Visualizations and Explanations

Computer Vision – Lecture 09

Further Reading

- Slides from L Fei-Fei.
- CVPR'18 Tutorial on Interpretable ML for CV
- Many posts on https://distill.pub


BBC

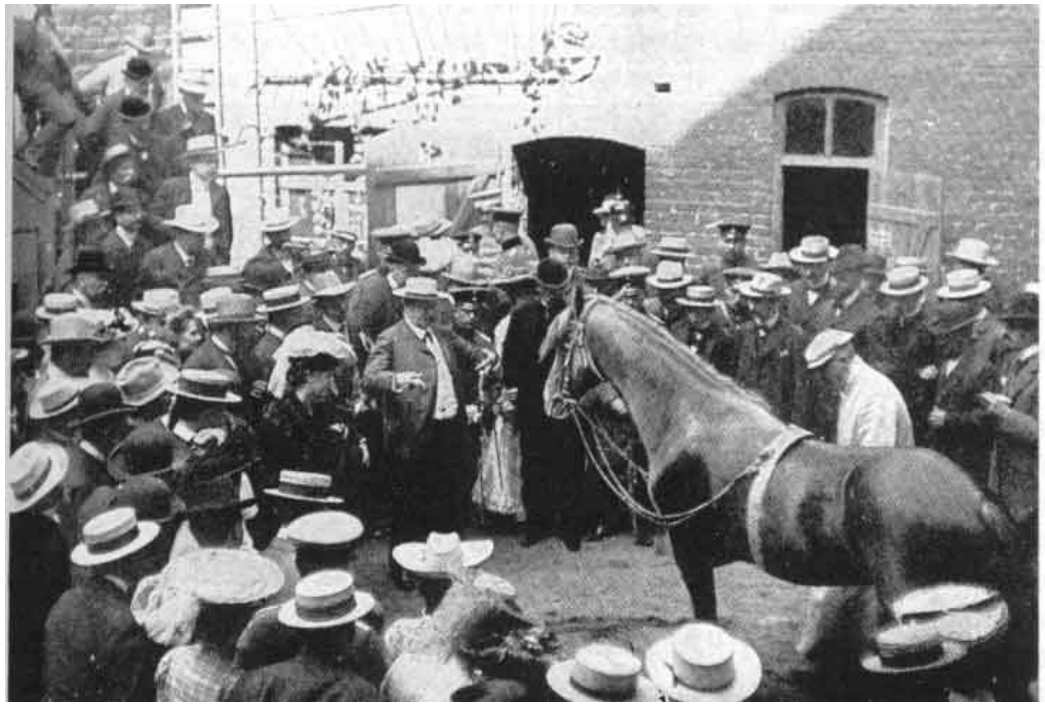
Which insights can we derive?

Does the orangutan know what a hammer is?

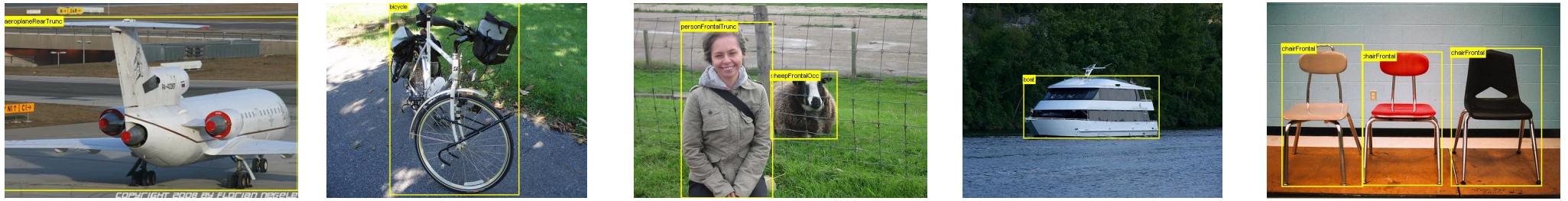
- Does she know *hammering* (Task)? Yes
- Can she *hammer* (Task)? No
- Does she know what a hammer is (Concept)? Maybe?

Clever Hans

- 1895-1916 German horse that was doing arithmetic.
- Formal investigation: horse was watching the reactions of his trainer.
- Trainer was entirely unaware that he was providing such cues.
- -> Clever Hans Effect.



Example: Image Classification



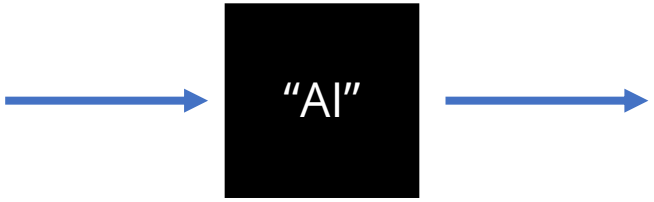
- Benchmark: PASCAL Visual Objects in Context (2004 - 2012)
- 20 classes:

  person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, tv/monitor

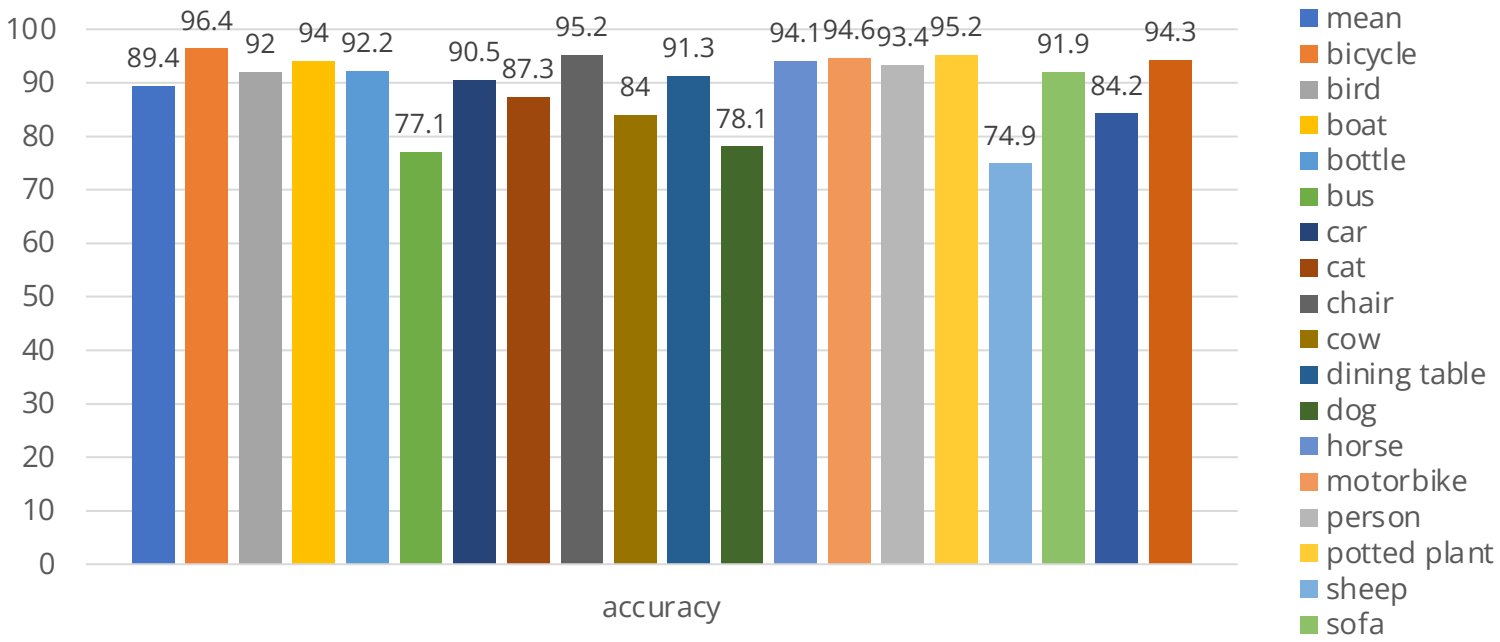
- 10,000 images with 25k objects

The PASCAL Visual Object Classes Challenge: A Retrospective
Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A.
International Journal of Computer Vision, 111(1), 98-136, 2015

Evaluation – test on unseen data



- horse
- person

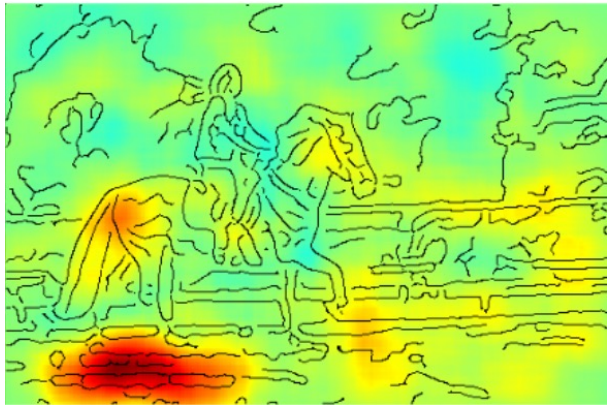


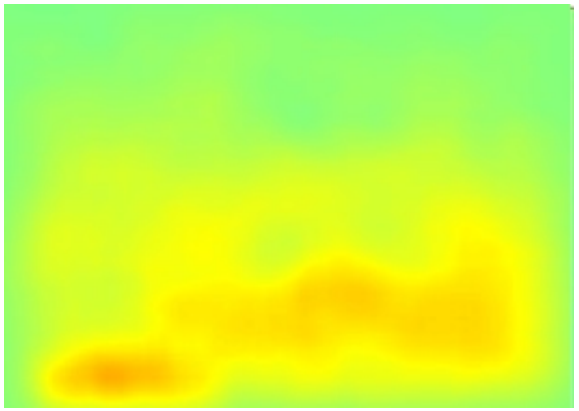
Jincao Yao
Cross-layer sparse atrous convolution network
http://host.robots.ox.ac.uk:8080/leaderboard/displaylb_main.php

Does the model know what a horse is?

- Evaluation accuracy is high - does that mean yes?

input


heatmap for “horse”


average heatmap for all horse images

- No! It uses spurious correlations: many horse images have a copyright notice in this dataset

Explainability

- Helps to uncover biases in data and models.
- Additional tool beyond test set evaluation.
- Another layer of verification.
- Help to create trust.
- Allows for non-expert interaction.
- Can lead to new insights.
- Is in the law: (GDPR Art. 13,14,22) “the right for explanation”.

Ke Jie vs. AlphaGo

(f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

GDPR, Art. 13.2.f

Explanations

Recipient

Explanations need to adapt to the recipient of the information.

- Developers
- Domain experts
- Users

Content

Explanations provide different types of information.

- Representations
- Individual predictions
- Behaviour
- Examples

Purpose

Explanations differ based on use-cases.

- What question is answered by the explanation?
- What is the explanation used for?

Towards explainable artificial intelligence
Samek Wojciech, and Klaus-Robert Müller
Explainable AI: interpreting, explaining and visualizing deep learning
Springer, Cham, 2019

Taxonomy - Approaches

Post-Hoc Analysis

Explanations are derived from a fixed, pre-trained model via analysis.

- No impact on performance
- Difficult
- Explanations are often *local* around predictions
- *Main focus today!*

Transparent Models

The model is constructed such that (some) mechanism have semantic meaning.

