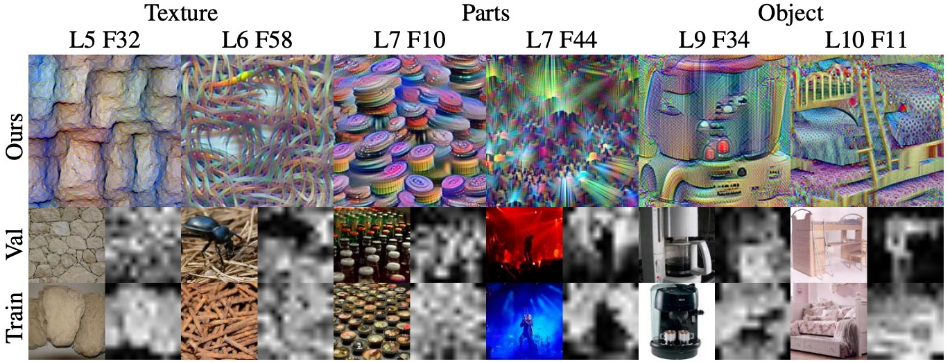


Accuracy	Natural Accuracy	Isolating CLS	Patch Average	Patch Maximum
Top 1	84.20	78.61	75.75	80.16
Top 5	97.16	94.18	90.99	95.65

- The resulting hacked network that only has CLS access in the last layer can still successfully classify 78.61% of the ImageNet validation set as shown in Table. From this result, they conclude that the CLS token captures global information mostly at the last layer, rather than building a global representation throughout the network.

Last-Layer Token Mixing

- Complexity of features vs depth in ViT B-32. Visualizations suggest that ViTs are similar to CNNs in that they show a feature progression from textures to parts to objects as we progress from shallow to deep features.



- We conclude by noting that the information structure of a ViT is remarkably similar to a CNN, in the sense that the information is positionally encoded and preserved until the final layer. Furthermore, the final layer in ViTs appears to behave as a learned global pooling operation that aggregates information from all patches, which is similar to its explicit averagepooling counterpart in CNNs.

Comparison Of VITs and CNNs

- An important observation is that in CNNs, early layers recognize color, edges, and texture, while deeper layers pick out increasingly complex structures eventually leading to entire objects. ViTs exhibit this kind of progressive specialization as well.
- On the other hand, they observe that there are also important differences between the ways CNNs and ViTs recognize images.

(a) (b)

Comparison Of VITs and CNNs

- ViTs more effectively correlate background information with correct class. Both foreground and background data are normalized by full image top-5 accuracy.

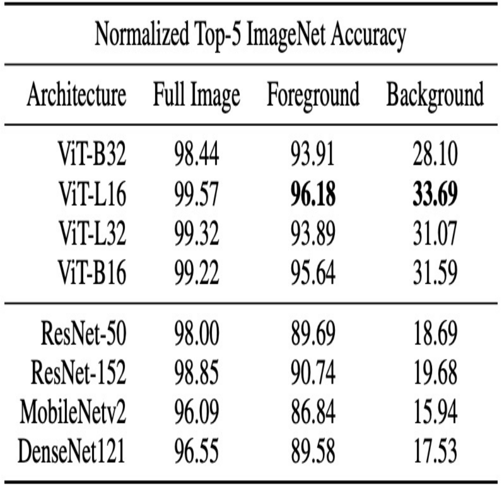
- To quantitatively assess each architecture's dependence on different parts of the image on the dataset level, they mask out the foreground or background on a set of evaluation images using the aforementioned ImageNet bounding boxes, and they measure the resulting change in top-5 accuracy.

- These tests are performed across a number of pretrained ViT models, and they compared to a set of common CNNs in Table

1785

Comparison Of VITs and CNNs

- They observe that ViTs are significantly better than CNNs at using the background information in an image to identify the correct class. At the same time, ViTs also suffer noticeably less from the removal of the background, and thus seem to depend less on the background information to make their classification.

