Doggo in the Machine? A Study of VGG16 Explainability

Hypothesis

H0: VGG16's neural network classification of a photo of a German Shephard is based on learned patterns that are dissimilar to or not understandable by humans.

H1: VGG16's neural network is trained to "look for" certain distinct visual features — like ears, snout, fur, paws — that are understandable by humans in its classification of a German Shephard.

Input Image

(Awwww. Stock photo from Google Images)

Approach

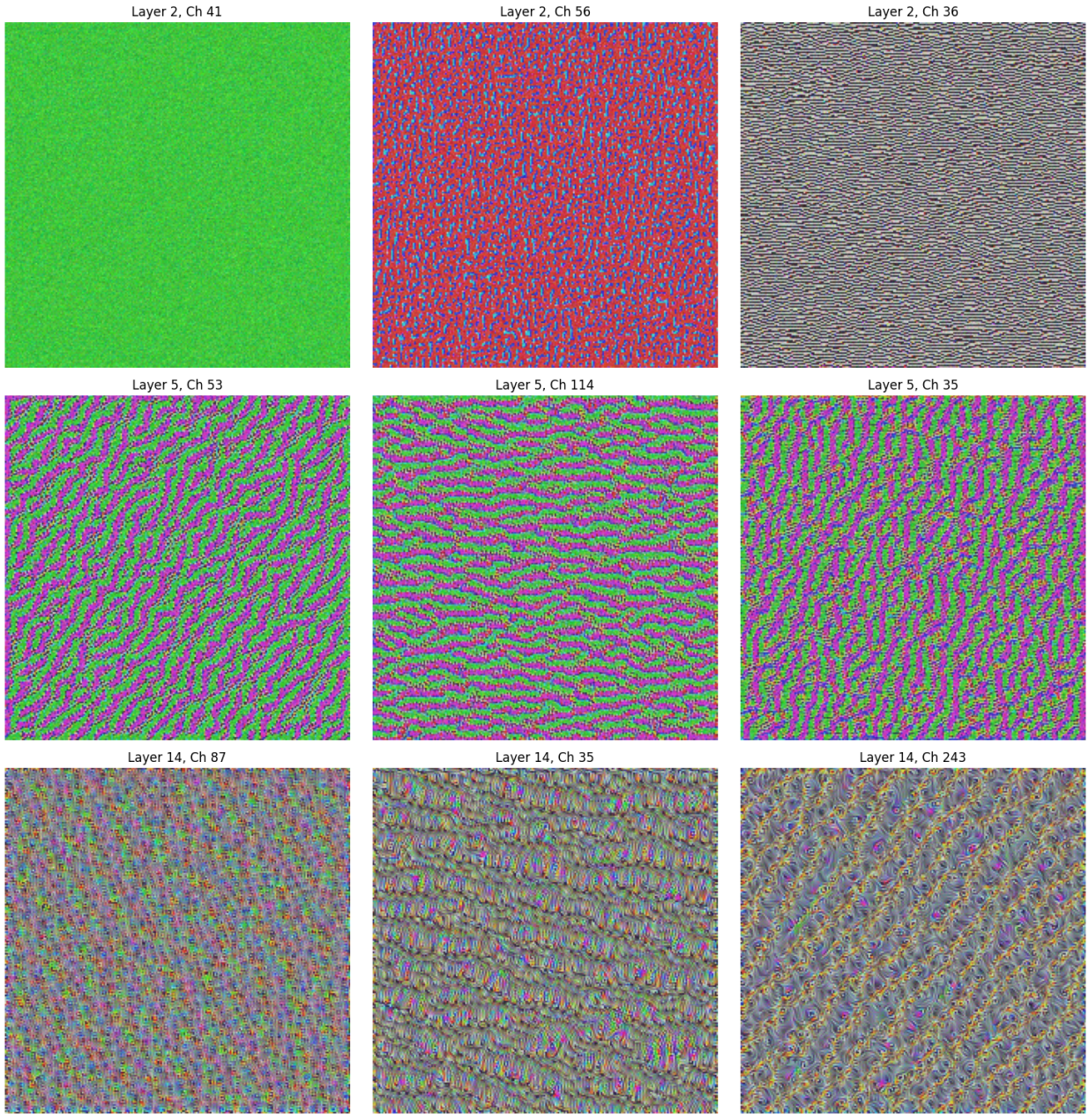
Given the computing and time restraints, I focused on six convolutional layers: 2, 5, 14, 19, 24, 26. This was so that there would be an even spread of early, middle, and end layers.