- Does not need post-hoc analysis
- Task-specific architecture
- Can affect performance

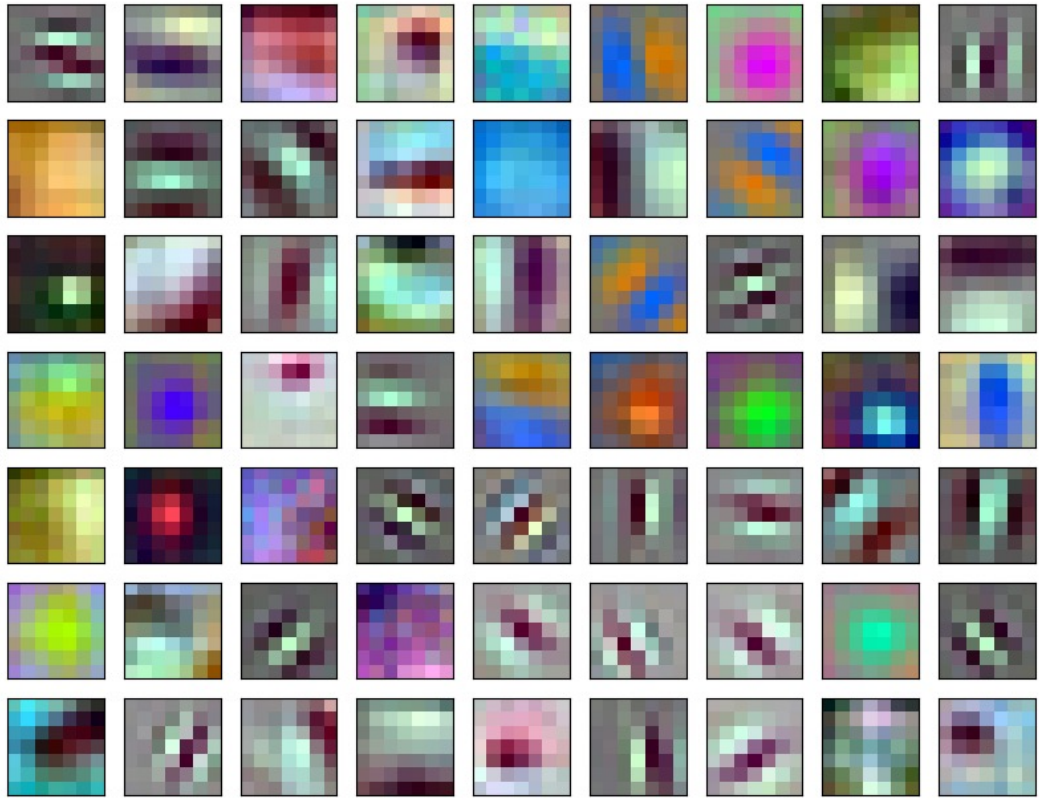
Learned Explanations

The model is trained to deliver explanations together with predictions.

- Explanations can be very semantic
- Might need meta-explanations
- Can affect performance

Post-Hoc Analysis: First Layer

First-layer filters from **ResNet18** ($[7\times7\times3]$ filters):

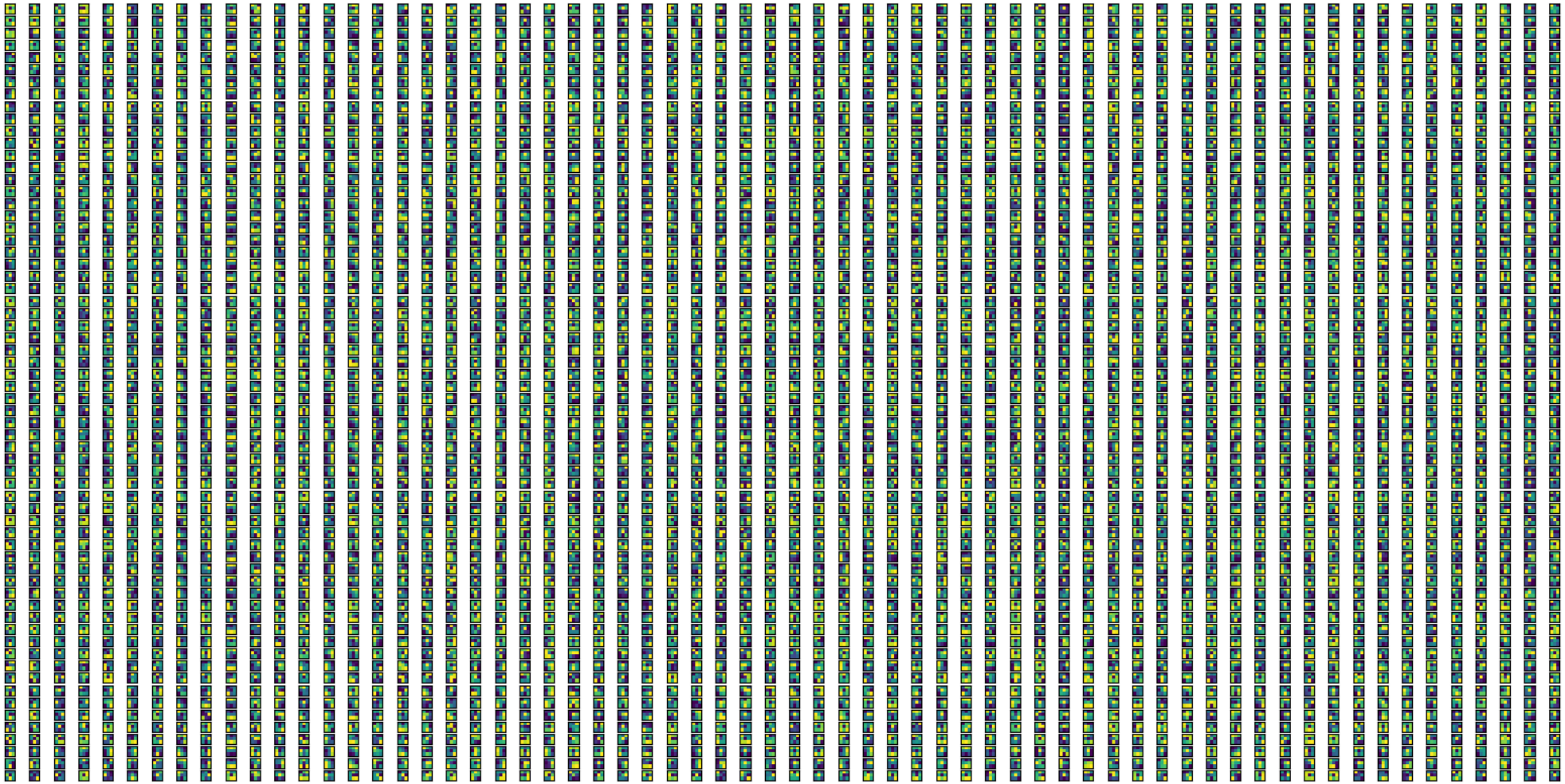


First-layer learned features include basic elements, such as edges, blobs, colors, etc.

Deeper layers depend on the features computed in the layers before: hard to directly understand the weights.

Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." NeurIPS 2012.

Second Layer

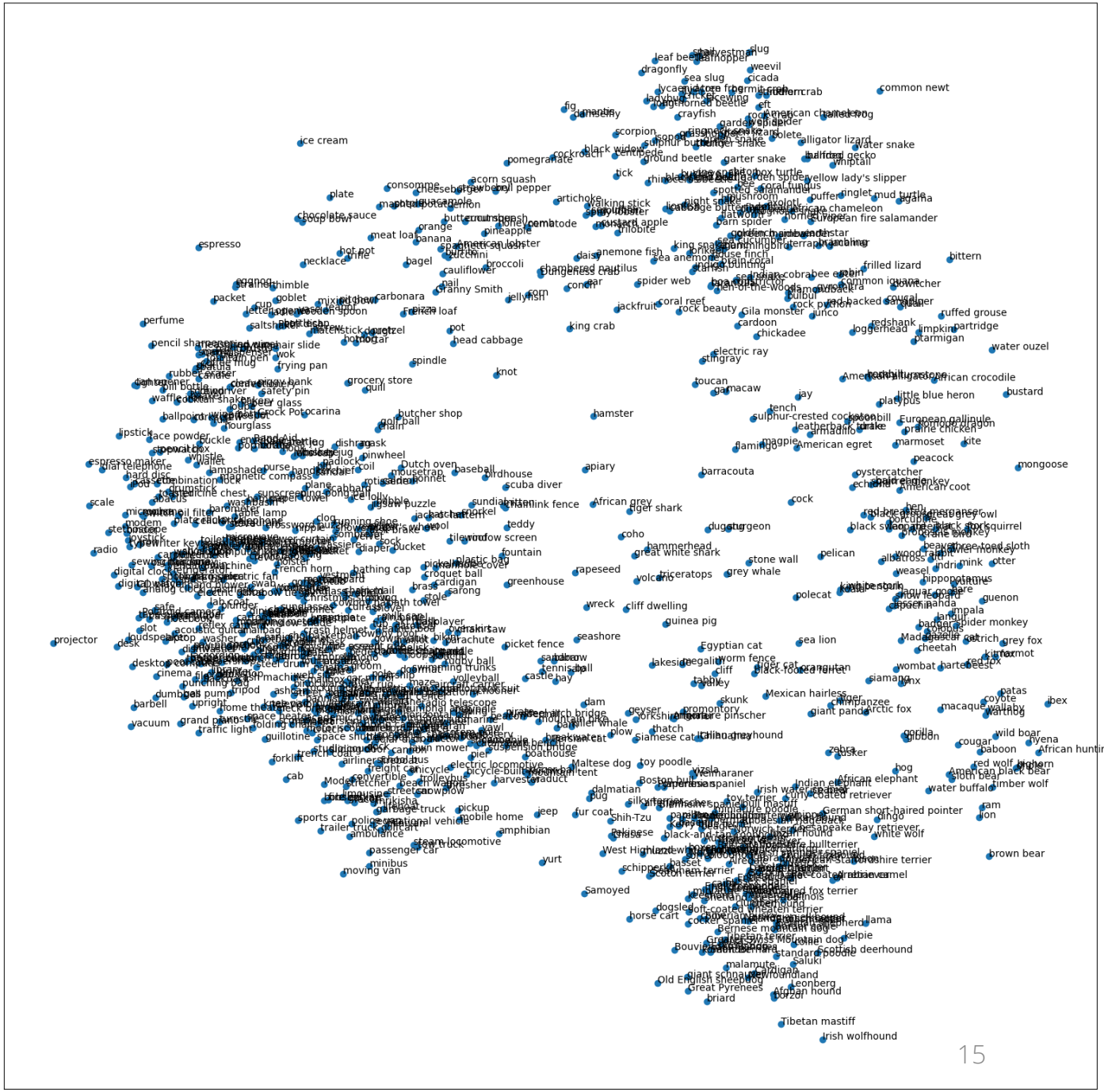


The second layer has 64 3x3 convolutions, each operates on 64 channels.

Not very interpretable!

Last Layer

- ResNet18: last layer 512x1000 (1000 class output)
- Dimensionality reduction with PCA (use the first 2 principal components)
- Observe groupings.



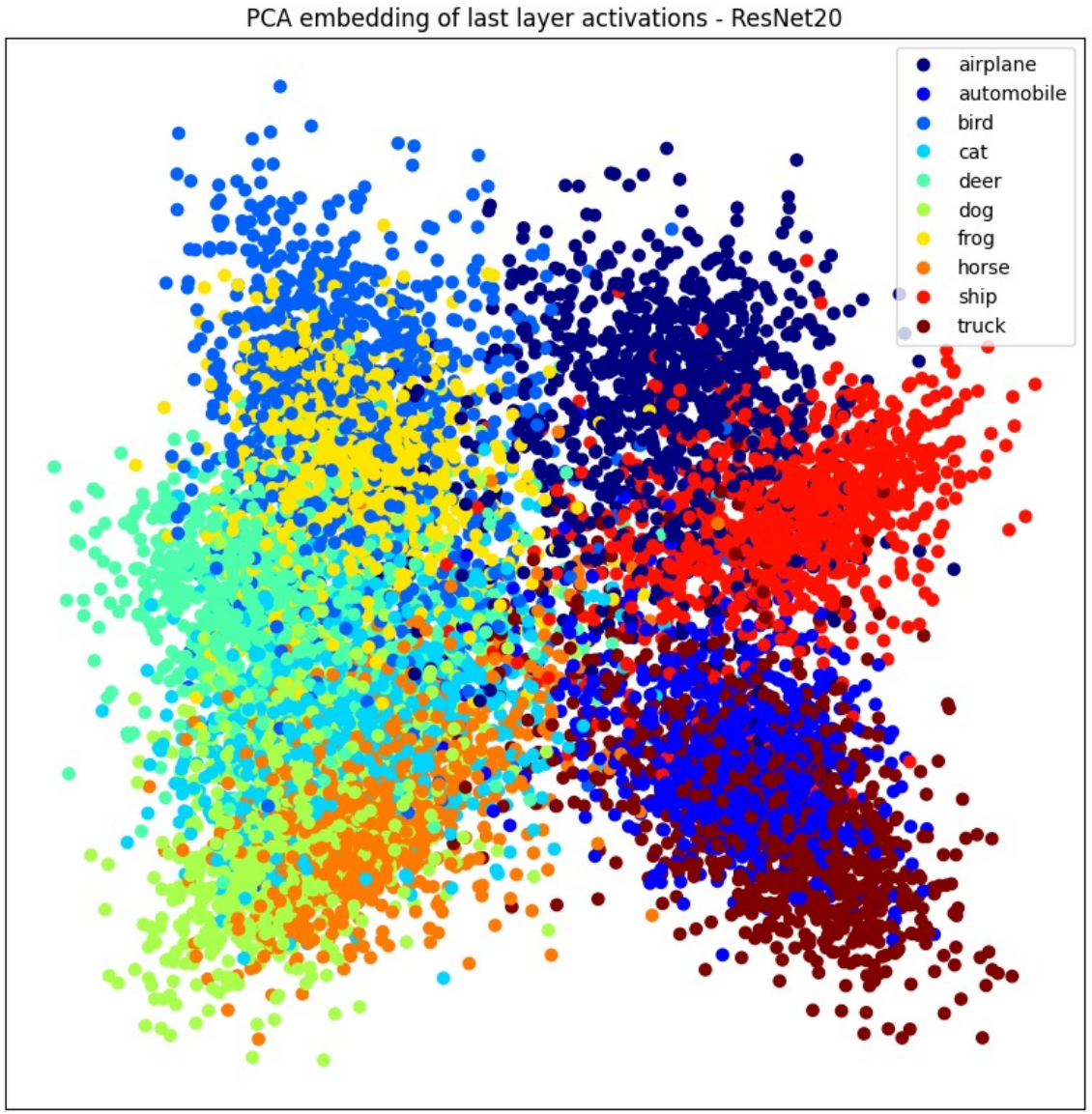
Post-Hoc Analysis

- So far: we have looked at the learned weights after training.
- This can show what the model has *learned*.
- We also want to understand what the model does with its inputs.

