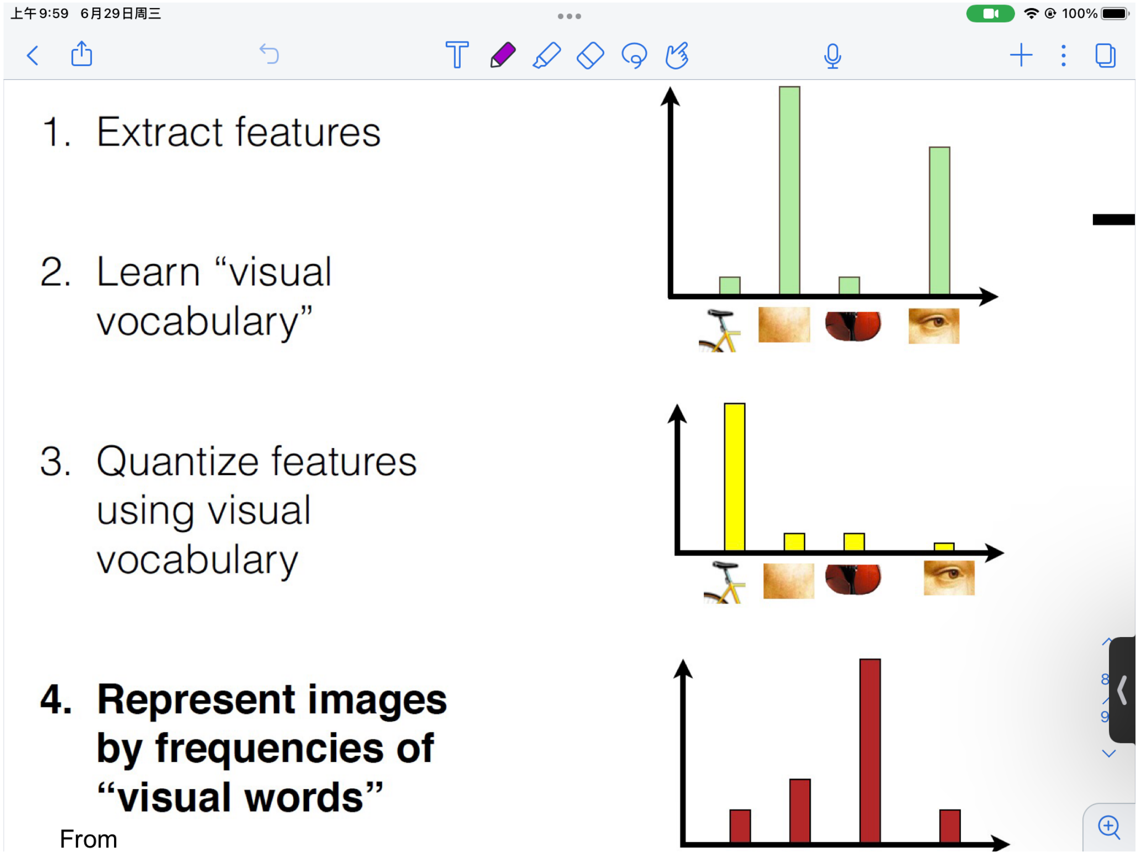


1. Extract features

2. Learn “visual vocabulary”

3. Quantize features using visual vocabulary

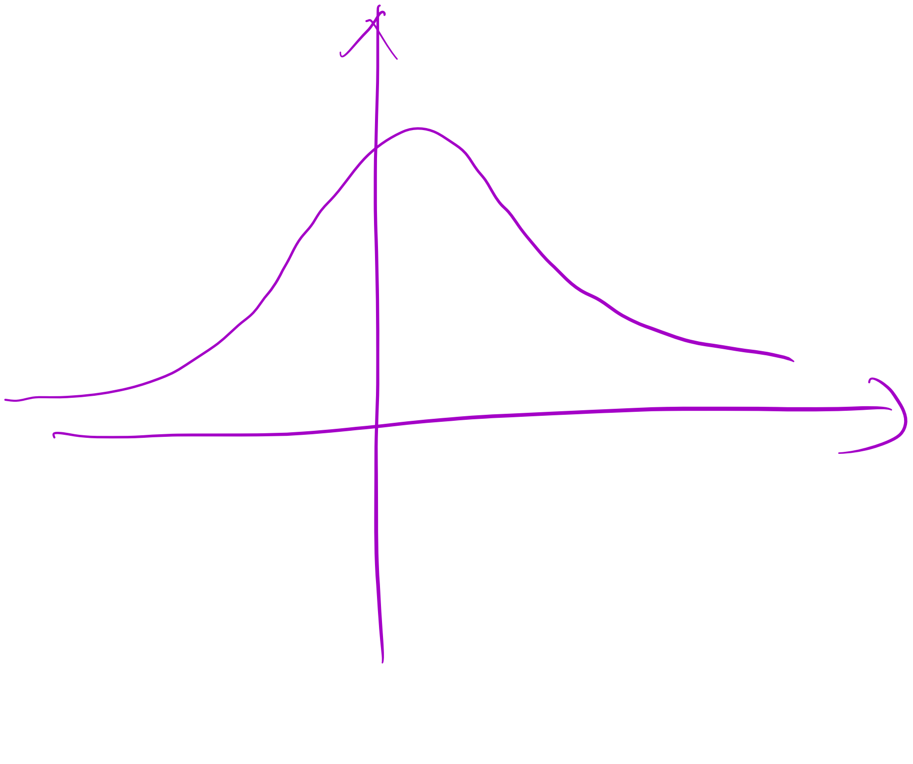
4. **Represent images by frequencies of “visual words”**

