
University of Colorado
Boulder

Deep Learning Applications for Computer Vision

Lecture 18: Visualizing CNNs

Visualizing the CNNs

w, b

- What are the parameters the model has learned?
  - What do the filters look like?

Visualizing the CNNs

- What are the parameters the model has learned?
  - What do the filters look like?
- What features did the model learn?
  - Edges, blobs, corners, … , eyes, fur, …

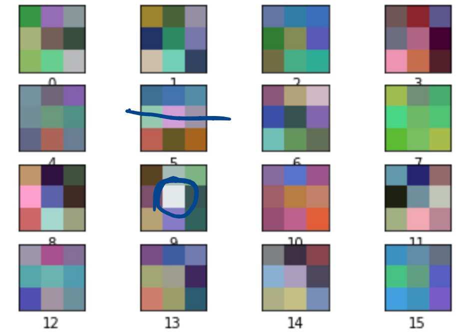
Visualizing the CNNs

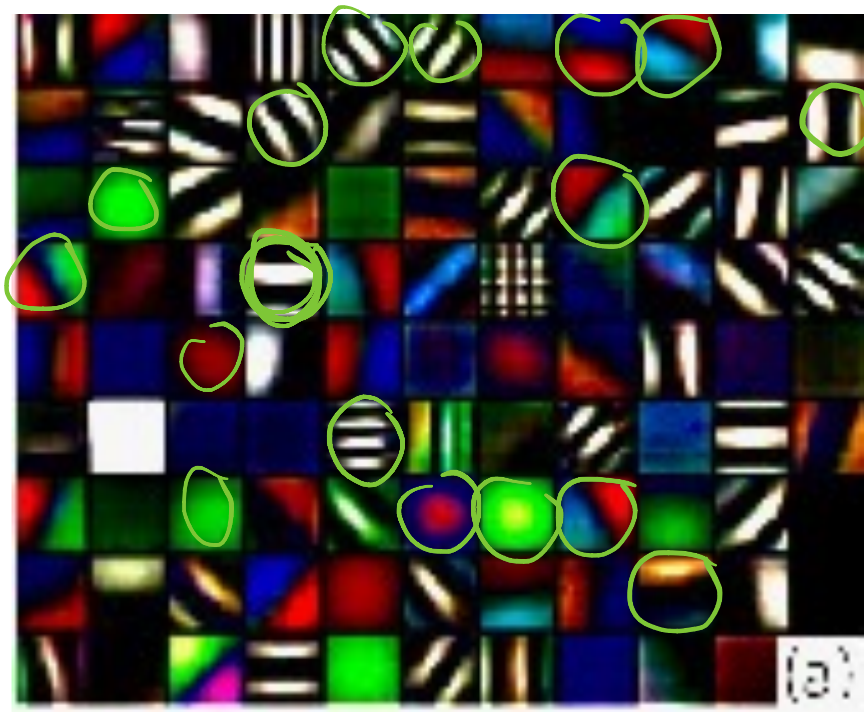
- What are the parameters the model has learned?
  - What do the filters look like?
- What features did the model learn?
  - Edges, blobs, corners, … , eyes, fur, …
- How does the model decide the category for a new image?
  - What do activation maps look like?
  - Which pixels in an image carry more "weight" in the classification decision?

Visualizing the filters

3x3x3

- For first convolutional layer
- AlexNet first layer:
  - 64 filters of size 11 x 11 x3





- edges - oriented
- blobs
- opposing colors

Figure 6 a) from Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Visualizing the filters