- We can also look at activations (=outputs of layers) instead.
- For that, we need to input data.
- Use data that was unseen during training: we want to understand generalisation.

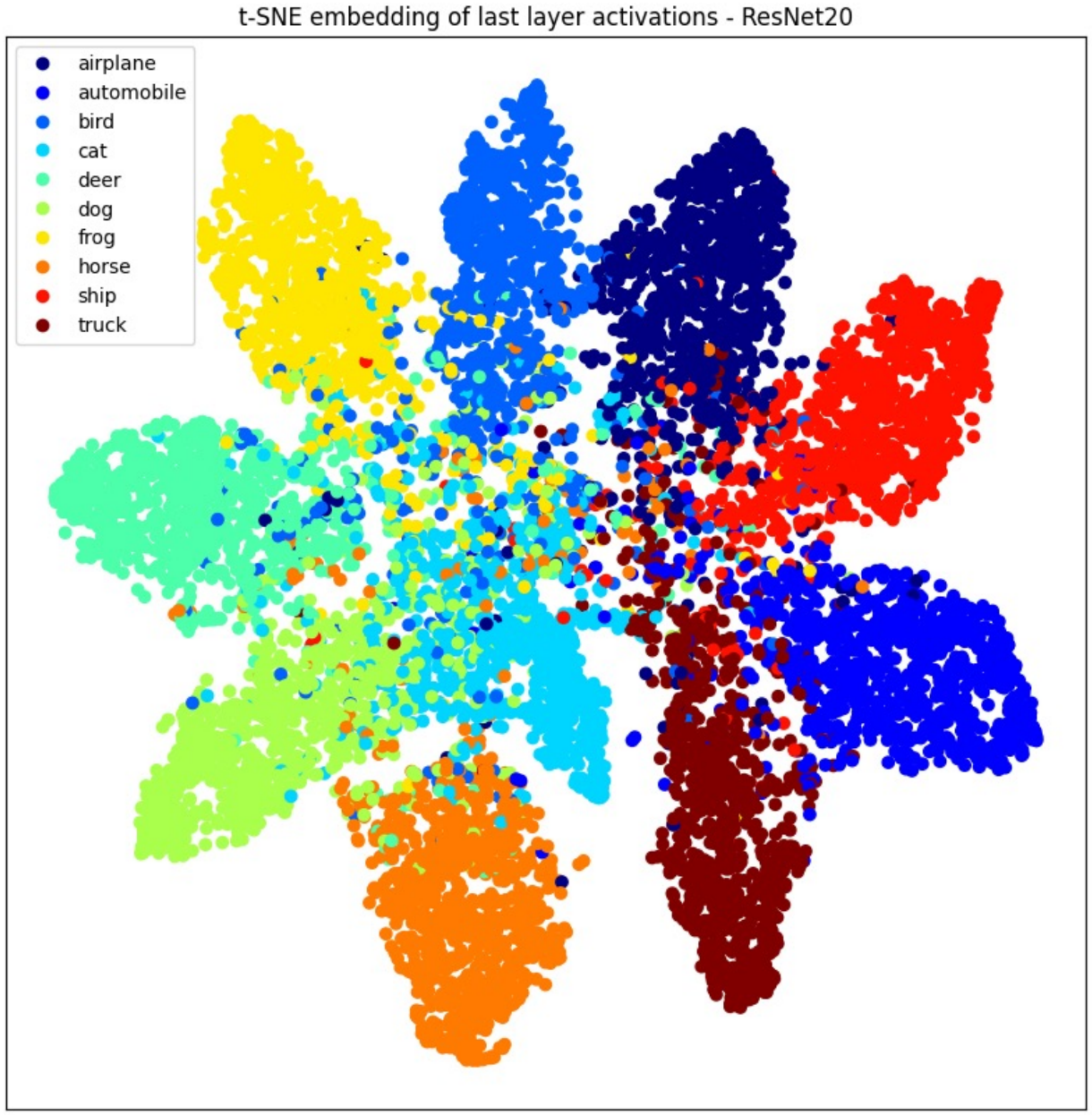
Last Layer Activations

- Compute inputs to the last layer of validation set images.
- Compute PCA.
- Visualise embedding with class labels.
- Last layer: linear+softmax, so we want linear separability.



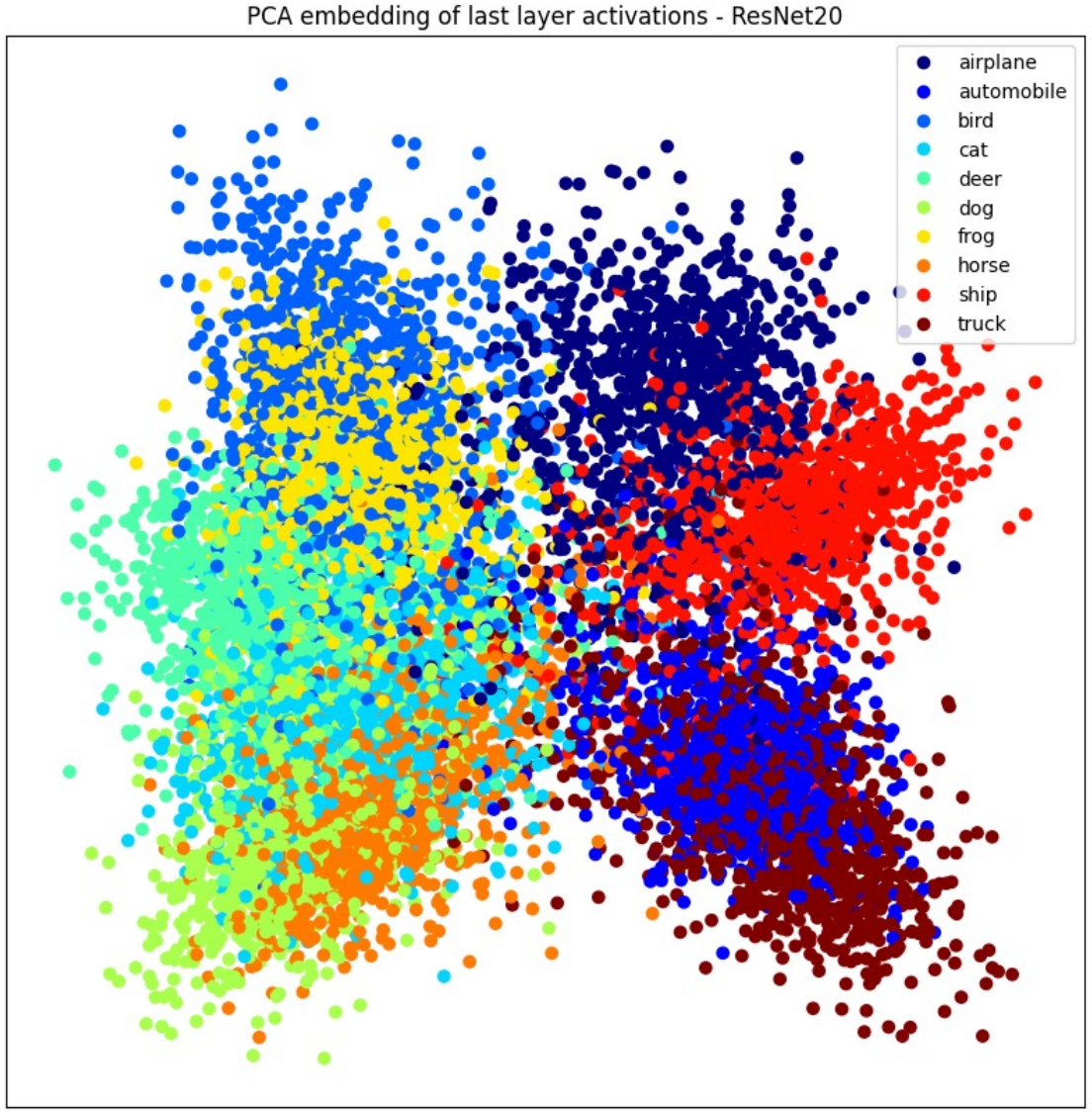
t-SNE Embedding

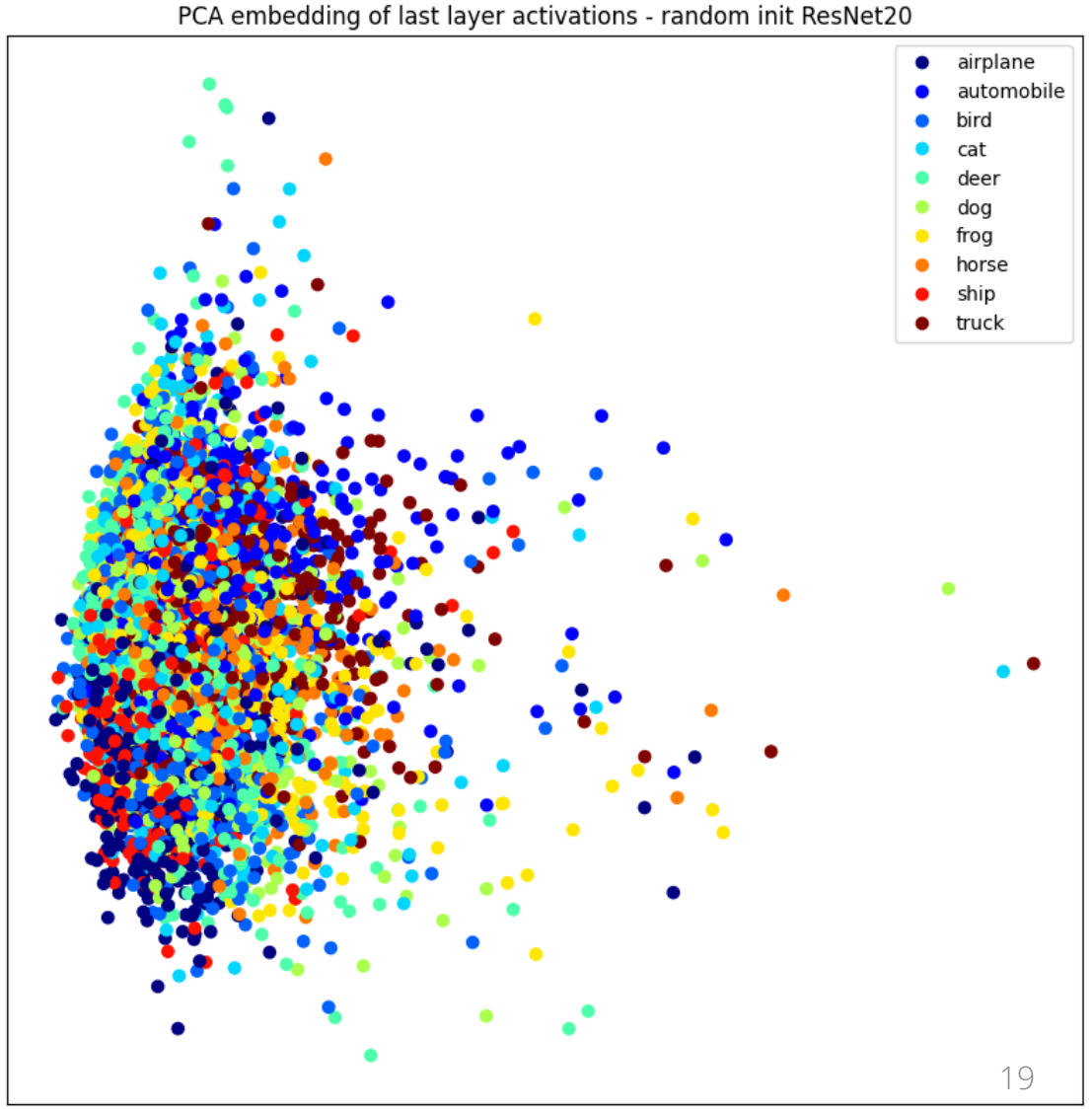
- PCA gives us a linear projection from a high dimensional space to 2 dimensions for visualisation.
- There are non-linear embedding techniques: e.g. t-SNE.
- Nicer plots, but less interpretable embedding.
- Further reading.



