What do Vision Transformers Learn? A Visual Exploration

Virgilio Strozzi

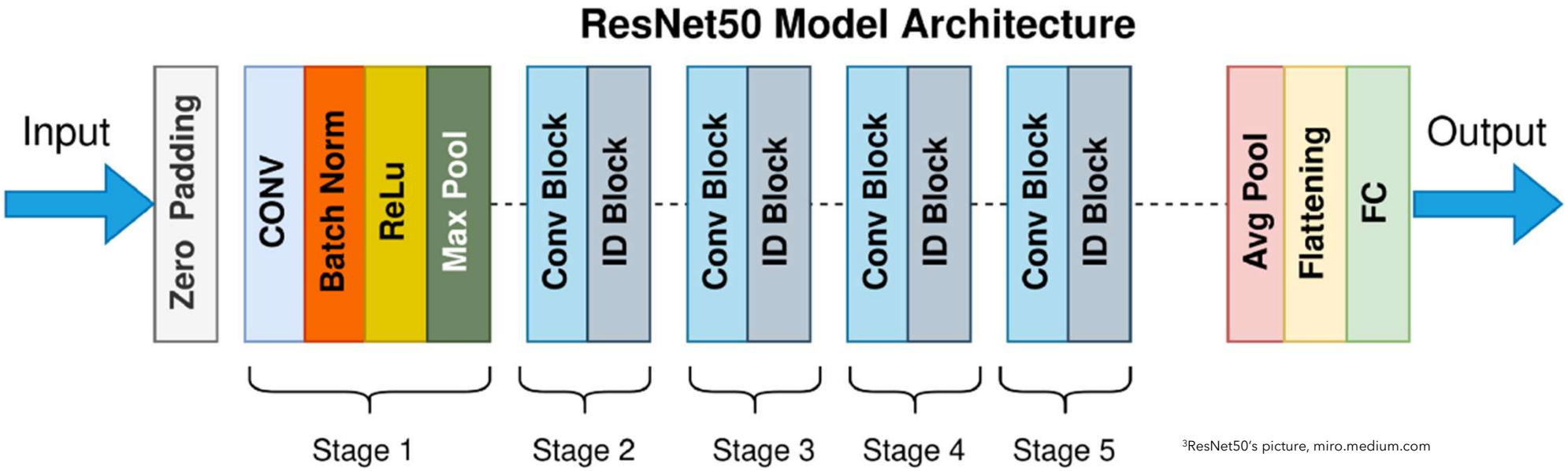
07.03.2023

Seminar

in Deep Neural Networks

Computer Vision: Image Classification

- Dominant architecture in Image Classification (before 2021)?
  - ResNet

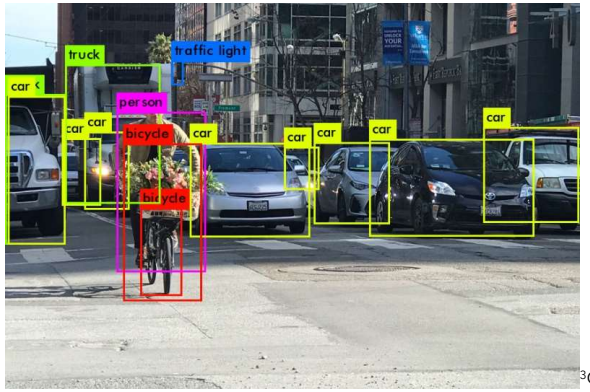


[3]ResNet50's picture, miro.medium.com

Computer Vision: Image Classification

- Advent of the Paper: *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (October 2020)*[2]
  - *New competition for CNN and RNN with Attention in CV*
  - *Transformers without CNN are good enough*

- Vast proliferation and state-of-the-art results in different tasks:

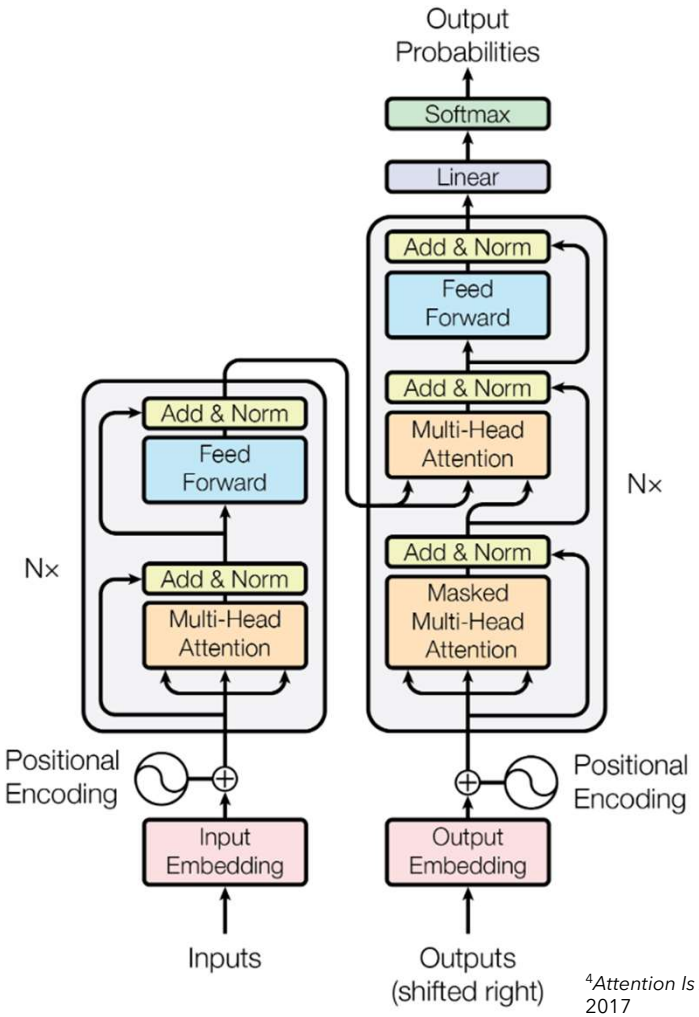


[3]Object Detection's picture. miro.medium.com

[3]Image Sementation's picture. miro.medium.com

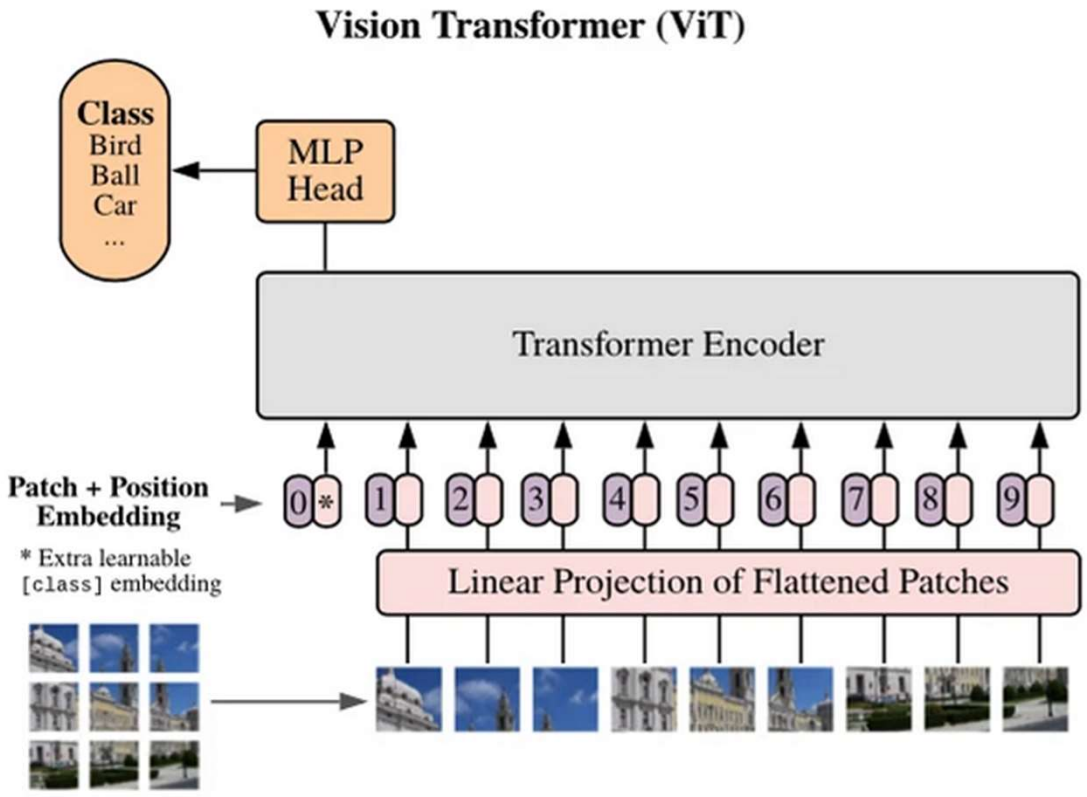
Transformers

- Sequence of tokens (all at once)
- Attention-Mechanism
- Positional Embeddings



[4] *Attention Is All You Need*, 2017

Vision Transformers

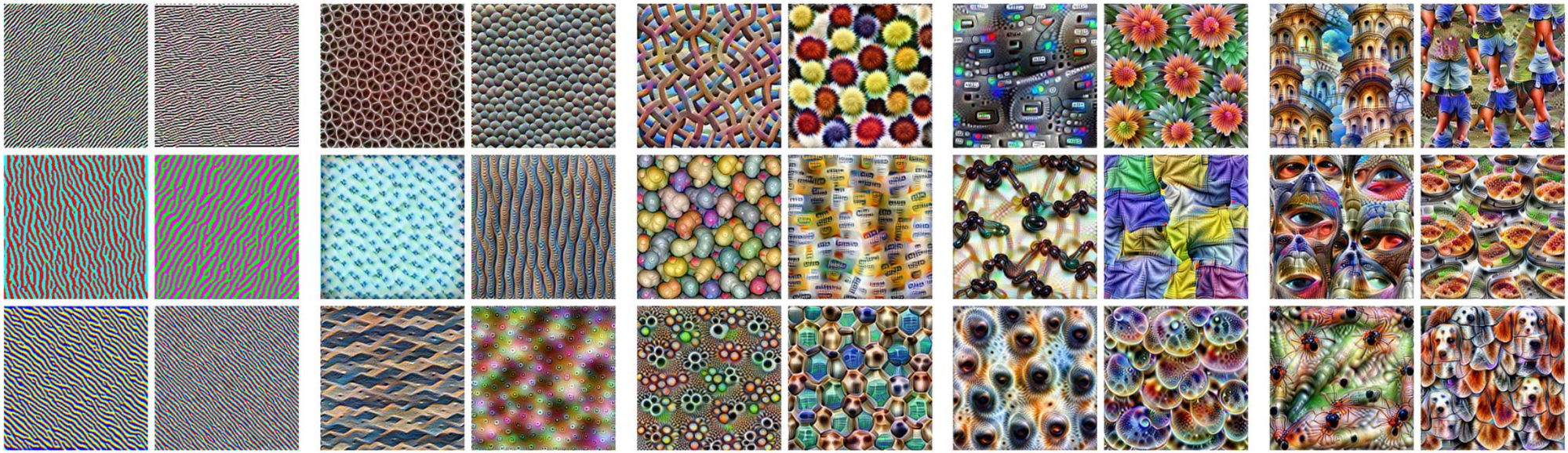


[2]*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021

Vision Transformers

[2] *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021

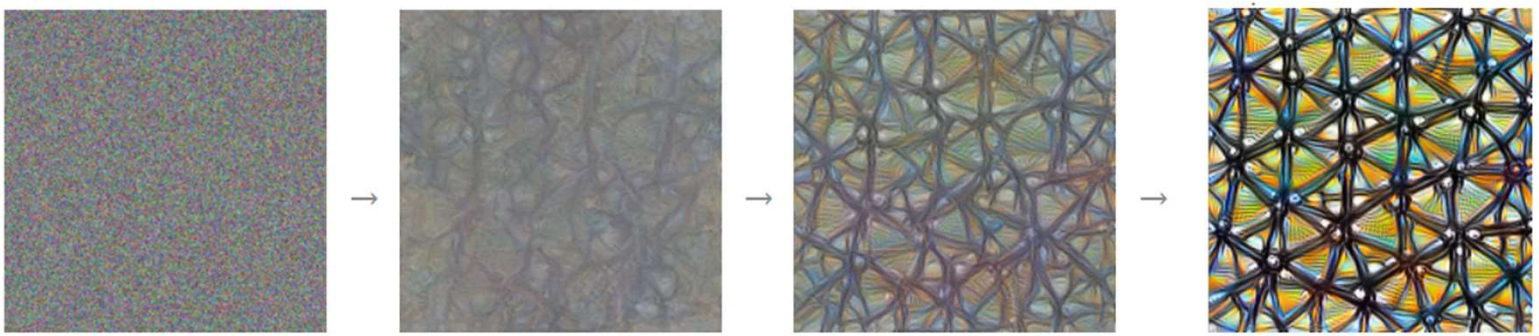