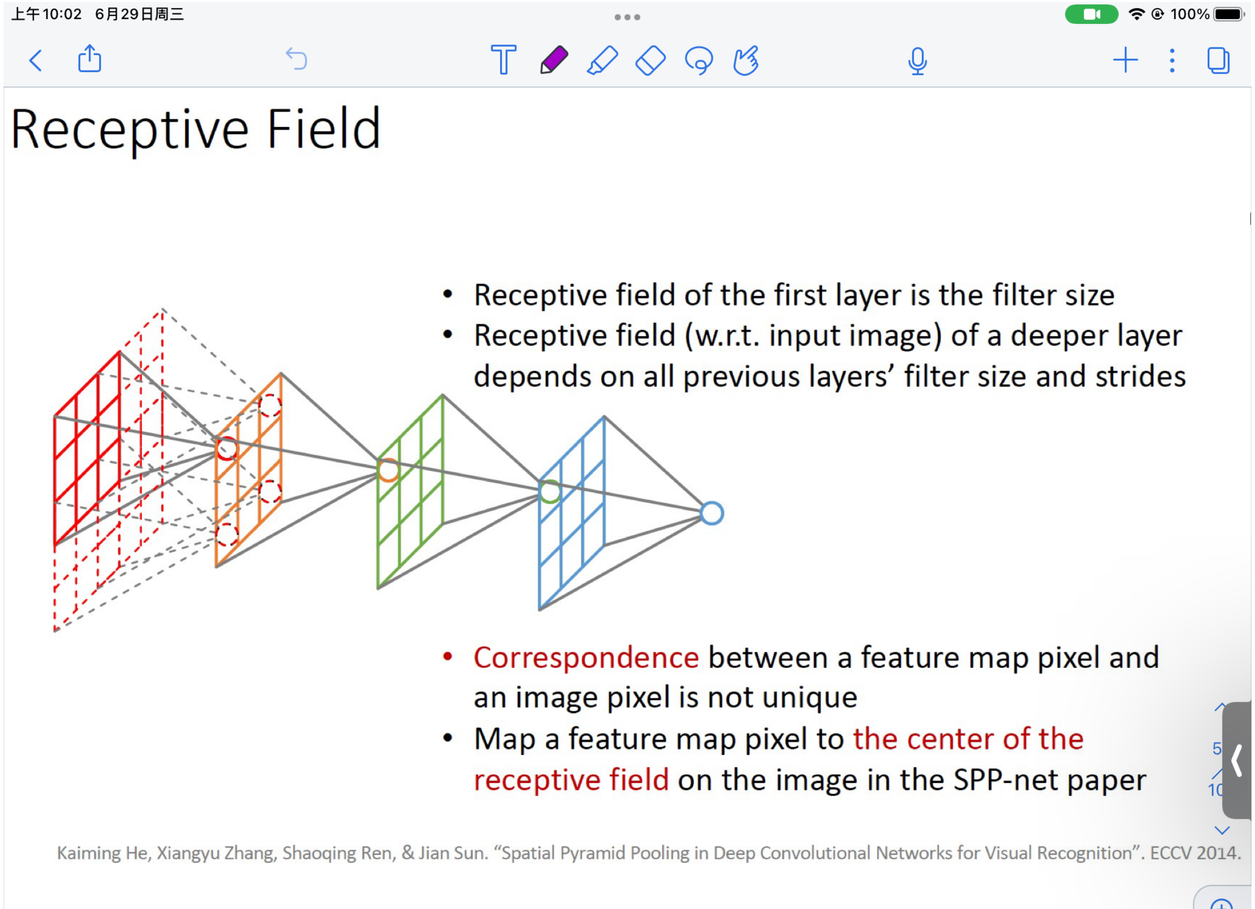
From

function $h(x)$. $f(x)$.

