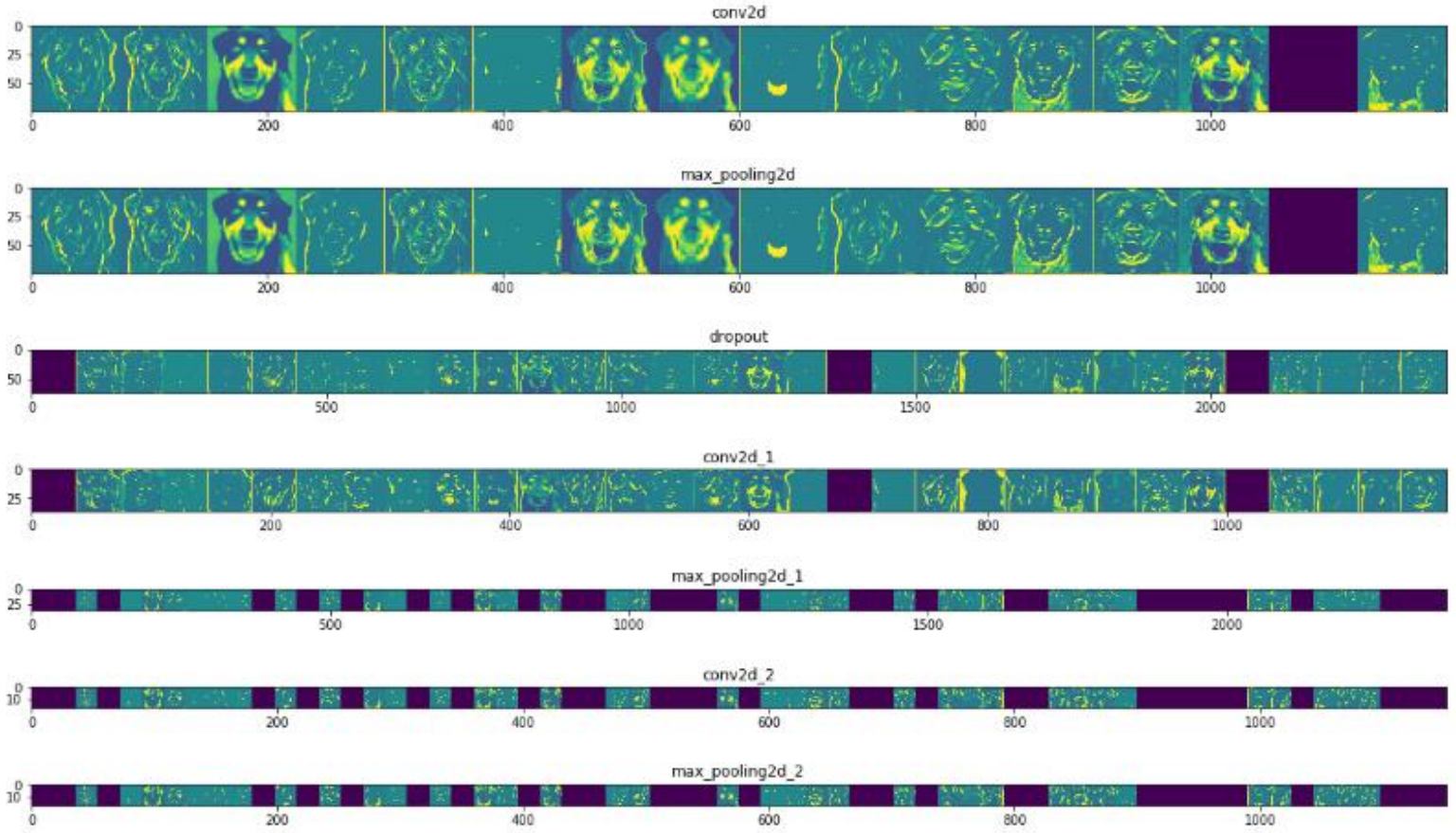
Visualizations



Feature maps

- Convolution activations == feature maps
- A deep network has several hierarchical layers
  - hence several hierarchical feature maps going from less to more abstract