Hinton, Geoffrey; Roweis, Sam. Stochastic neighbor embedding. NeurIPS, 2021

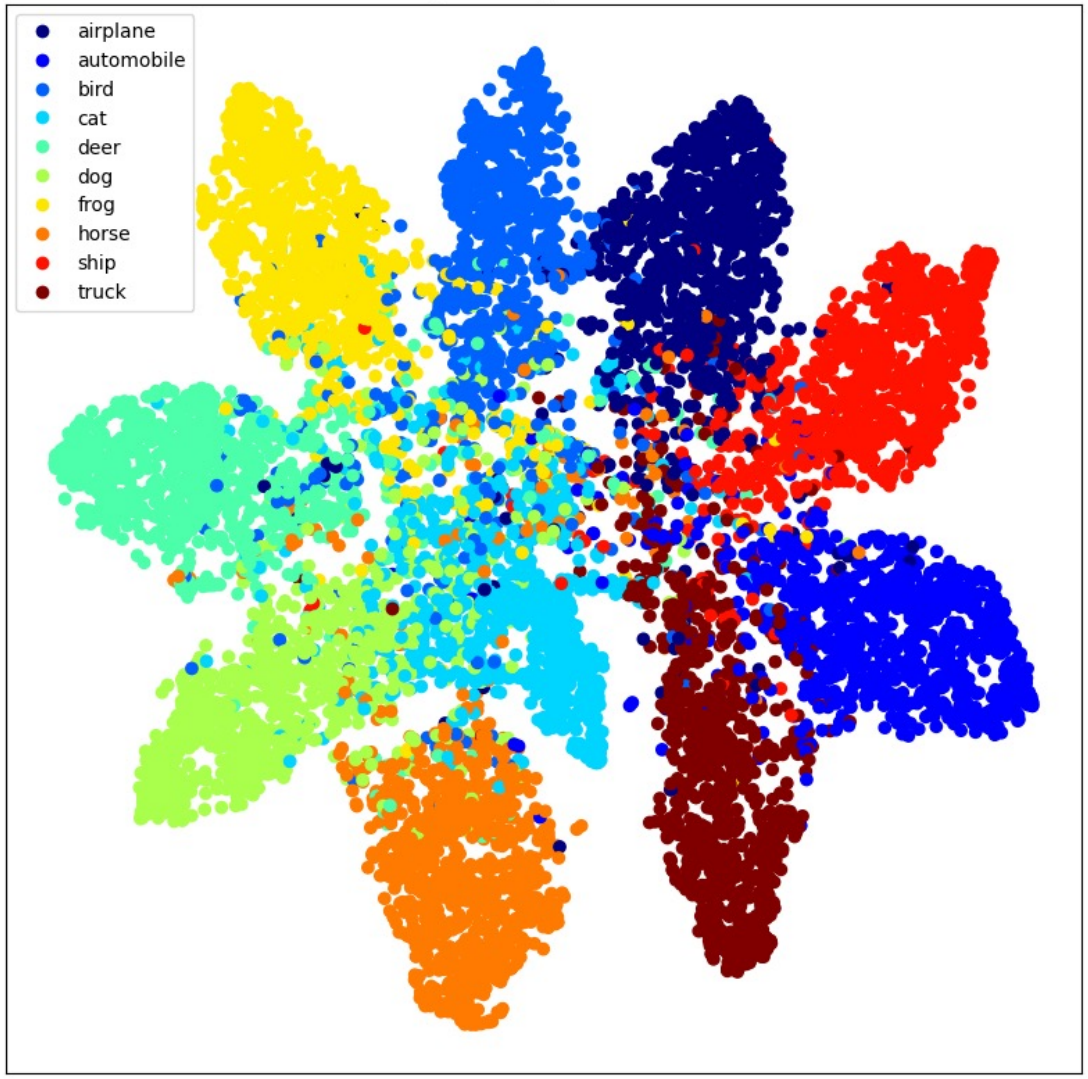
Comparisons



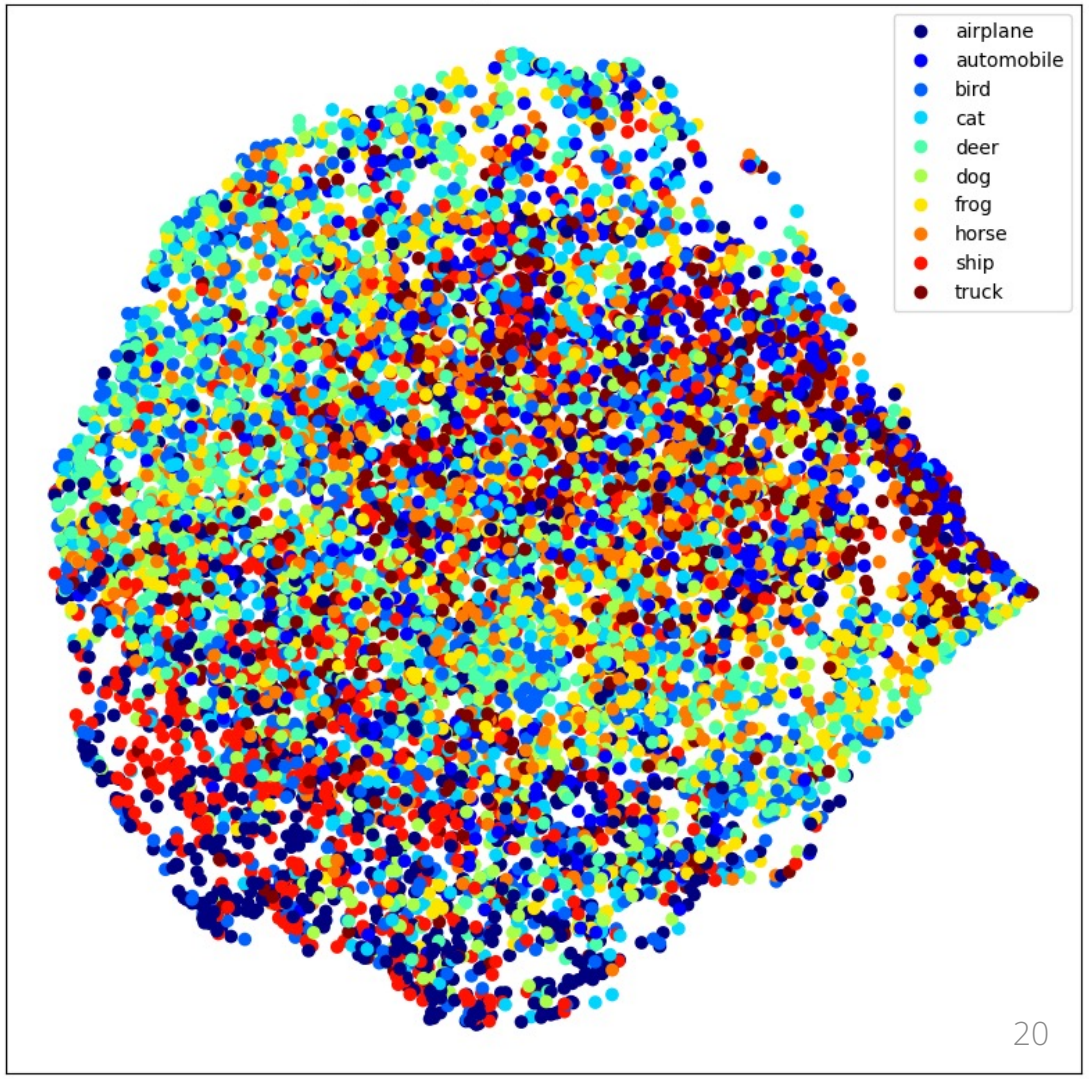


Comparisons

t-SNE embedding of last layer activations - ResNet20

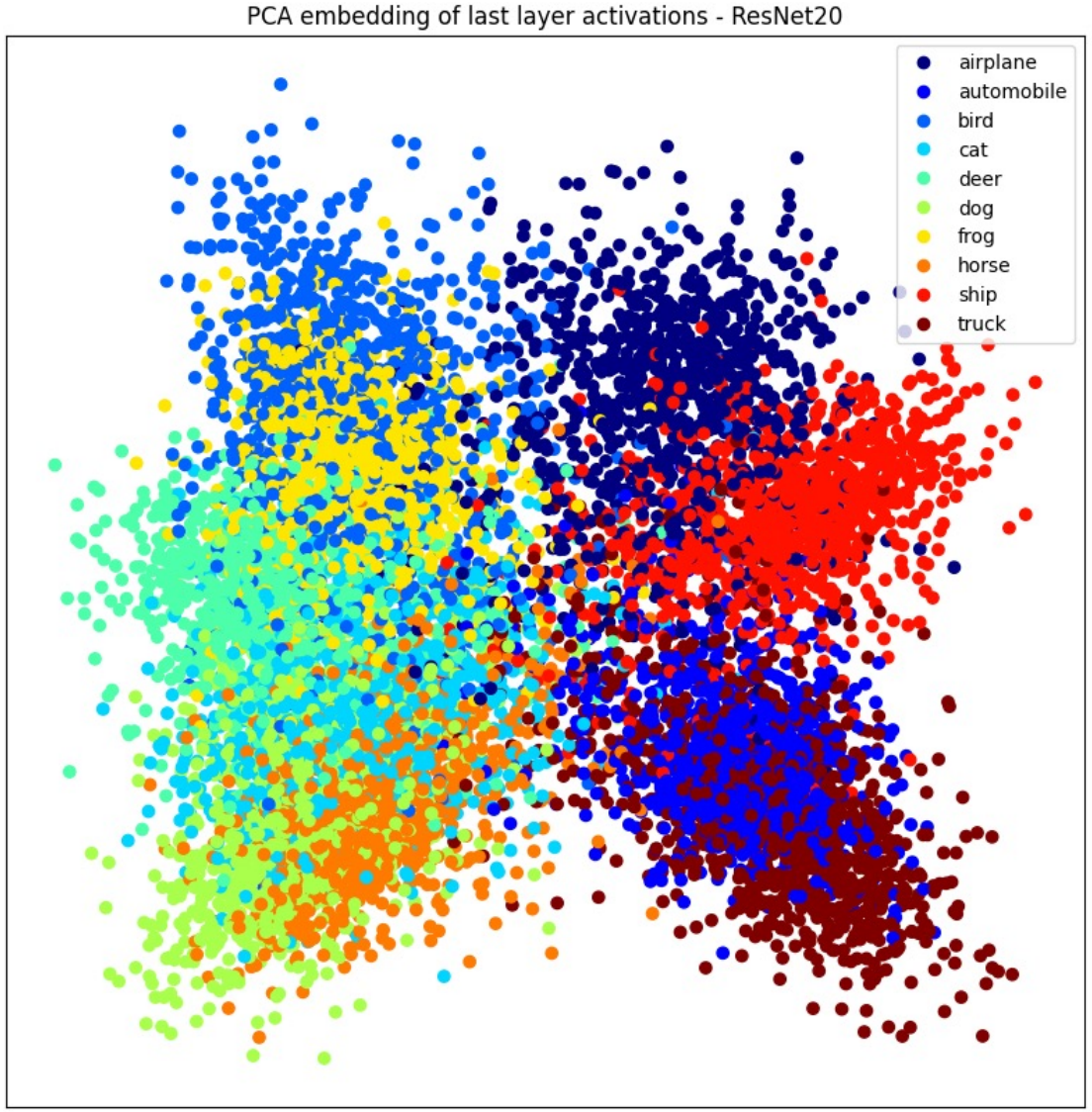


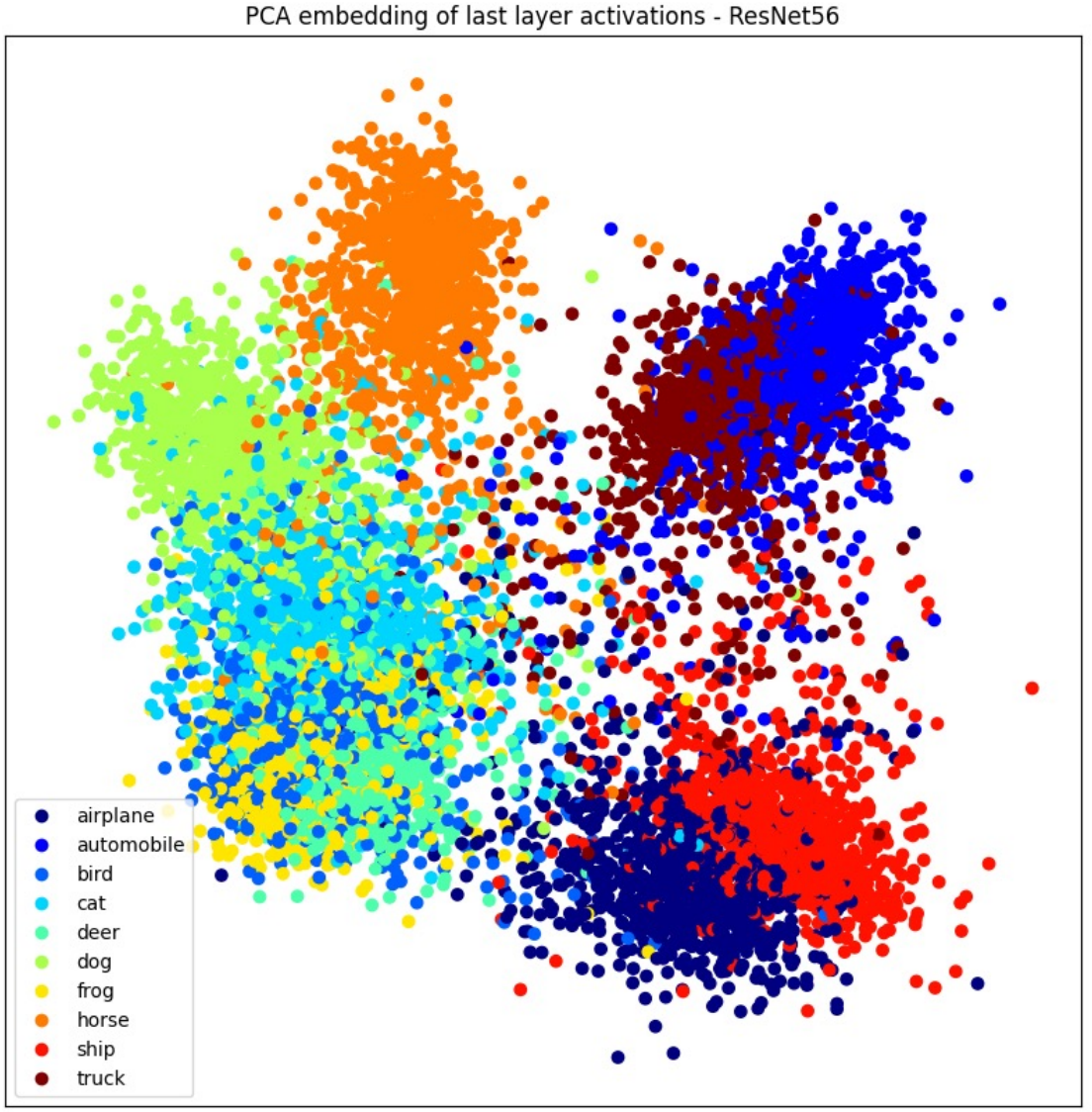
t-SNE embedding of last layer activations - random init ResNet20