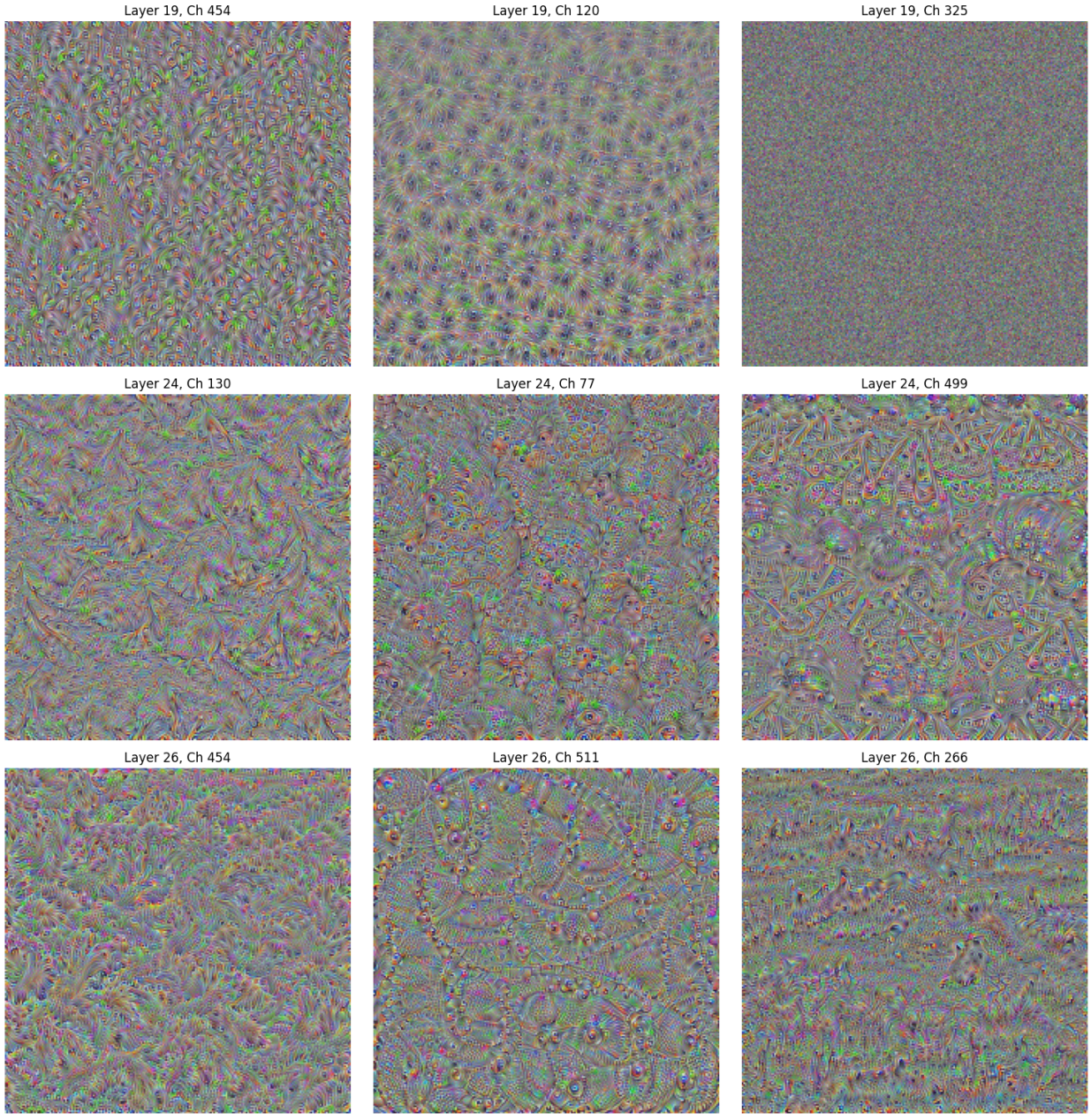
After passing the image through the model, for each channel in these layers, an importance score was calculated based on the highest summed gradient magnitudes. This identifies the channels in the convolutional layer that are the most "influential." The top three most important channels for any given layer were identified.

```
Layer 2: [41 56 36]
Layer 5: [53 114 35]
Layer 14: [87 35 243]
Layer 19: [454 120 325]
Layer 24: [130 77 499]
Layer 26: [454 511 266]
```