Visualization of Features: GoogleNet on ImageNet



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

[5]Feature Visualization, distill.pub

Feature Visualization: GoogleNet on ImageNet



[5]Feature Visualization, distill.pub

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\ s.t.\ ||\mathbf{x}||=\rho} h_{ij}(\theta, \mathbf{x})$$

[6]*Visualizing Higher-Layer Features of a Deep Network*, 2009

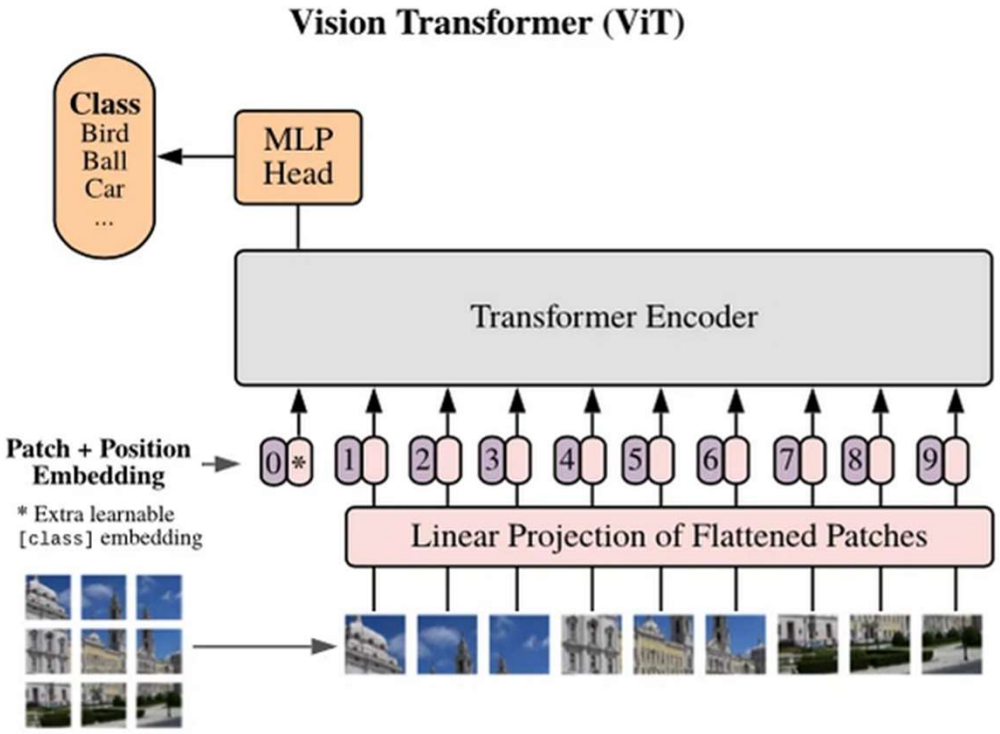
Visualization of Features: ViT?

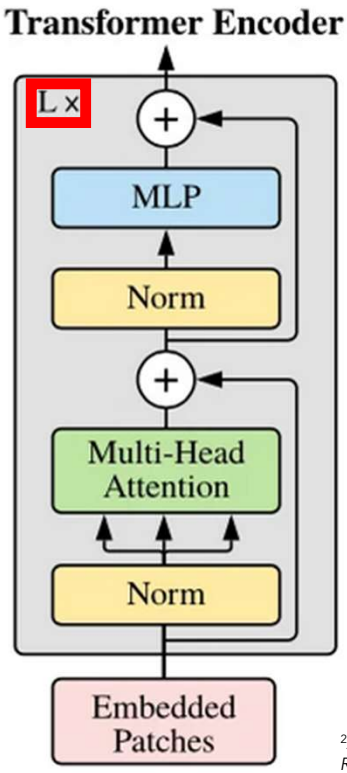
- Not yet explored!