ResNet20 – 92.6% Accuracy
ResNet56 – 94.4% Accuracy

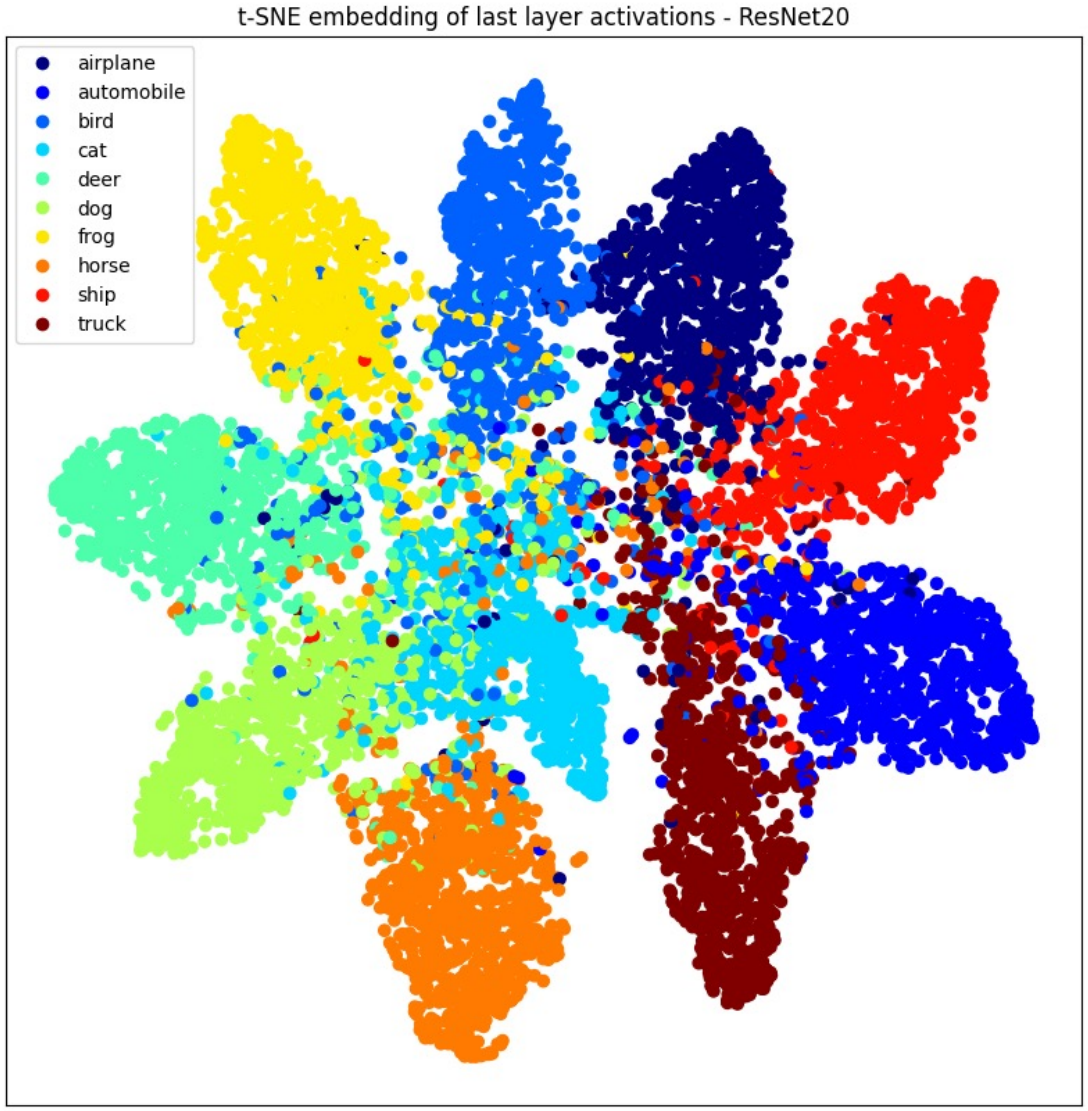
Comparisons

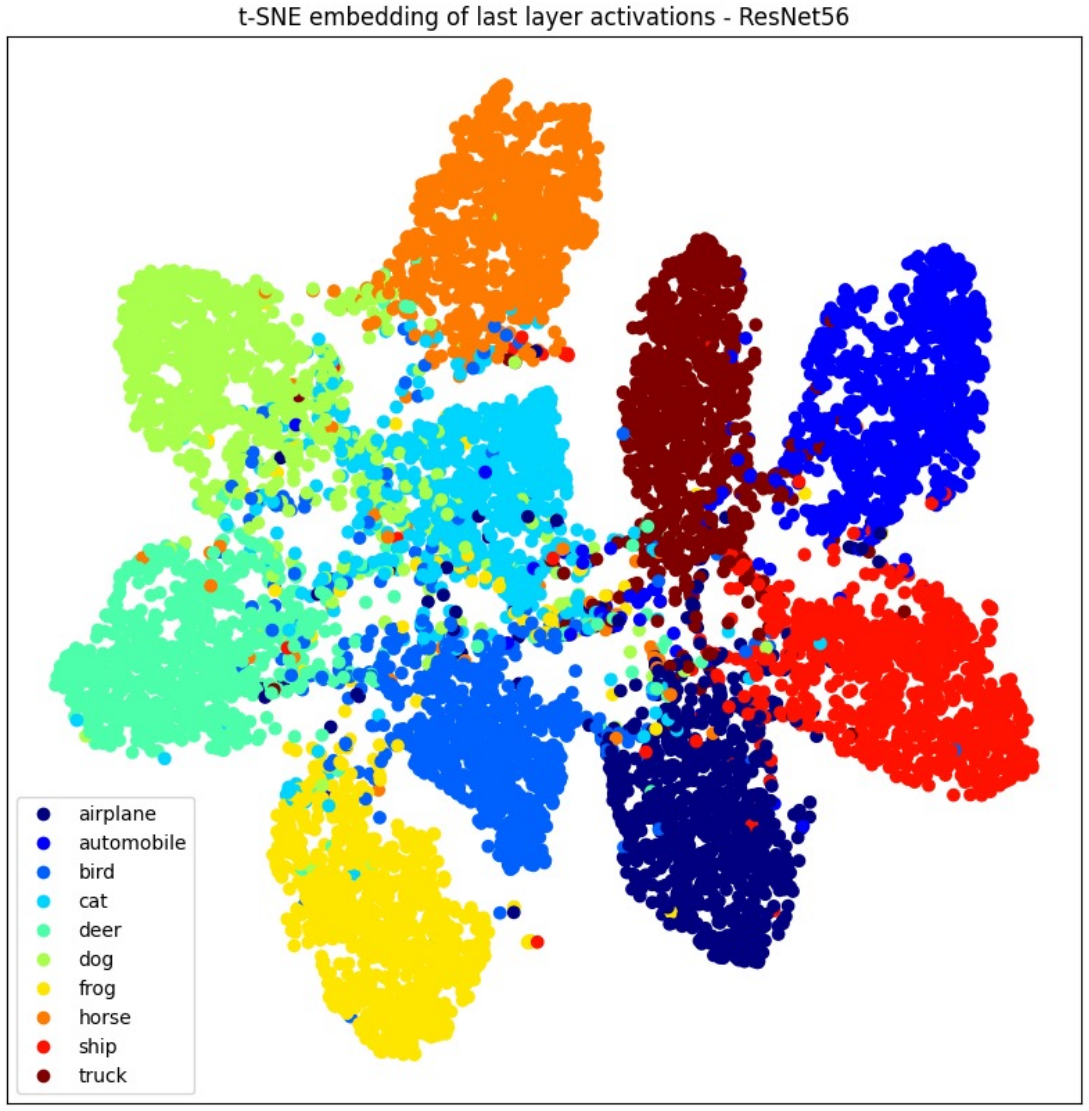




ResNet20 – 92.6% Accuracy
ResNet56 – 94.4% Accuracy

Comparisons



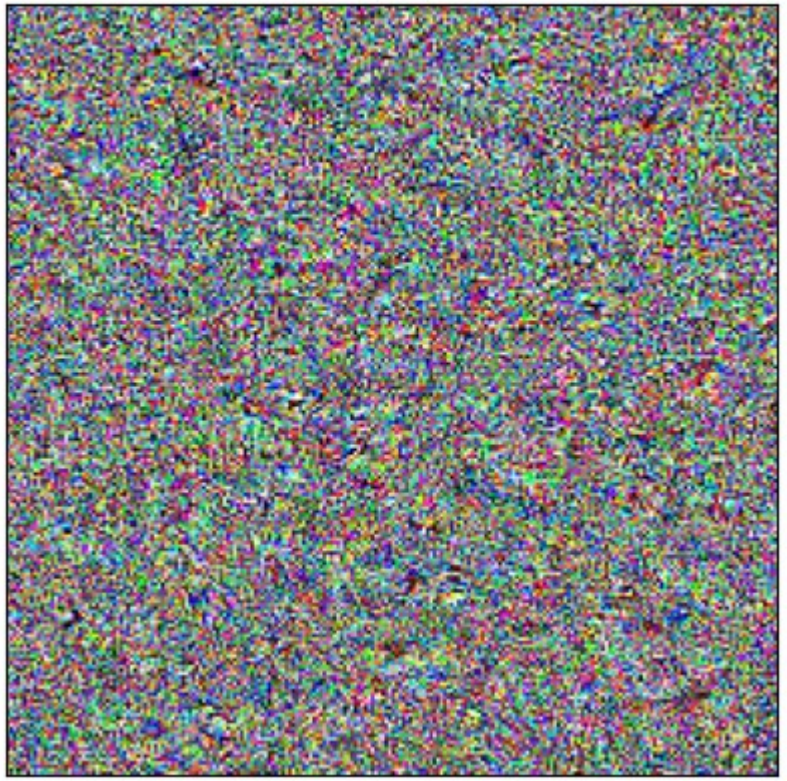


Comparisons

- Large visual difference to random network – we have clearly learned *something*!
- Differences between trained networks small, and hard to interpret.
- Careful: t-SNE uses randomness – every run will show you a different embedding.

Input Reconstruction

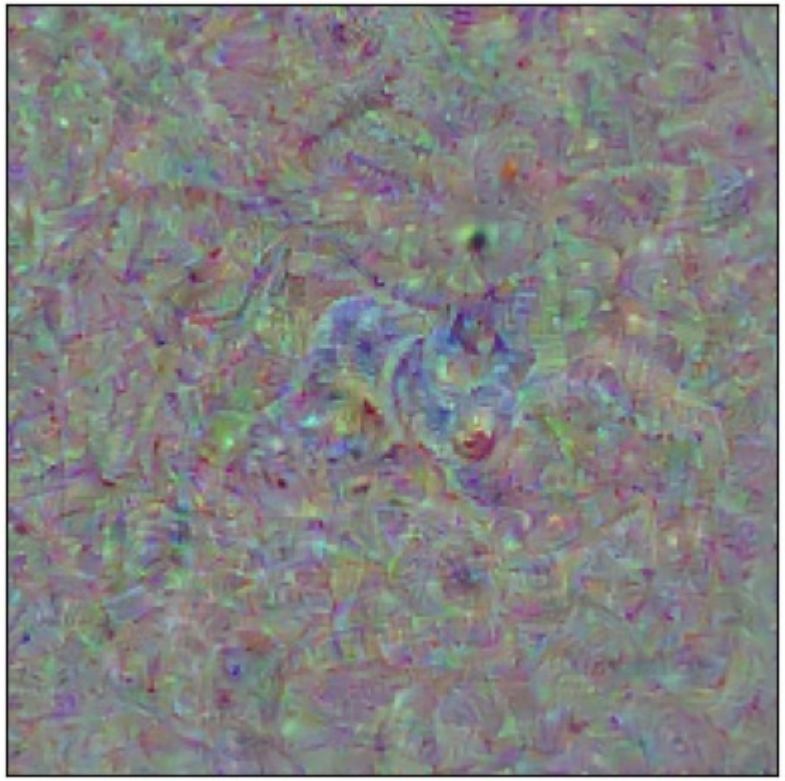
- To understand what a model has learned, we can also search for an input that maximised a class probability.
- Search with gradient ascent on the image.
- But: generates adversarial example.



Input Reconstruction - Tricks

- Regulariser: smoothness (total variation = L1 on image gradients).
- Image jittering: randomly move image by some pixels at every step.
- Better regularisers: better reconstructions.
- Works also for intermediate neurons.
- https://distill.pub/2017/feature-visualization/