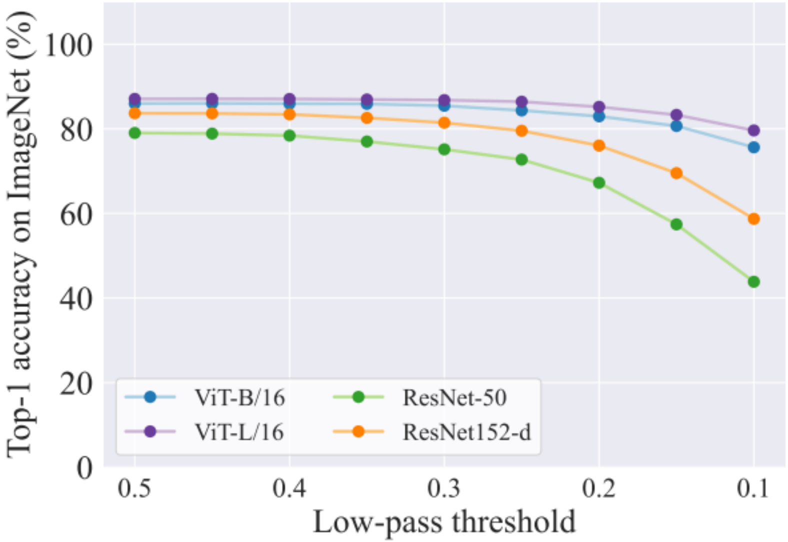


Normalized Top-5 ImageNet Accuracy

Architecture	Full Image	Foreground	Background
ViT-B32	98.44	93.91	28.10
ViT-L16	99.57	**96.18**	**33.69**
ViT-L32	99.32	93.89	31.07
ViT-B16	99.22	95.64	31.59
ResNet-50	98.00	89.69	18.69
ResNet-152	98.85	90.74	19.68
MobileNetv2	96.09	86.84	15.94
DenseNet121	96.55	89.58	17.53

Comparison Of VITs and CNNs

- Effect of low pass filtering:

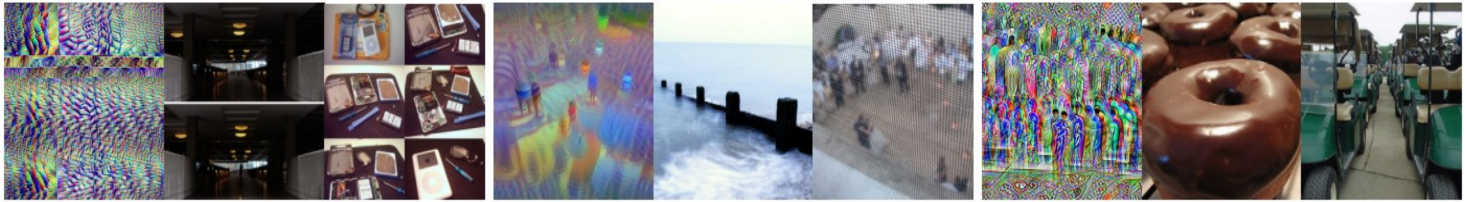


VITs With Language Model Supervision

- Recently, ViTs have been used as a backbone to develop image classifiers trained with natural language supervision and contrastive learning techniques
- The CLIP models are state-of-the-art in transfer learning to unseen datasets.
- The zero-shot ImageNet accuracy of these models is even competitive with traditionally trained ResNet-50 competitors.
- They compare the feature visualizations for ViT models with and without CLIP training to study the effect of natural language supervision on the behavior of the transformer-based backbone

VITs With Language Model Supervision

- Figure (a) shows the image optimized to maximally activate a feature in the fifth layer of a ViT CLIP model alongside its two highest activating examples from the ImageNet dataset.
- The fact that all three images share sharp boundaries indicates this feature might be responsible for detecting caption texts relating to a progression of images.