convo $= h(x) \otimes f(x)$

$$= \int_{-\infty}^{+\infty} h(\tau) \cdot f(t-\tau)) \, d\tau.$$





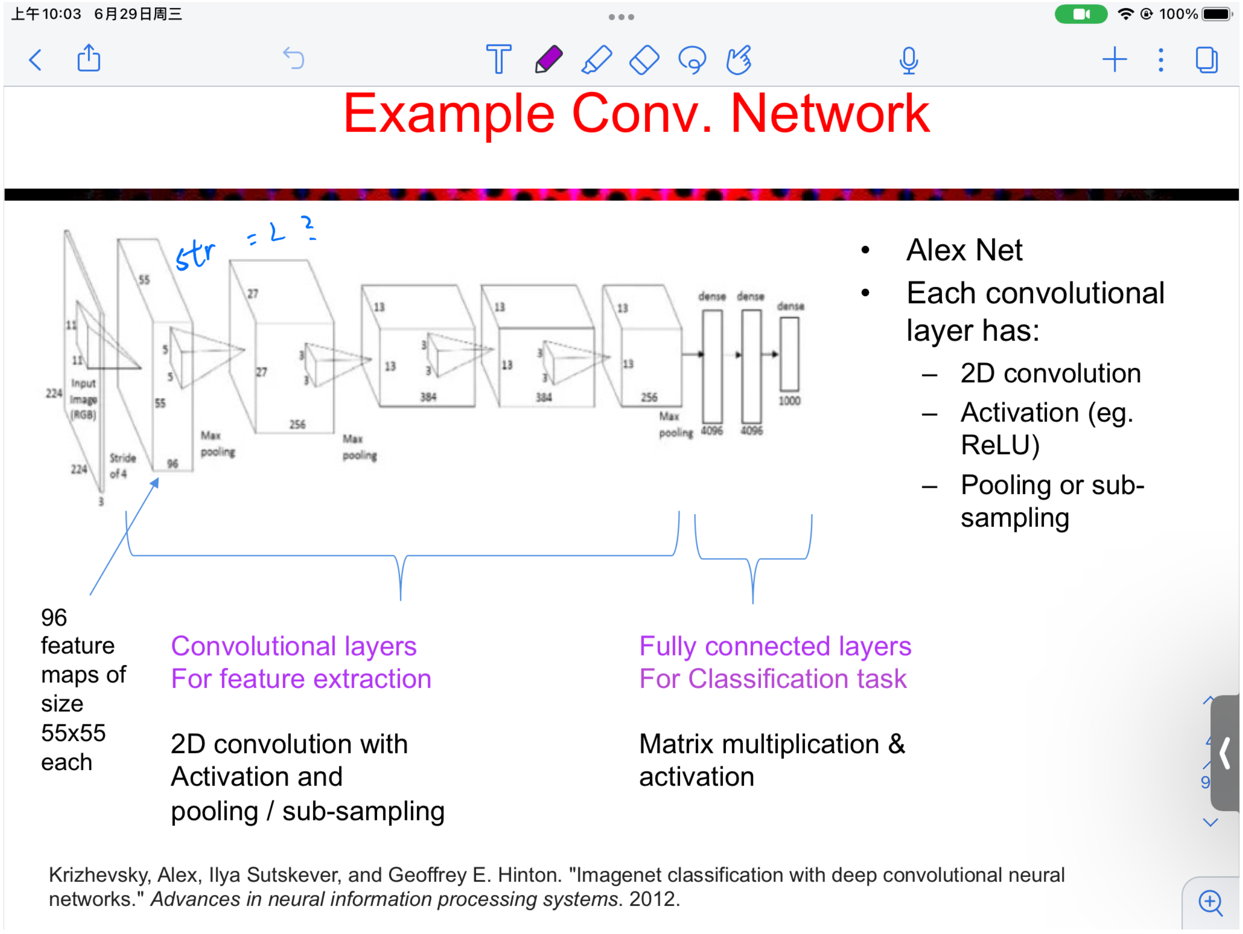
Receptive Field

- Receptive field of the first layer is the filter size
- Receptive field (w.r.t. input image) of a deeper layer depends on all previous layers' filter size and strides

- Correspondence between a feature map pixel and an image pixel is not unique