Visualization

Then, visualizations were generated by via optimization for each of the identified channels. The results are as follows:



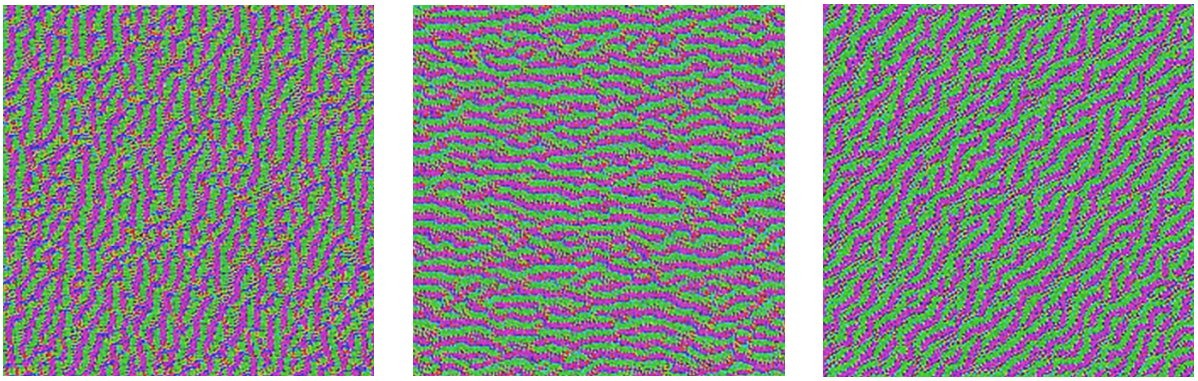


Early Mistakes

You will find in my Colab Notebook that it took a bit of workshopping to fully understand how to use optimization properly for feature visualization. Initial missteps were largely driven by conflating *visualization* and *attribution*. Early efforts started with randomly selecting a single channel to sample, and then randomly sampling from each convolutional layer. After reviewing the Olah, et al. paper, it became clear that the visualization process simply activated a channel, but did not identify relevant channels. (Hence, earlier efforts were visualizing random channels within the neural network). Thus, a two-part process was required: first to identify the most relevant channels, then to optimize those features.