- Some work related to ViTs' understanding:
  - ViTs robust to many kind of adversarial perturbation and corruption[7]
  - ViTs low-pass filters[8]
  - ViTs resistant to high-frequency removal[8]
  - …

- What Features do they tend to learn?

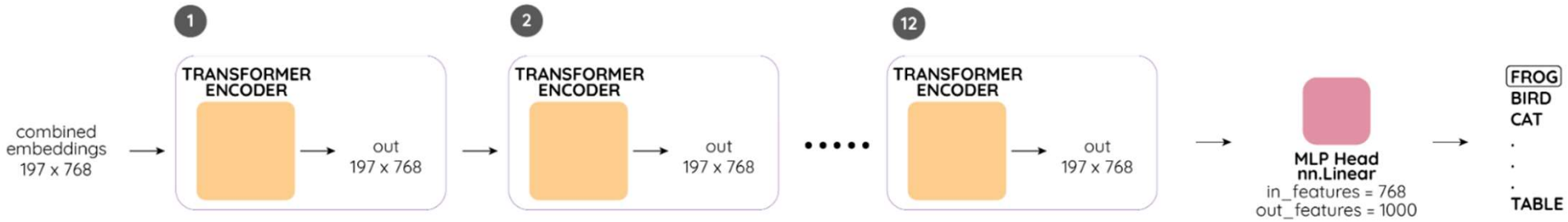
Visualization of Features: Gradient Steps





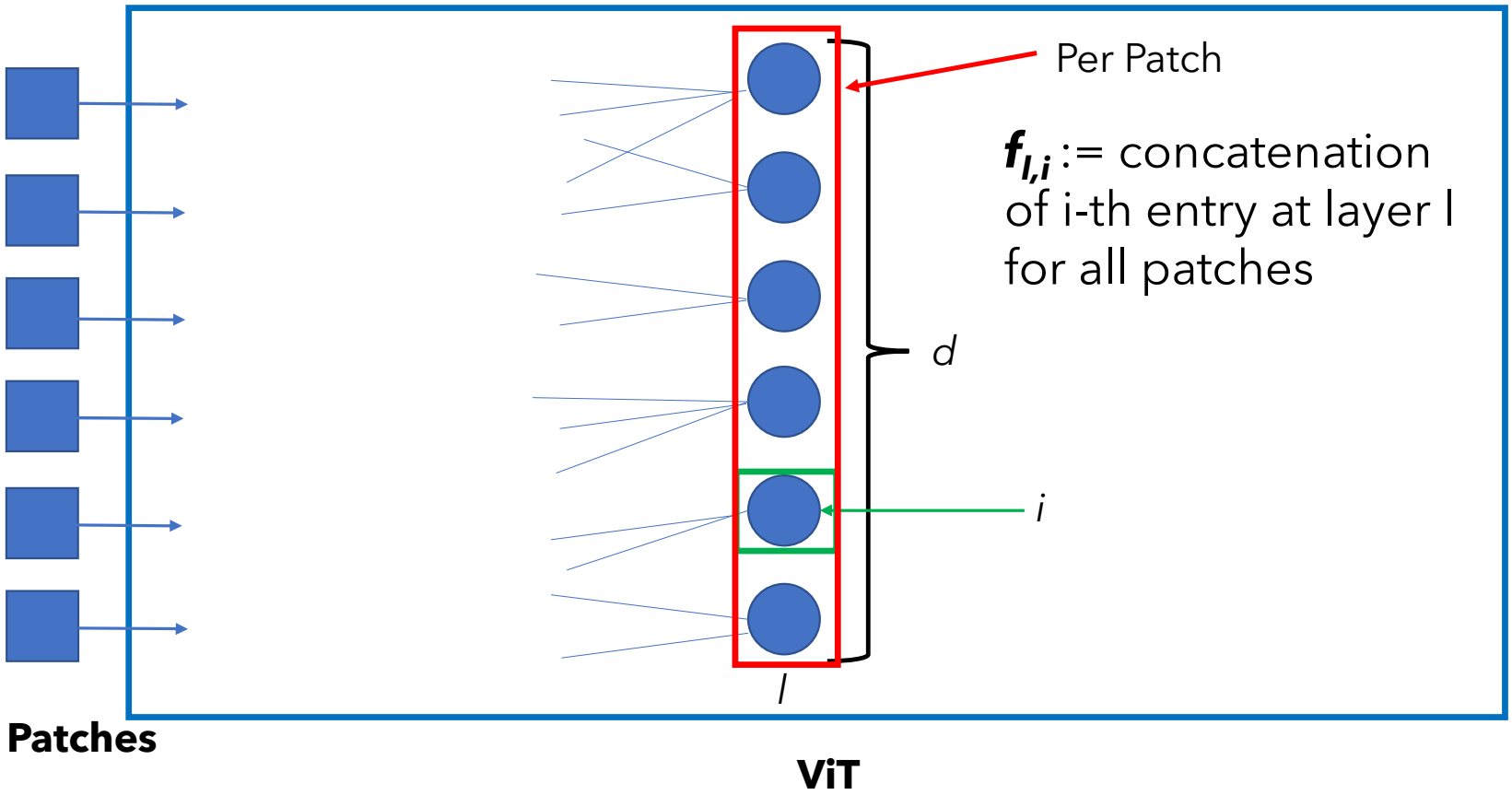
[2] *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021