- Map a feature map pixel to the center of the receptive field on the image in the SPP-net paper

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". ECCV 2014.

From Fergus: https://cs.nyu.edu/~fergus/teaching/vision/3_convnets.pdf



Example Conv. Network

- Alex Net
- Each convolutional layer has:
  - 2D convolution
  - Activation (eg. ReLU)
  - Pooling or sub-sampling

96 feature maps of size 55x55 each

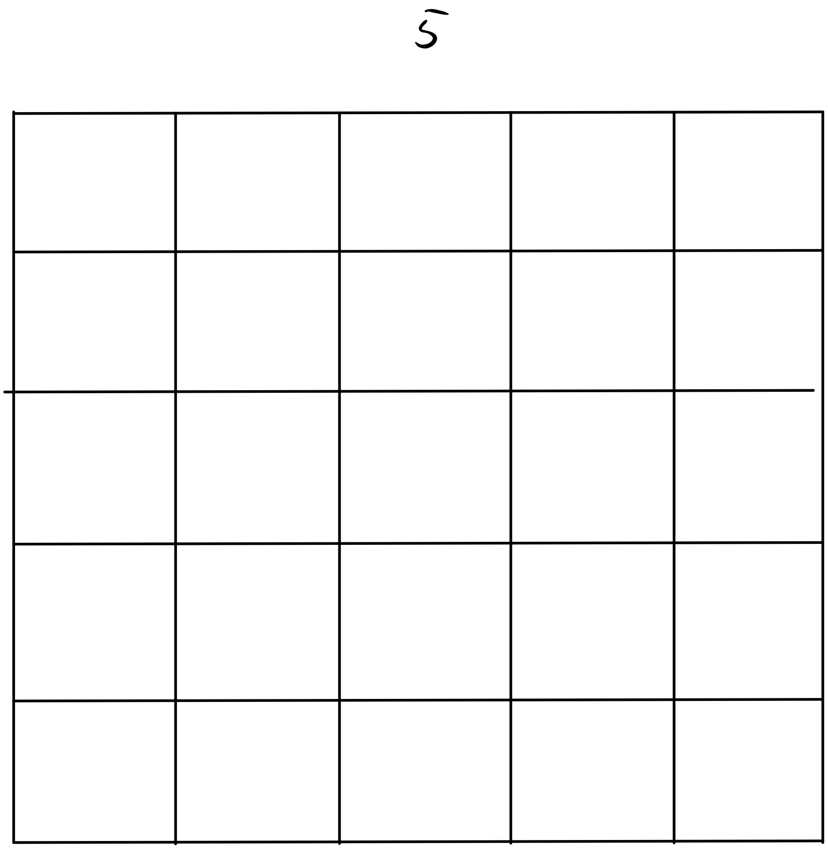
Convolutional layers
For feature extraction