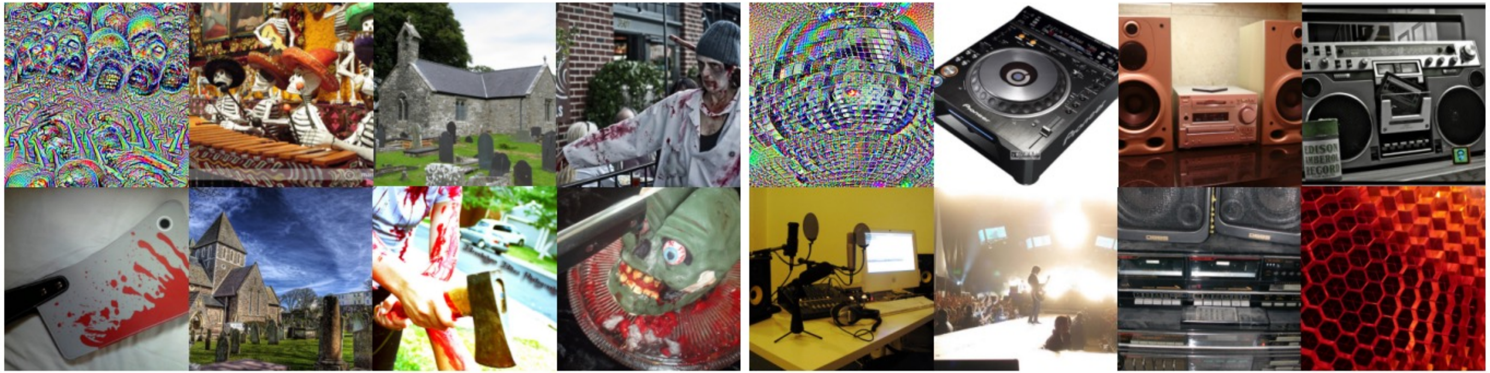
(a) Before and after/Step-by-step

(b) From above

(c) Many

VITs With Language Model Supervision

- The presence of features that represent conceptual categories is another consequence of CLIP training. Unlike ViTs trained as classifiers, in which features detect single objects or common background information, CLIP-trained ViTs produce features in deeper layers activated by objects in clearly discernible conceptual categories.



(a) Category of morbidity (b) Category of music

VITs With Language Model Supervision

- Given that the space of possible captions for images is substantially larger than the mere one thousand classes in the ImageNet dataset, high performing CLIP models understandably require higher level organization for the objects they recognize.

- Moreover, the CLIP dataset is scraped from the internet, where captions are often more descriptive than simple class labels.

Discussion

- In order to dissect the inner workings of vision transformers, we introduce a framework for optimization-based feature visualization.
- They found that the high-dimensional inner projection of the feed-forward layer is suitable for producing interpretable images, while the key, query, and value features of self-attention are not.
- After identifying the suitable components, they applied the framework to analyze how spatial information is preserved in ViTs. They found that ViTs preserve spatial information of the patches even for individual channels across all layers, except for the last layer.
- This suggests that the networks learn spatial relationships from scratch.
- Finally, they showed that the sudden disappearance of localization information in the last attention layer is due to a learned token mixing behavior that resembles average pooling. In other words, the last layer of the ViT already performs a mixing operation, making global pooling unnecessary.

Discussion

- In comparing CNNs and ViTs, we find that ViTs make better use of background information and are able to make vastly superior predictions relative to CNNs when exposed only to image backgrounds despite the seemingly counter-intuitive property that ViTs are not as sensitive as CNNs to the loss of high-frequency information, which one might expect to be critical for making effective use of background.

- We also conclude that the two architectures share a common property whereby earlier layers learn textural attributes, whereas deeper layers learn high level object features or abstract concepts. Finally, they show that ViTs trained with language model supervision learn more semantic and conceptual features, rather than object-specific visual features as is typical of classifiers.

Thank You

UNIVERSITY OF GEORGIA

1785 ™