Input maximisation class dingo



Understanding Samples

- We can also try to understand the decision process for a single sample.
- ResNet50: “Dingo”
- ResNet18: “Bucket”
- ConvNeXt_Large: “Dingo”

Input image

Dingo

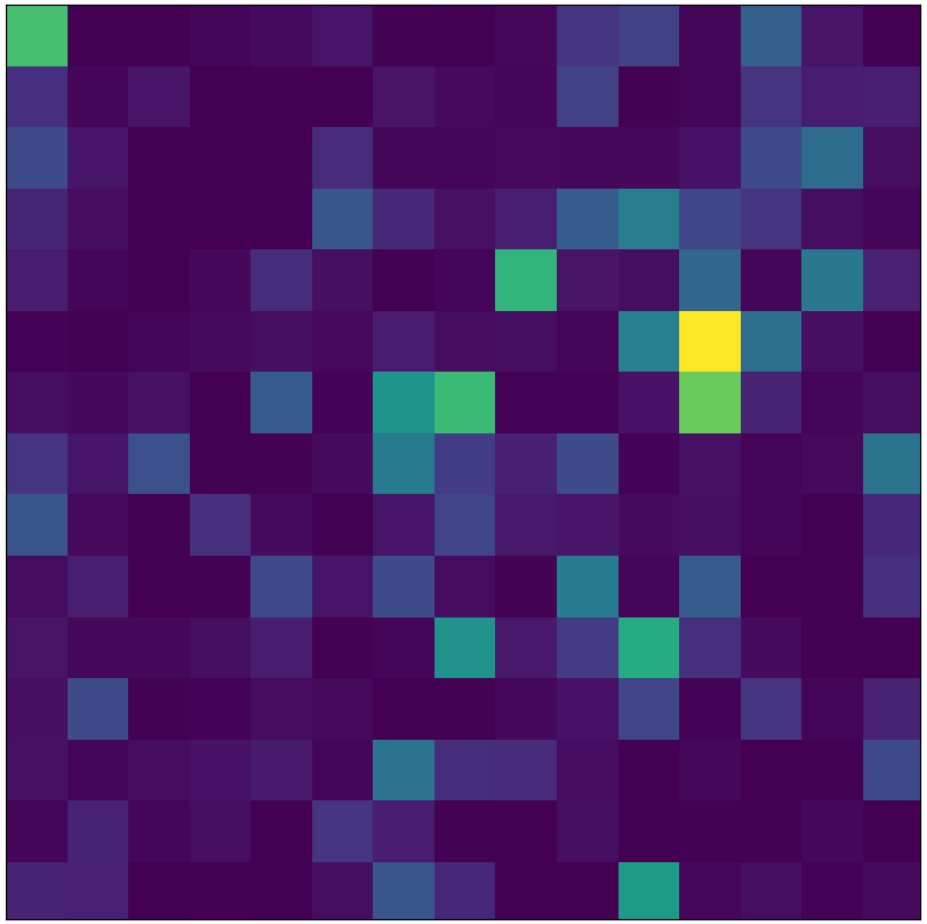
https://en.wikipedia.org/wiki/Dingo

Black-Box Attribution

- No access to model itself, observe only input/output.
- Idea: make changes to the input and observe what happens.
- Occlusion method (Zeiler & Fergus, 2015)
  - Occlude a part of the image and measure the change in response.
  - The bigger the change, the more important was the occluded region.
  - Measure the change in target/predicted class probability (other classes can change too, but do not matter)

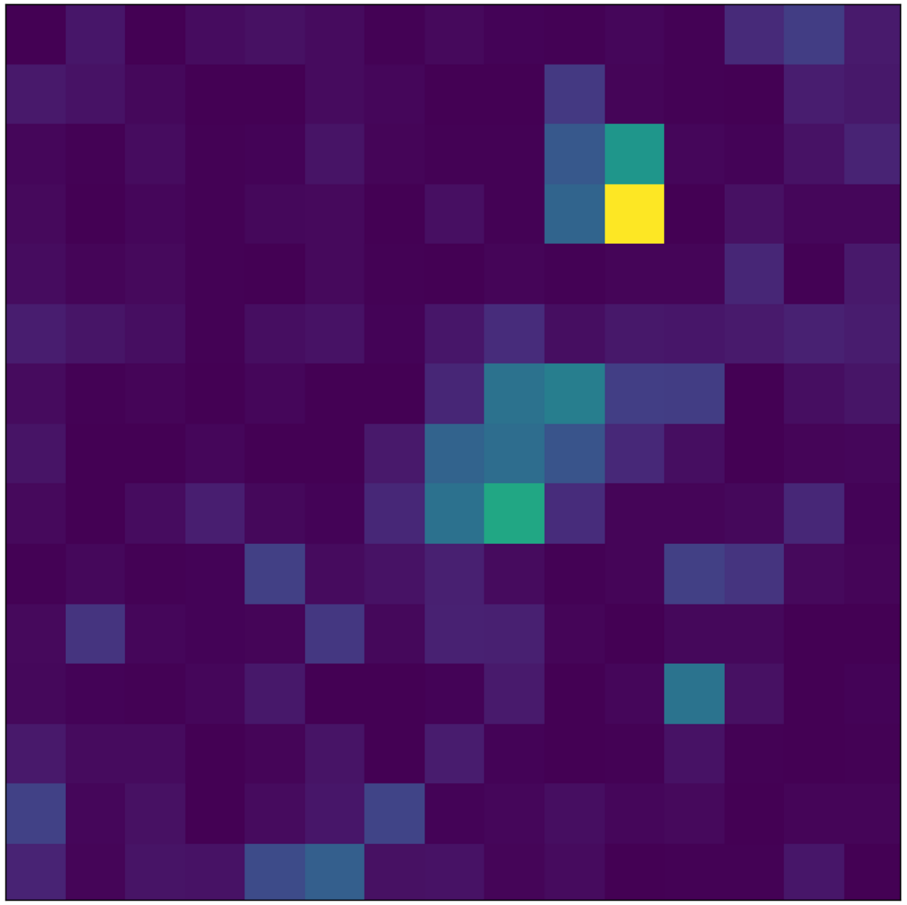
Occlusion Method

ResNet50: “Dingo”



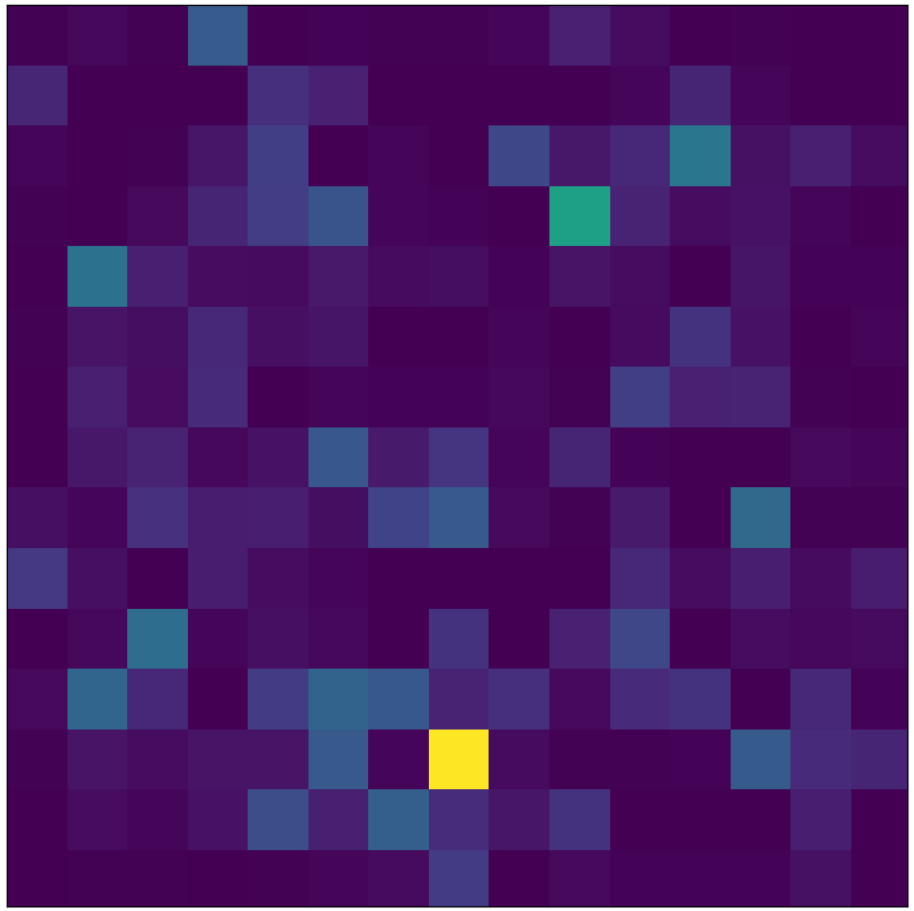
Occlusion Method

ResNet18: “Bucket”



Occlusion Method

ResNet18 (random init.): “Gong”



Occlusion Method

- Depends on this size of the occlusion.
- What do we fill in when we occlude? (0, random noise, avg, …)
- Is a square occlusion meaningful?