2D convolution with
Activation and
pooling / sub-sampling

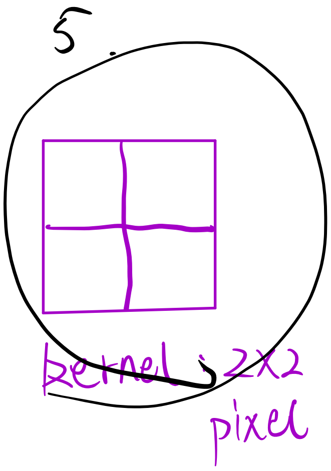
Fully connected layers
For Classification task

Matrix multiplication &
activation

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

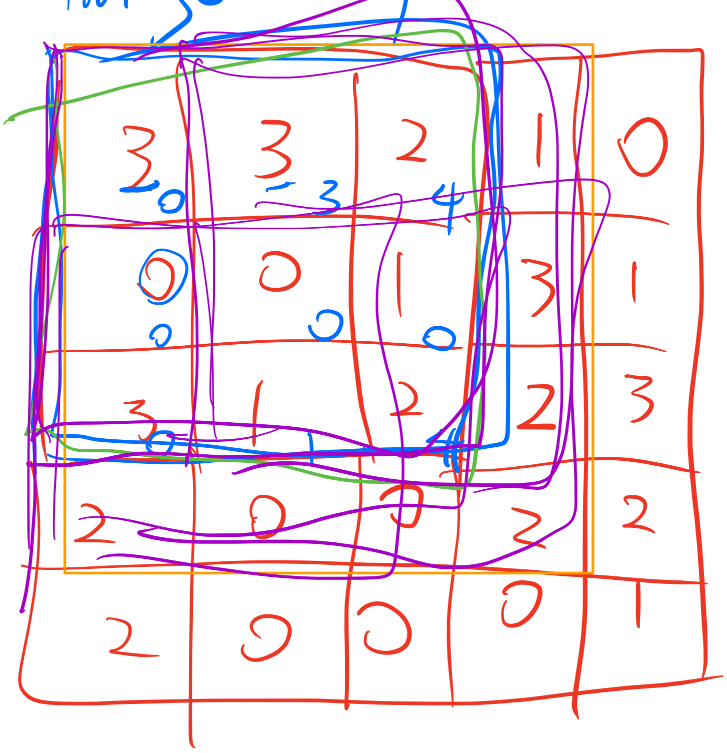


5x5 pixels



$$w = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

image = input



3_0	3_3	2_4	1	0
0_0	0_0	1_0	3	1
3_0	1_1	2_4	2	3
2	0	0	2	2
2	0	0	0	1

$$W = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

~~3+4~~

12

$$W = \begin{bmatrix} 0 & 1 & 2 & 4 \\ 2 & 2 & 0 & 1 \\ 0 & 1 & 2 & 3 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

$$\begin{bmatrix} 25 & 21 \\ 1 & 2 \end{bmatrix} = \text{output}.$$

image = input.

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$W = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}.$$

$$W = \begin{bmatrix} \quad \end{bmatrix}_{5\times5}.$$

$$\begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ \vdots \end{bmatrix} \rightarrow \begin{bmatrix} \text{Apple} \\ \text{human} \\ \vdots \end{bmatrix}$$

output of the first con

$$\underbrace{\begin{bmatrix} 12, & 12 & 17 \\ 10, & 17 & 19 \\ 9, & 6 & 14 \end{bmatrix}}_{\text{input of the second}}$$

$$W_2 = \begin{bmatrix} -1, & 0 \\ 0, & 1 \end{bmatrix}$$

$$\text{output}_2 = \begin{bmatrix} 29 & 31 \\ 16 & 31 \end{bmatrix}$$

$k = \boxed{3 \times 3}\ \boxed{5 \times 5}$

$\boxed{31 \times 31}$

$\underbrace{\boxed{1 + 2 +}}_{3}\ 3 = 6.$

$\underbrace{\boxed{1 + 2 +}}_{3}\ \boxed{4} = 7.$

$3 + 4$

$\underbrace{\boxed{1 + 2 +}}_{3}\ \boxed{100} = 103.$

$3. + 100$