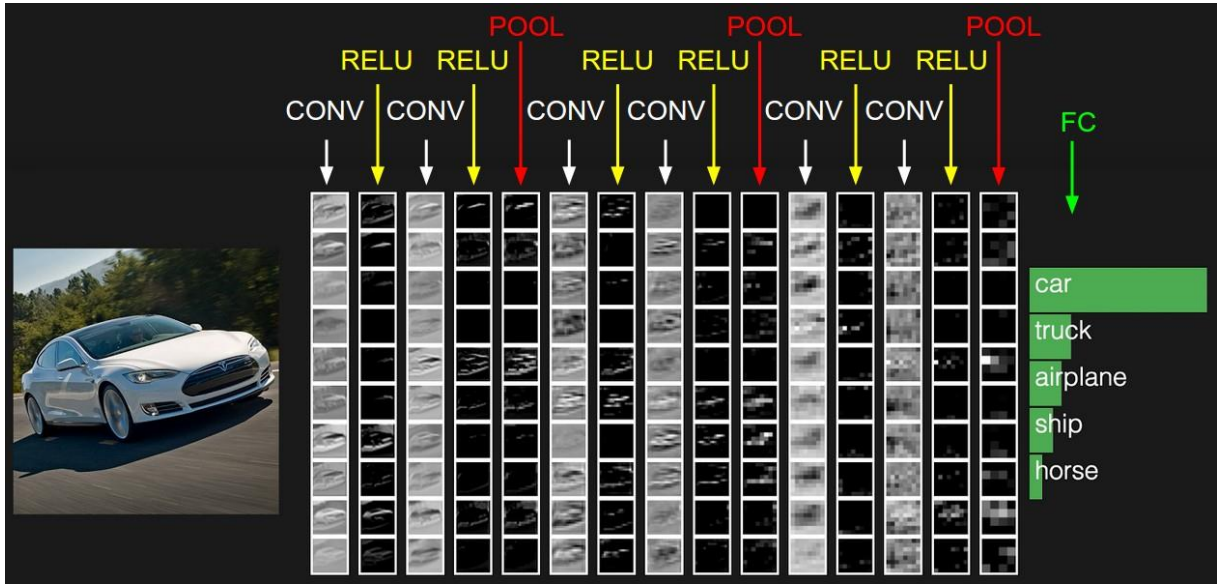
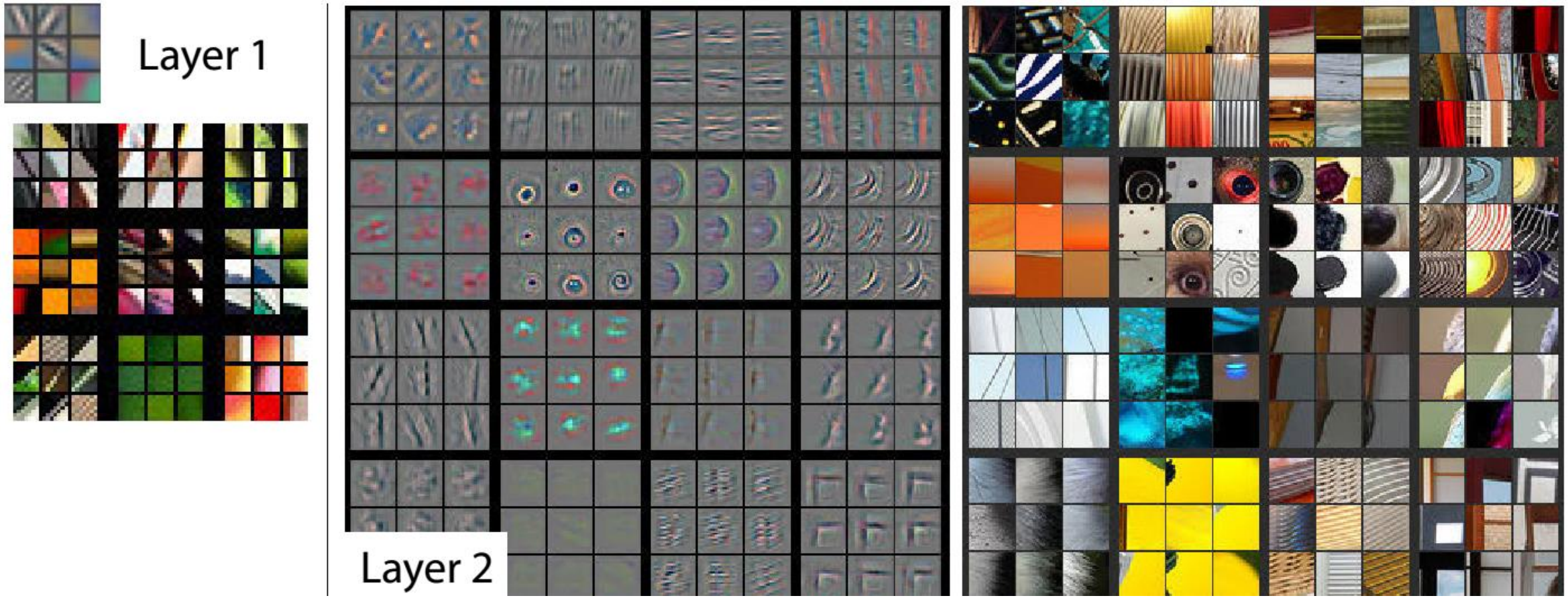