Visualization of Features: Gradient Steps



[15]Visual Exploration Vision Transformers, amaarora.github.io

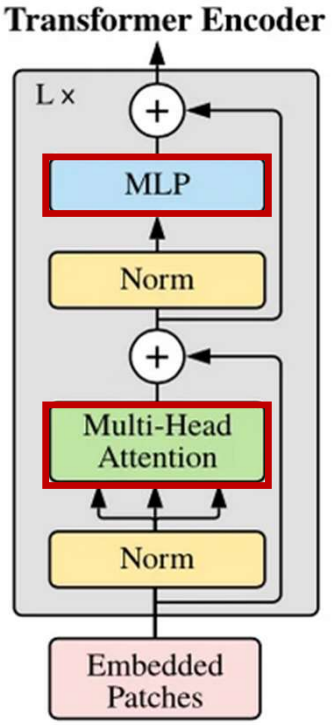
Feature Vector: $f_{l,i}$



Visualization of Features: Gradient Steps

$$\mathcal{L}_{\text{main}}(x, l, i) = \sum_{p} (f_{l,i})_p$$ [1]

max



[2] *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021

Visualization of Features: Gradient Steps

- **Optimization Problem (Quality)**

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

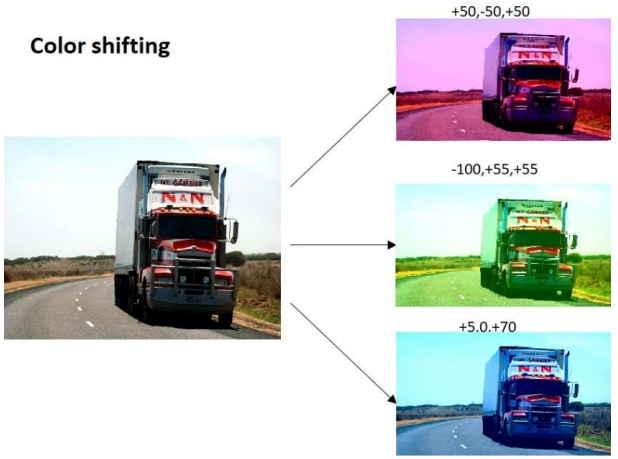
1

- TV :=Total Variation
- $a_k \in A := GS(CS(Jitter(x)))$

Best Visualization

Total Variation denoising [10]Total Variation Denoising, wikipedia.org

Gaussian Smoothing [11]Gaussian Smoothing, media5.datahacker.rs



[12]Color shifting augmentation, 3.bp.blogspot.com

Color Shifting augmentation

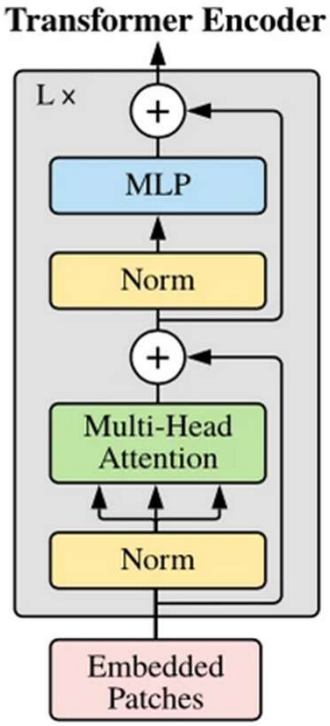
$$x^* = \arg\max_x \sum_k \mathcal{L}_{\text{main}}(a_k(x), l, i) + \lambda TV(a_k(x))$$

$$A \coloneqq GS(CS(Jitter(x)))$$

Visualizations

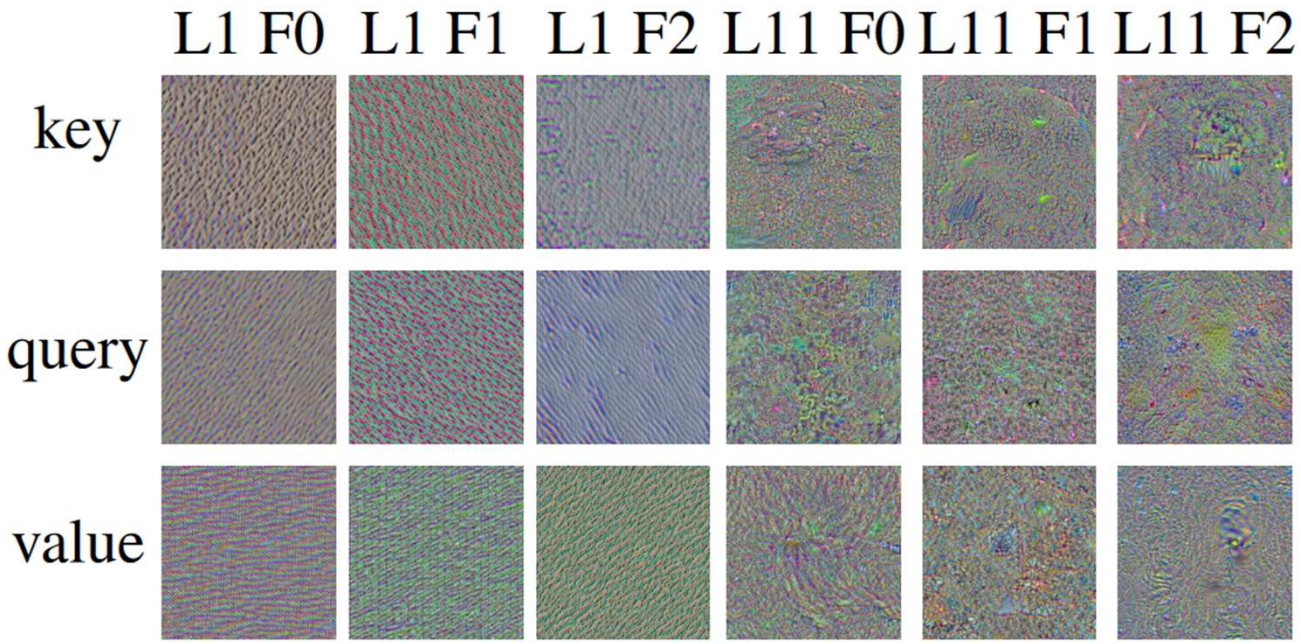
- 38 ViT models' variants tested

- ViT-B16 Model[2]
  - ImageNet
  - 12 Blocks:
    - MH-Attention layers (768 size)
    - Projection layers for Mixing
    - Feed-forward layer (3072 size)