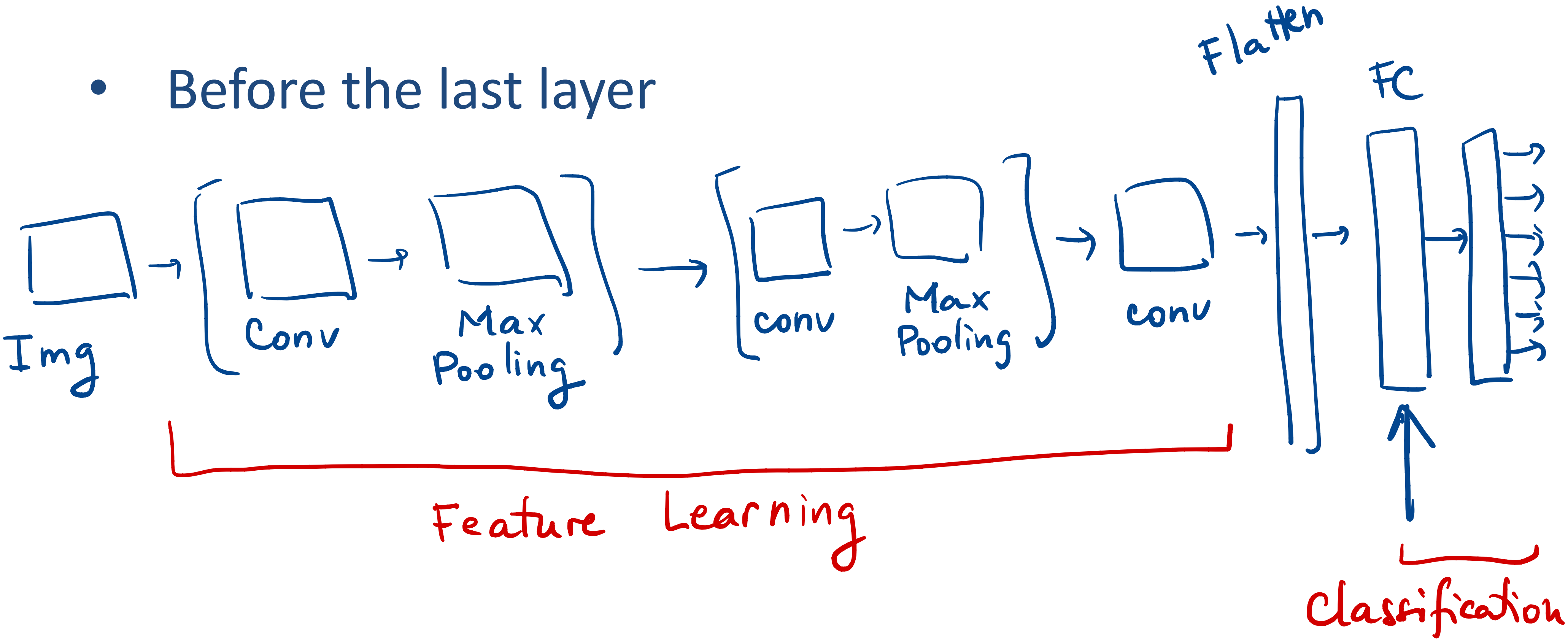
- For second convolutional layer
  - does not have the shape of an image

  (in our last tutorial second layer: 32 filters of size 3 x 3 x 32)
- We could visualize 32 * 32 greyscale images of size 3 x 3 …
- The input into the second conv layer is not an image …

Difficult to interpret

The final “Feature vectors”

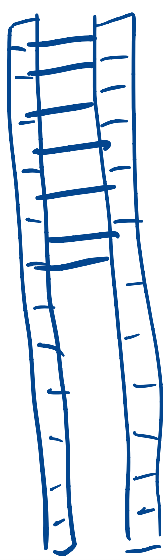
- Before the last layer



The final “Feature vectors”

- Before the last layer

- Flatten to 1024, 4096 column vector



Clustering : use distance metric

Nearest neighbor clustering

- Semantic clustering: images that are different in pixel intensities cluster together in feature space

Classification decisions

- What pixels influence the classification decision most?
- Is the model truly identifying the object, or using surrounding data

Occlusion Sensitivity

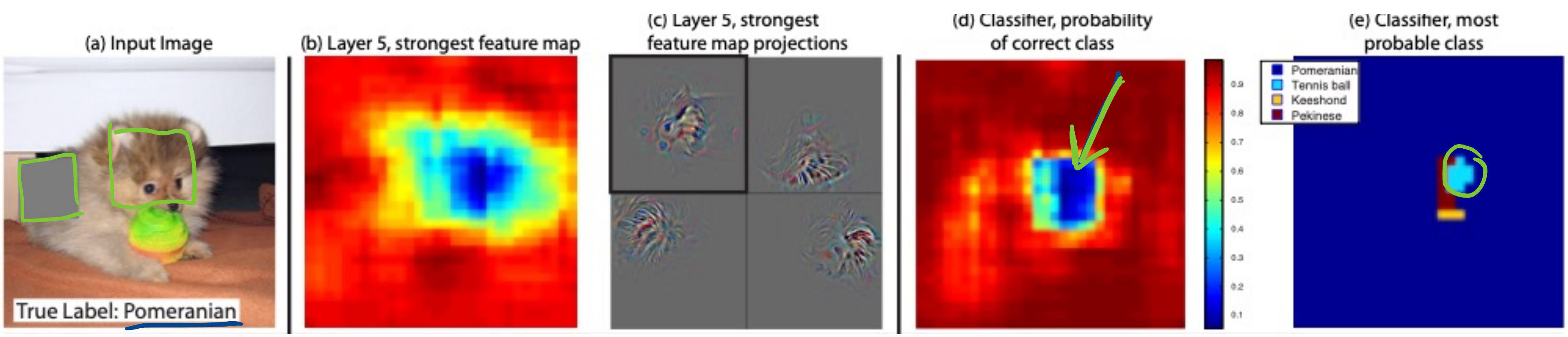


Figure 7 (first row) from Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

95 %

Occlusion Sensitivity

- Bad data example:

Husky vs Wolf

! All wolf pictures had snow

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Figure 11 and Table 2 from Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin - *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 2016