Image borrowed by A. Karpathy

What excites feature maps?

"Given a random feature map what are the top 9 activations?"



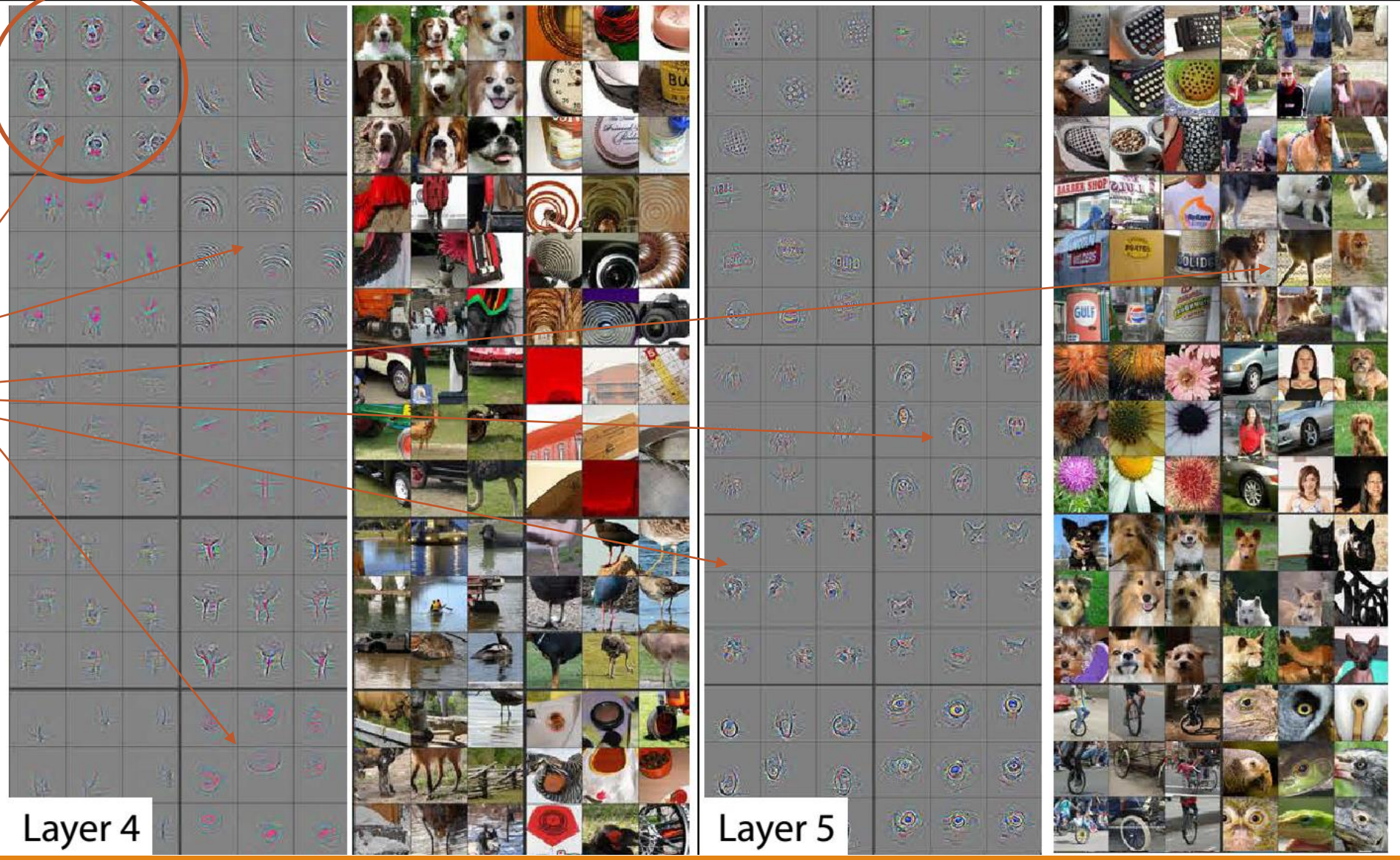
What excites feature maps?