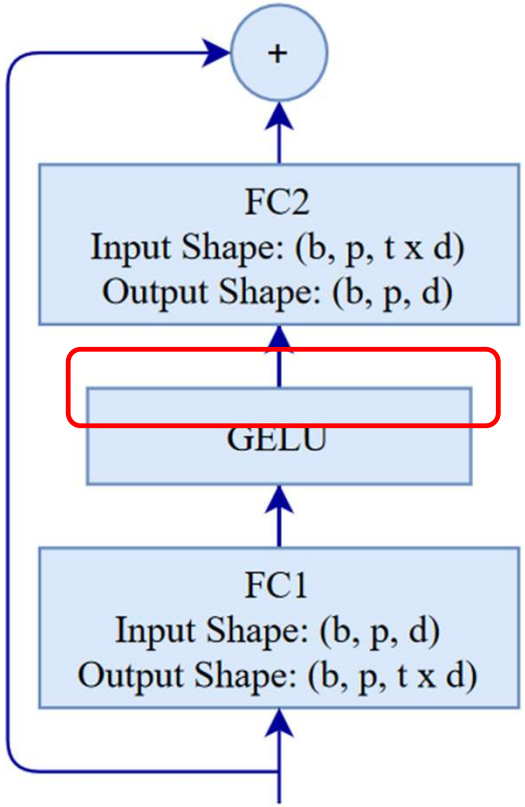
[2] *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021

Visualizations: Multi-Head Attention



1

Visualizations: Feed-Forward layers



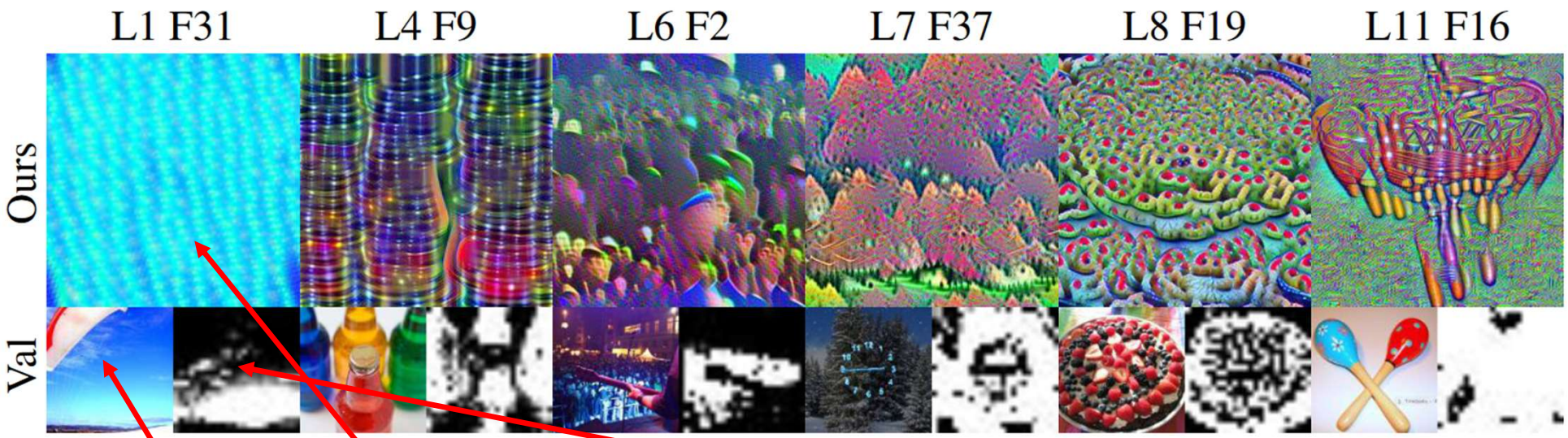
1

Visualizations: Feed-Forward layers

[1]

