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

ResNet50 Model Architecture

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:
Transformers

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

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

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)

Feature Visualization, distill.pub
Feature Visualization: GoogleNet on ImageNet

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

*Feature Visualization, distill.pub

*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\textsuperscript{7}
  • ViTs low-pass filters\textsuperscript{8}
  • ViTs resistant to high-frequency removal\textsuperscript{8}
  • ...

• What Features do they tend to learn?
Visualization of Features: Gradient Steps

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021
Visualization of Features: Gradient Steps

1. Combined embeddings: 197 x 768
   → Transformer Encoder: out 197 x 768
   → Transformer Encoder: out 197 x 768
   → ... → Transformer Encoder: out 197 x 768

   MLP Head
   nn.Linear
   in_features = 768
   out_features = 1000

   → FROG
   BIRD
   CAT
   ...
   TABLE

\[1\text{Visual Exploration Vision Transformers, amaarora.github.io} \]
Feature Vector: $f_{l,i}$

$f_{l,i} := \text{concatenation of } i\text{-th entry at layer } l \text{ for all patches}$
Visualization of Features: Gradient Steps

\[ L_{\text{main}}(x, l, i) = \max_p \sum (f_{l,i})_p \]

\[ ^2 \text{An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021} \]
Visualization of Features: Gradient Steps

- Optimization Problem (Quality)

\[ x^* = \arg \max_x \sum_k L_{\text{main}}(a_k(x), l, i) + \lambda TV(a_k(x)) \]

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

**Total Variation denoising**

**Gaussian Smoothing**

**Color Shifting augmentation**

\[ x^* = \arg \max_x \sum_k L_{\text{main}}(a_k(x), l, i) + \lambda T V(a_k(x)) \]

\[ A := GS(CS(Jitter(x))) \]
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

<table>
<thead>
<tr>
<th></th>
<th>L1 F0</th>
<th>L1 F1</th>
<th>L1 F2</th>
<th>L11 F0</th>
<th>L11 F1</th>
<th>L11 F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>key</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>query</td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td>value</td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Visualizations: Feed-Forward layers
Visualizations: Feed-Forward layers

\[ x^* = \arg \max_x \sum_k \mathcal{L}_{\text{main}}(a_k(x), l, i) + \lambda TV(a_k(x)) \]
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
CNN vs ViT: Progressive specialization

- CNN exhibits a progressive features’ specialization
CNN vs ViT: Background & Foreground

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

• **Experiment:** Images from ImageNet → Hide Foreground/Background
CNN vs ViT: Background & Foreground

- **Results:**

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Full Image</th>
<th>Foreground</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B32</td>
<td>98.44</td>
<td>93.91</td>
<td>28.10</td>
</tr>
<tr>
<td>ViT-L16</td>
<td>99.57</td>
<td><strong>96.18</strong></td>
<td><strong>33.69</strong></td>
</tr>
<tr>
<td>ViT-L32</td>
<td>99.32</td>
<td>93.89</td>
<td>31.07</td>
</tr>
<tr>
<td>ViT-B16</td>
<td>99.22</td>
<td>95.64</td>
<td>31.59</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>98.00</td>
<td>89.69</td>
<td>18.69</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>98.85</td>
<td>90.74</td>
<td>19.68</td>
</tr>
<tr>
<td>MobileNetv2</td>
<td>96.09</td>
<td>86.84</td>
<td>15.94</td>
</tr>
<tr>
<td>DenseNet121</td>
<td>96.55</td>
<td>89.58</td>
<td>17.53</td>
</tr>
</tbody>
</table>
CNN vs ViT: Low-pass filtering
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
ViTs with Language Model Supervision

- Different Features!
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


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


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