- Slow: needs many network evaluations – one for each patch.

Several improvements, for example:

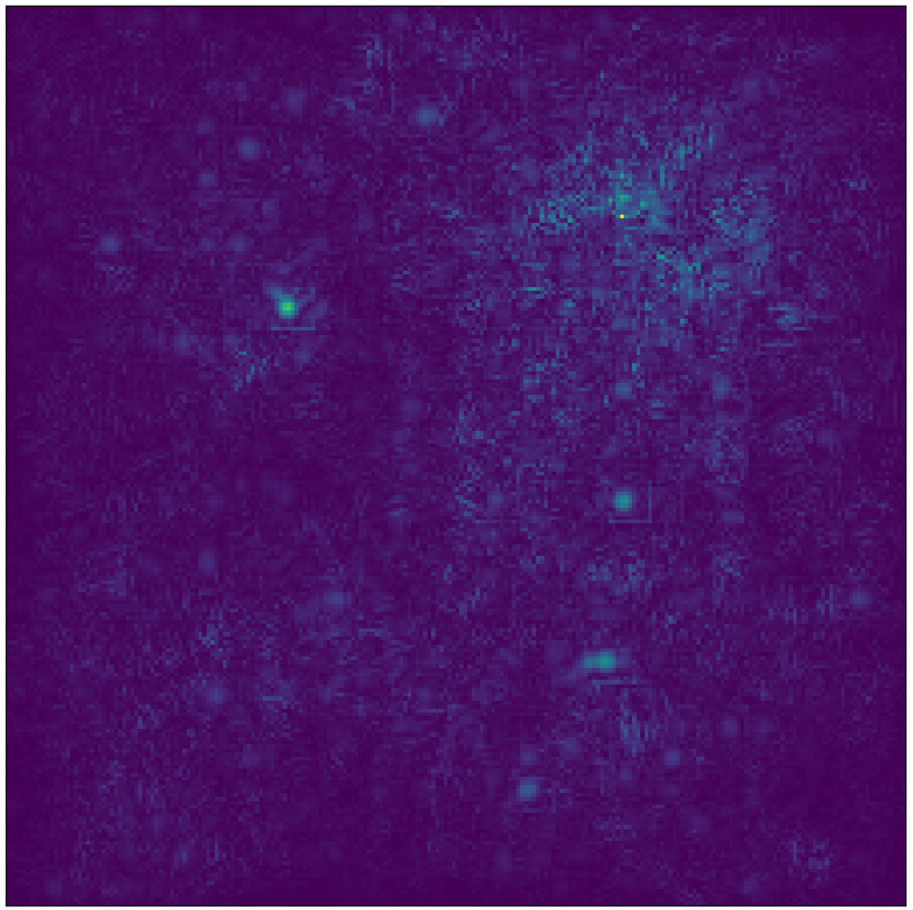
- Fong, Ruth C., and Andrea Vedaldi, Interpretable explanations of black boxes by meaningful perturbation ICCV, 2017

White-Box Attribution

- We do have access to the weights and computations inside the model.
- How can we use this information to extract understanding.
- Idea: use the gradient magnitude $|\nabla_x f(x)|_1$.
- “Which direction does the input need to change to affect the output the most.”

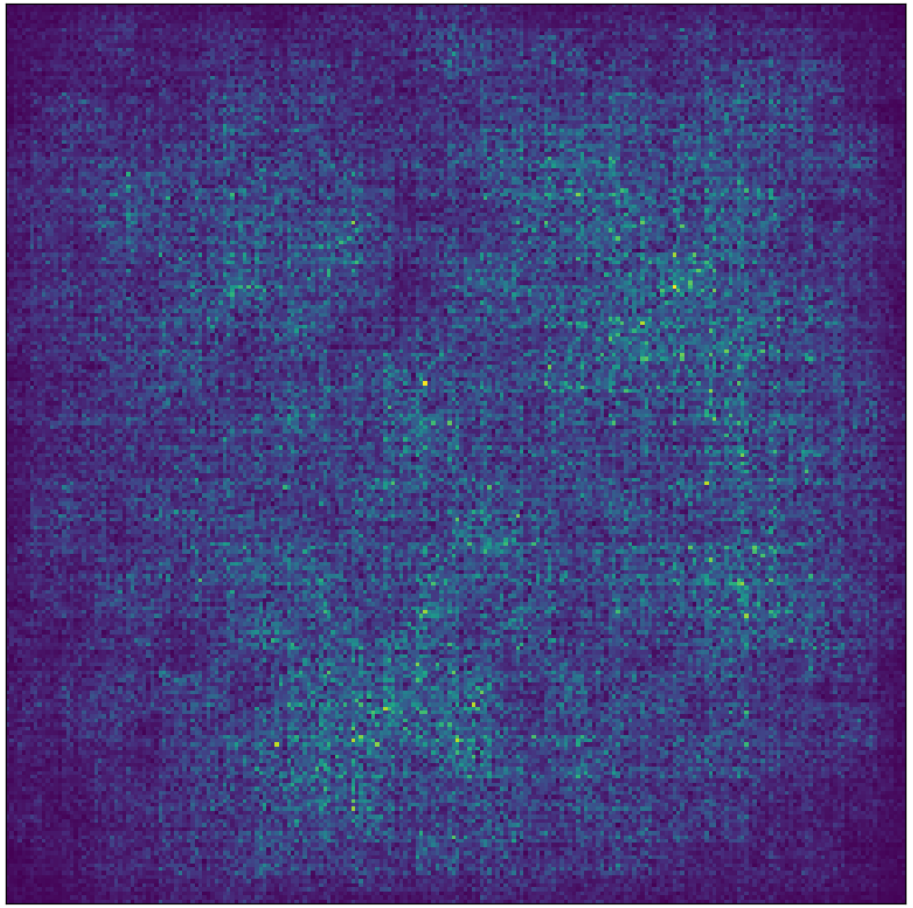
Gradient Method

AlexNet: “Mountain Lion”



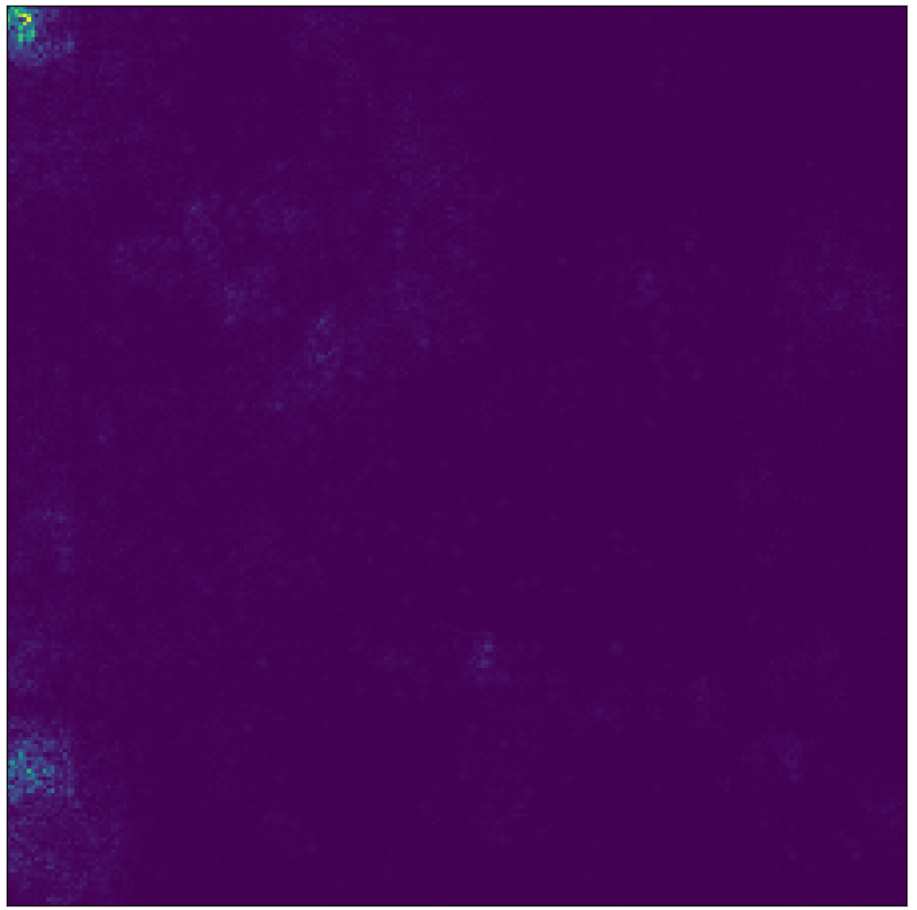
Gradient Method

AlexNet (random init)



Gradient Method

ResNet50 “Dingo”



Gradient Methods

- Not limited to last layer.

- Several improved variants.
- Mainly: ideas how to deal with ReLU and pooling layers.

- Observation: lower dependence on network weights.

- How can we benchmark visualisation techniques?

Attribution Methods

- Visualisation techniques that highlight which input pixels are important are often called *Attribution Methods* or *Saliency Methods*.

- The (un) reliability of saliency methods, Kindermans, Hooker, et al., 2017

- A benchmark for interpretability methods in deep neural networks, Hooker et al., 2019

ROAR: Remove and Retrain

- Run your attribution method on the train & test set.
- For each image: sort all pixels by attribution performance.
- Delete X% of most important pixels.
- Retrain your network on this new data.

- Measure performance change on test set.
- If you removed many critical pixels, the performance will be lower.
- Need for re-training: images look very different after deletion.