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

Visualizations: Feed-Forward layers



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

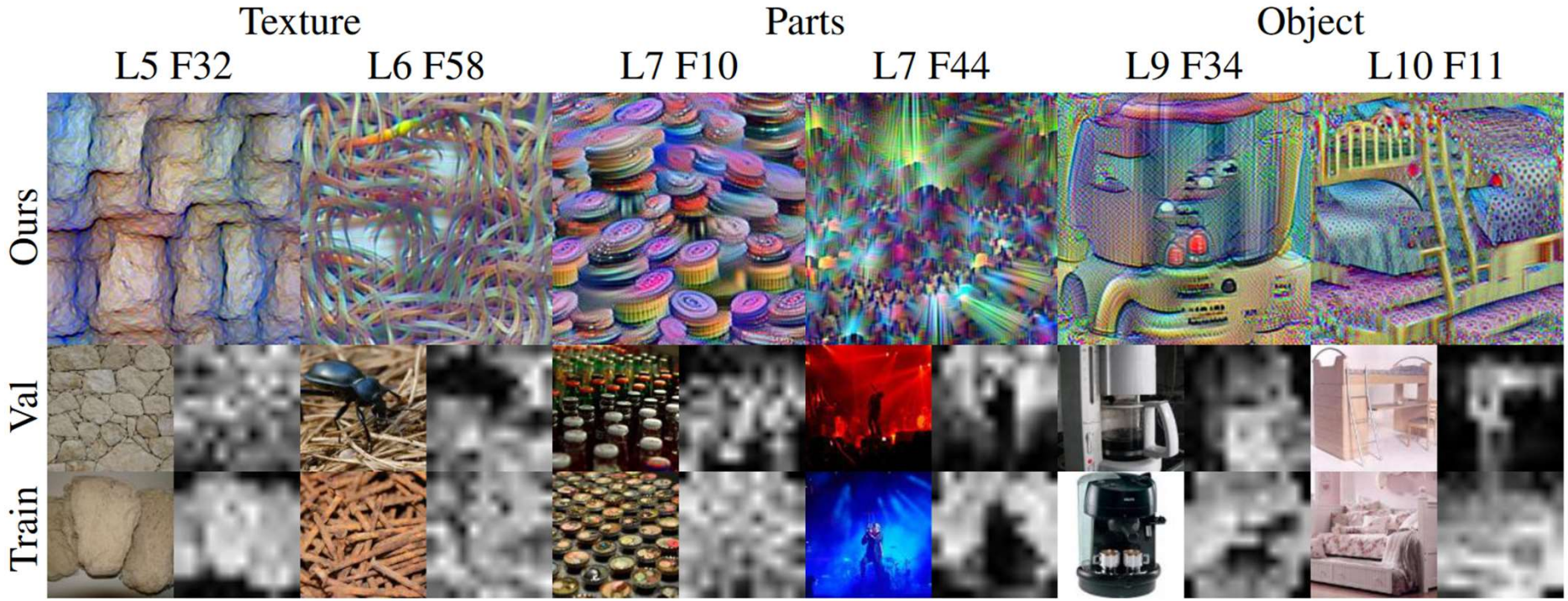
Patch-Wise spatial information preservation

- Every patch *can* influence the representation of every other patch
  - Yet the representation remains local

Vits retains spatial information[1]

CNN vs ViT: Progressive specialization

- CNN exhibits a progressive features' specialization



1

CNN vs ViT: Background & Foreground

- **Hypothesis:** ViT better at using background features

[1]

- **Experiment:** Images from ImageNet → Hide Foreground/Background

[1]

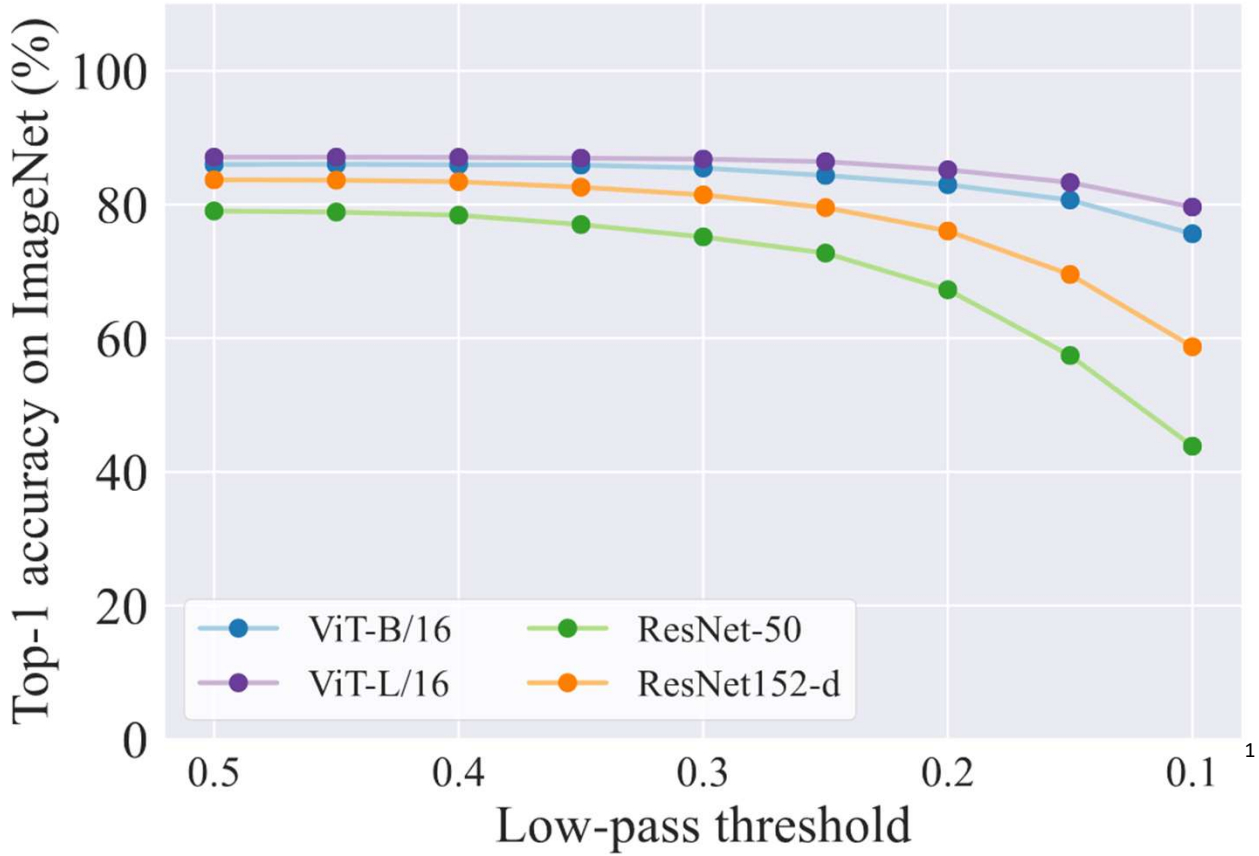
CNN vs ViT: Background & Foreground

- **Results**:

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

1

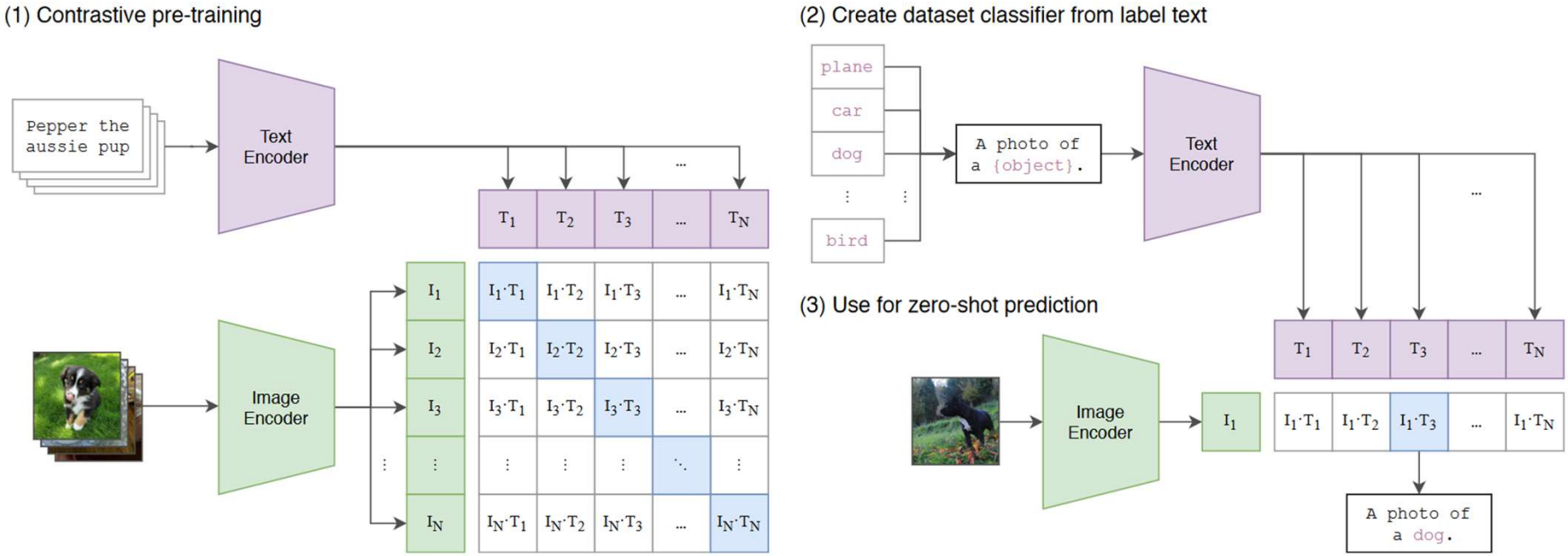
CNN vs ViT: Low-pass filtering



[1]

ViTs with Language Model Supervision

- CLIP multimodal (NLP and CV) model is state-of-the-art in transfer learning to unseen data (Zero-Shot)



[14] *Learning Transferable Visual Models From Natural Language Supervision*, 2021

ViTs with Language Model Supervision

- Different Features!

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

1

ViTs with Language Model Supervision

- Different Features!

(a) Before and after/Step-by-step (b) From above (c) Many

1

Summary

- Understand how to apply Feature Visualization on ViT
  - Feed-Forward Layer
- Features' visualization of 38 models
- ViTs retain positional relationship between patches
- ViTs make better use backgrounds' information compared to CNNs
- ViTs rely less on high-frequency, textural attributes
- ViTs exhibit progressive feature specialization
- ViTs trained with Natural Language Supervision (CLIP) learn semantical and conceptual features rather than object-specific features

Personal Opinion

- + Lots of Pictures and Visualizations
- + Easy to follow and understand
- + Scientific approach in the experiments
- - Not super innovative (reuse existing methods)
- - Some experiments are not completely convincing

QUESTIONS?

Bibliography/Sitography

- [1] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, Tom Goldstein. *What do Vision Transformers Learn? A Visual Exploration*, 2022
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021
- [3] ResNet-50's, Semantic Segmentation's, Object Detection's picture. miro.medium.com, Accessed: 28.02.2023
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. *Attention Is All You Need*, 2017
- [5] Website of Feature Visualization. distill.pub/2017/feature-visualization, Accessed: 28.02.2023
- [6] Erhan, Dumitru & Bengio, Y. & Courville, Aaron & Vincent, Pascal. *Visualizing Higher-Layer Features of a Deep Network*, 2009
- [7] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. *On the adversarial robustness of visual transformers*, 2021
- [8] Namuk Park and Songkuk Kim. *How do vision transformers work?*, 2022
- [9] Amazing Vision transformers explanation: amaarora.github.io/2021/01/18/ViT.html, Accessed: 28.02.2023
- [10]Total Variation Picture: wikipedia.org/wiki/Total_variation, Accessed: 28.02.2023
- [11]ColorShift Picture: https://media5.datahacker.rs, Accessed: 28.02.2023
- [12]Gaussian Smoothing Picture: 3.bp.blogspot.com, Accessed: 28.02.2023
- [13]Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, Wieland Brendel. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*, 2018
- [14]Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*, 2021
- [15]Visual Exploration Vision Transformers: amaarora.github.io, Accessed: 28.02.023