Similar activations from lower level visual patterns

What excites feature maps? [Zeiler2014]

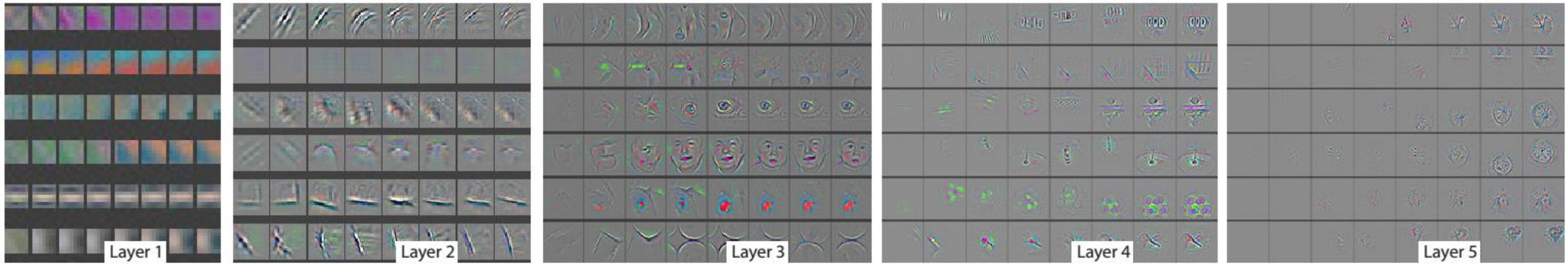
Similar activations from semantically similar pictures

Visual patterns become more and more intricate and specific (greater invariance)