ROAR

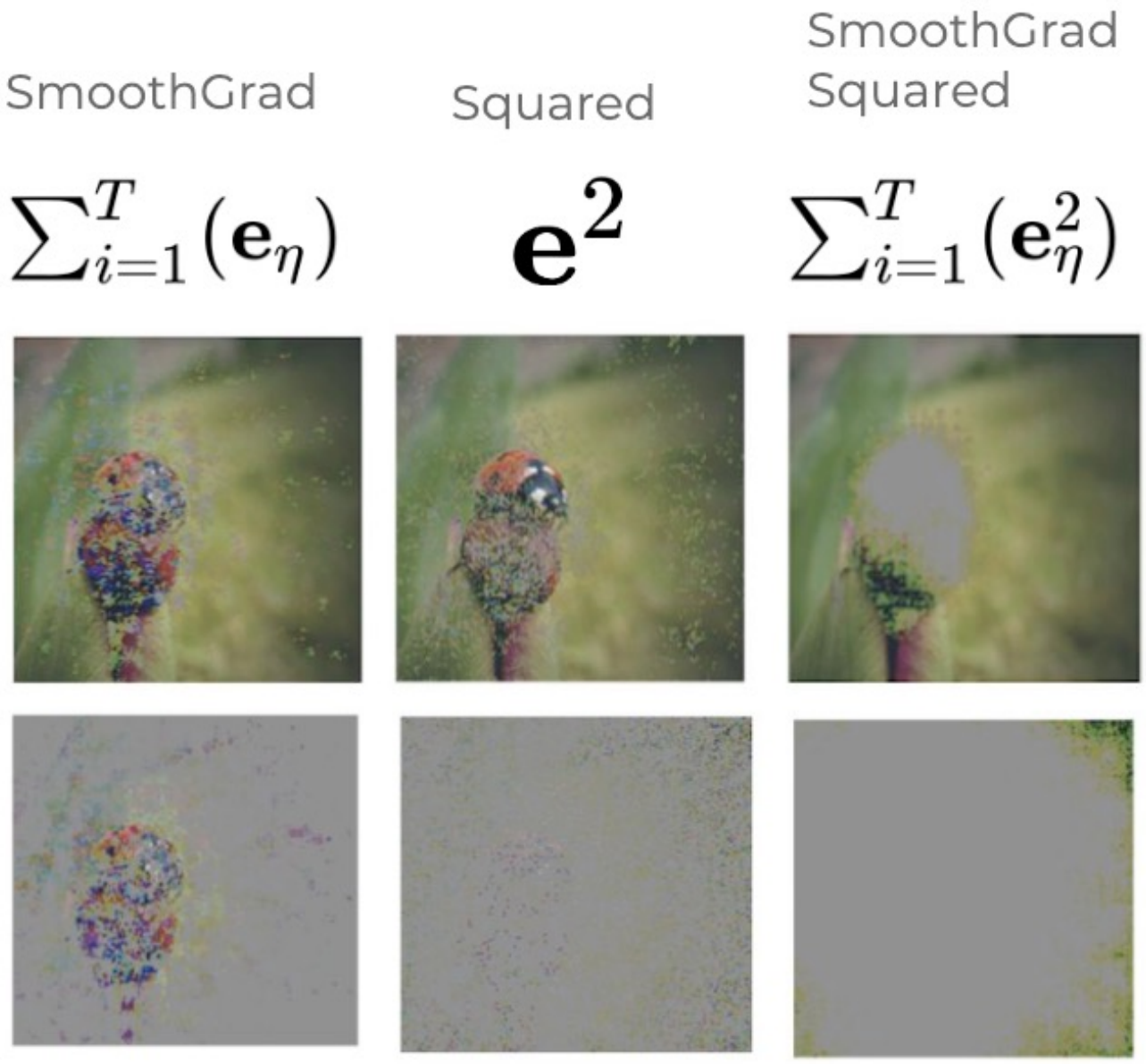
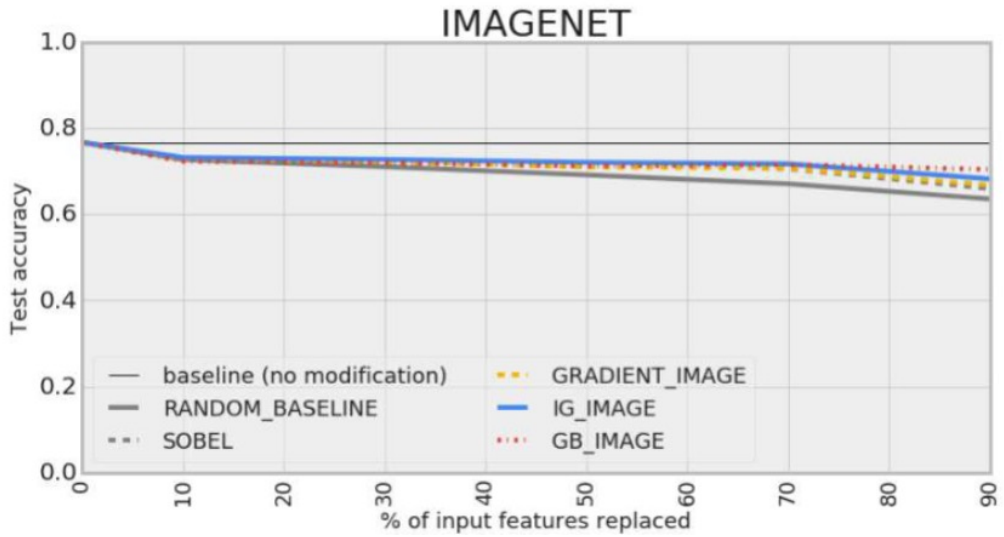


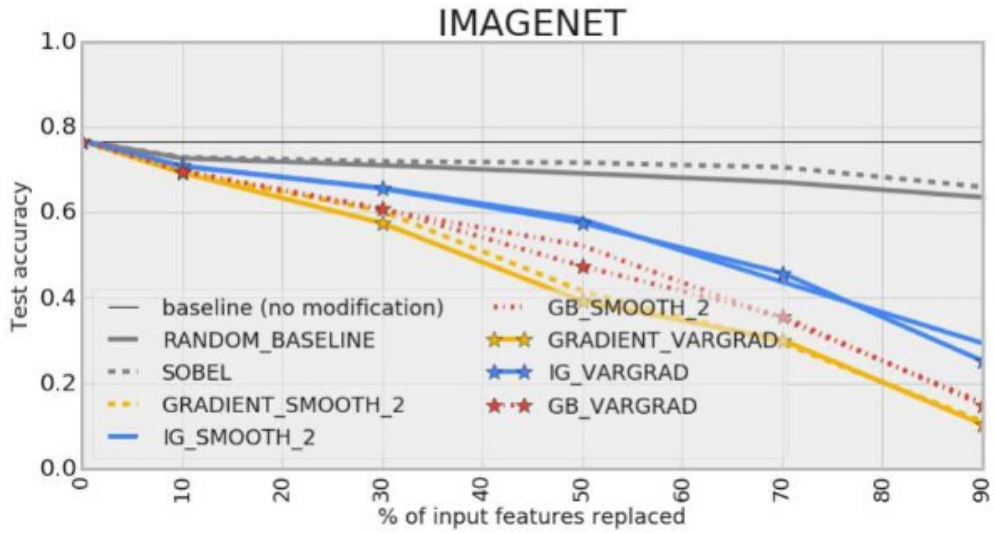
Image source

ROAR

- Gradient Image works even slightly worse than randomly deleting pixels.

- Ensemble approaches are much better: average the gradients over many small perturbations (add noise to the image).





A benchmark for interpretability methods in deep neural networks, Hooker et al., 2019

Sanity Checks

Sanity Checks for Saliency Maps, Adebayo et al, 2018.

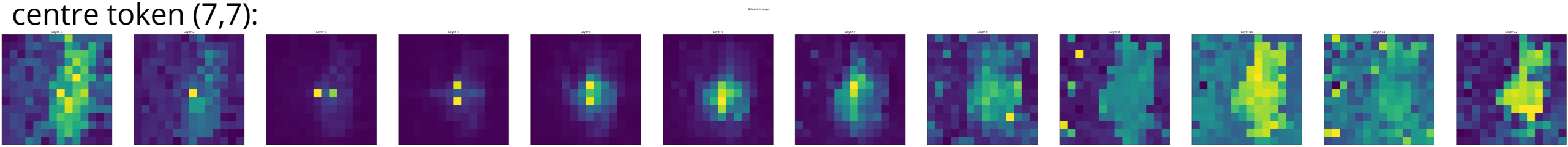
- Test 1: randomising the model weights should affect the attribution method. (Otherwise we are not visualising what the model has learned)

- Test 2: Train another model on the same data but random labels. This should also affect the visualisations.

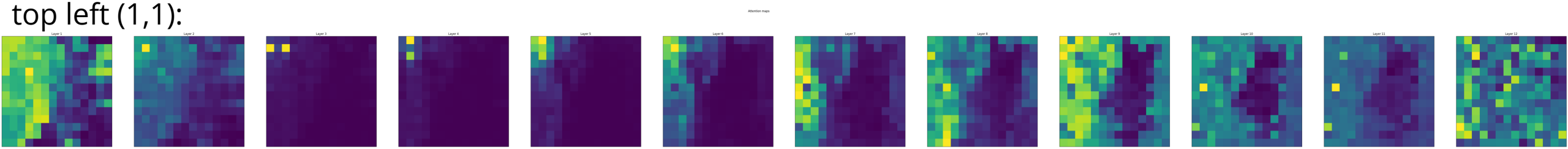
Attention

- Self-Attention on 14x14 patches means attention weights are a matrix of size 196x196 (or equiv. 14x14x14x14)
- For every token, 14x14 attention map for each layer (12).

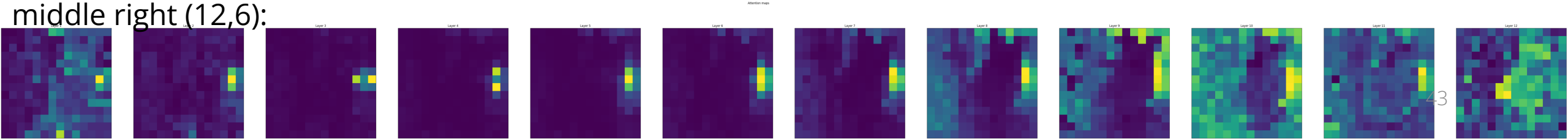
centre token (7,7):



top left (1,1):



middle right (12,6):



Visualising Attention

- Many choices: layer, token, MHA head.
- Trained models seem to do the “right thing”.
- Last layer looks task focused: most attention is on the object independent of which token we are looking at.
- Often difficult to choose what to visualise, some choice will always look similar to what you are looking for -> confirmation bias.

Taxonomy - Approaches

Post-Hoc Analysis

Explanations are derived from a fixed, pre-trained model via analysis.

- No impact on performance
- Difficult
- Explanations are often *local* around predictions
- *Main focus today!*

Transparent Models

The model is constructed such that (some) mechanism have semantic meaning.

- Does not need post-hoc analysis
- Task-specific architecture
- Can affect performance

Learned Explanations

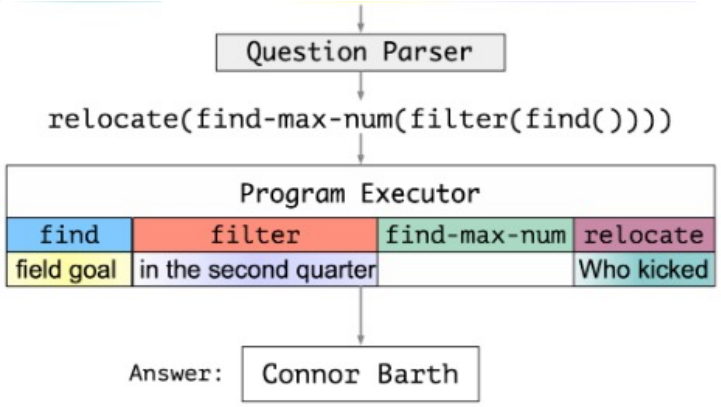
The model is trained to deliver explanations together with predictions.

- Explanations can be very semantic
- Might need meta-explanations
- Can affect performance

Example: Transparent Models

In the first quarter, Buffalo trailed early as Chiefs QB Tyler Thigpen completed a 36-yard TD pass to RB Jamaal Charles. The Bills responded with RB Marshawn Lynch getting a 1-yard TD run. In the second quarter, Buffalo took the lead as kicker Rian Lindell made a 21-yard field goal. Kansas City answered with Thigpen completing a 2-yard TD pass to TE Tony Gonzalez. Buffalo regained the lead as Lindell got a 39-yard field goal, while rookie CB Leodis McKelvin returned an interception 64 yards for a touchdown. The Chiefs struck back with kicker Connor Barth getting a 45-yard field goal, yet the Bills continued their offensive explosion as Lindell got a 34-yard field goal, along with QB Trent Edwards getting a 15-yard TD run.
In the third quarter, Buffalo continued its poundings with Edwards getting a 5-yard TD run, while Lindell got himself a 38-yard field goal. Kansas City tried to rally as Thigpen completed a 45-yard TD pass to WR Mark Bradley, yet the Bills replied with Edwards completing an 8-yard TD pass to WR Josh Reed. In the fourth quarter, Buffalo pulled away as Edwards completed a 17-yard TD pass to TE Derek Schouman.

Who kicked the longest field goal in the second quarter?



- Decompose the problem in smaller parts that can be interpreted individually
- Increases interpretability of the whole system
- Some steps might need further decomposition/explanation

Neural Module Networks for Reasoning over Text
Gupta, N., Lin, K., Roth, D., Singh, S., & Gardner, M.
ICLR 2019

Example: Learned Explanations

*This is a **Downy Woodpecker** because...*

Explanation: this is a black and white bird with a **red spot** on its crown.

*This is a **Downy Woodpecker** because...*

Explanation: this is a white bird with a black wing and a black and white striped head.

- The model predicts an explanation
- Training contains explanations together with input-output pairs
- Explanation needs to be both:
  - input specific
  - output specific
- How do we explain the explanation?

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T.
Generating visual explanations
European Conference on Computer Vision, 2016