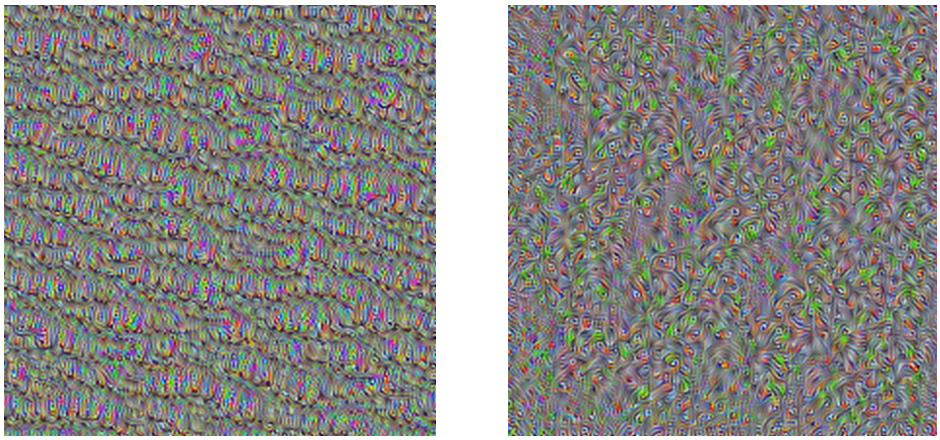
Results

Channels in earlier convolution layers seemed to be “looking for” certain edges. For example, channels 35, 114, and 53 in layer 5 seem to be activated by vertical, horizontal, and sloping edges, respectively.



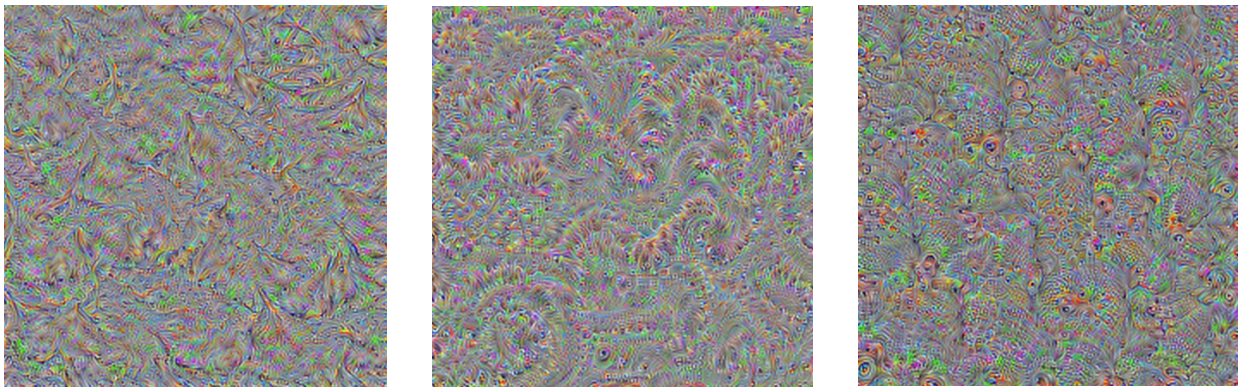
Channels 35, 114, and 53 in Layer 5.

Later convolutional layers seem to be detecting patterns of certain shape profiles.



Channel 14 in Layer 35, Channel 454 in Layer 19

In the final convolutional layers, more abstract — and potentially typological — shapes are emerging in its channels. See the Channel 130 in Layer 24 that resembles dog ears, Channel 454 in Layer 26 that resembles tails, and Channel 77 in Layer 24 that resembles a combination of eyes and snouts.



From Left to Right: Channel 130 in Layer 24 (“Dog Ears”), Channel 454 in Layer 26 (“Tails”), Channel 77 in Layer 24 (“Eyes and Snouts”)

Conclusion and Discussion

There is *some* evidence that convolutional layers — particularly ones later in the pipeline — are activated by categorical features that are understandable by humans. However, the visualizations are so abstract and that the potential for confirmation is so strong that I hesitate to reject the null hypothesis.

Further Research

Based on the results of this constrained experiment, one hypothetical direction to take future research (for more brilliant minds) is to explore if there is a way to create stepwise “composite” visualizations of top channels as a way to visualize how the model is drawing from various feature to build an understanding of the input for its classification. A stepwise process may illuminate how each channel adds to this understanding. If that composite is more understandable by humans, it may us closer to rejecting the null hypothesis.

References

- Google Colab Notebook
- ChatGPT(o1, 4o) Queries
- Reference Paper: Olah, et al., "Feature Visualization", Distill, 2017. https://distill.pub/2017/feature-visualization/