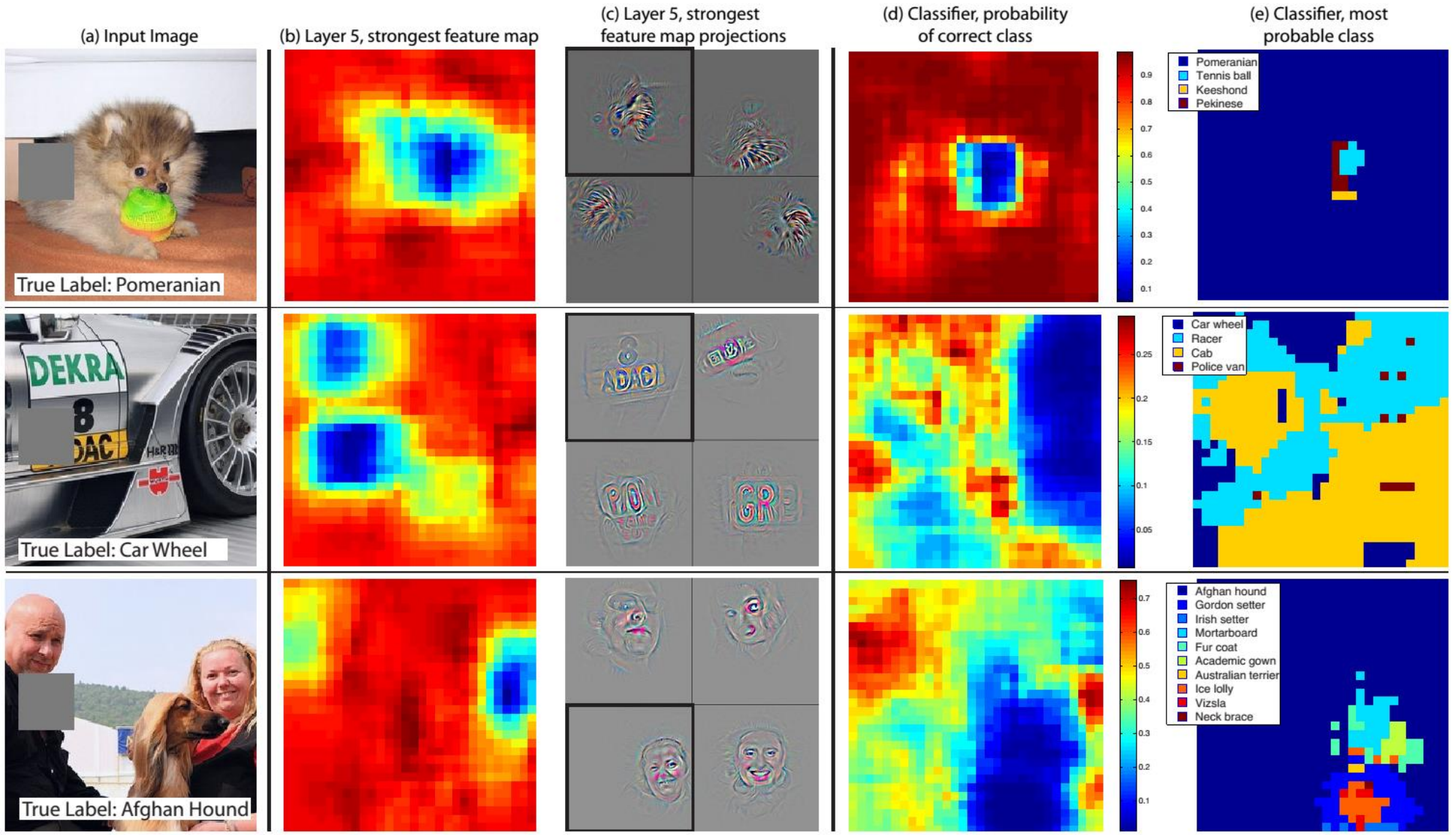


Feature evolution over training

- Given a neuron (outputs a single feature map)
  - Strongest activation during training for epochs 1, 2, 5, 10, 20, 30, 40, 64



But does a Convnet really learn the object?



What is a “Convnet dog”, however? [Simonyan2014]

- Find the most “dumbell”/”cup”/”dalmatian”/… image

$$\arg\max_I \text{Score}(I) - \lambda \|I\|_2^2$$

- Can be adapted for image-specific class saliency $\propto \left.\frac{\partial Score}{\partial I}\right|_{I_0}$