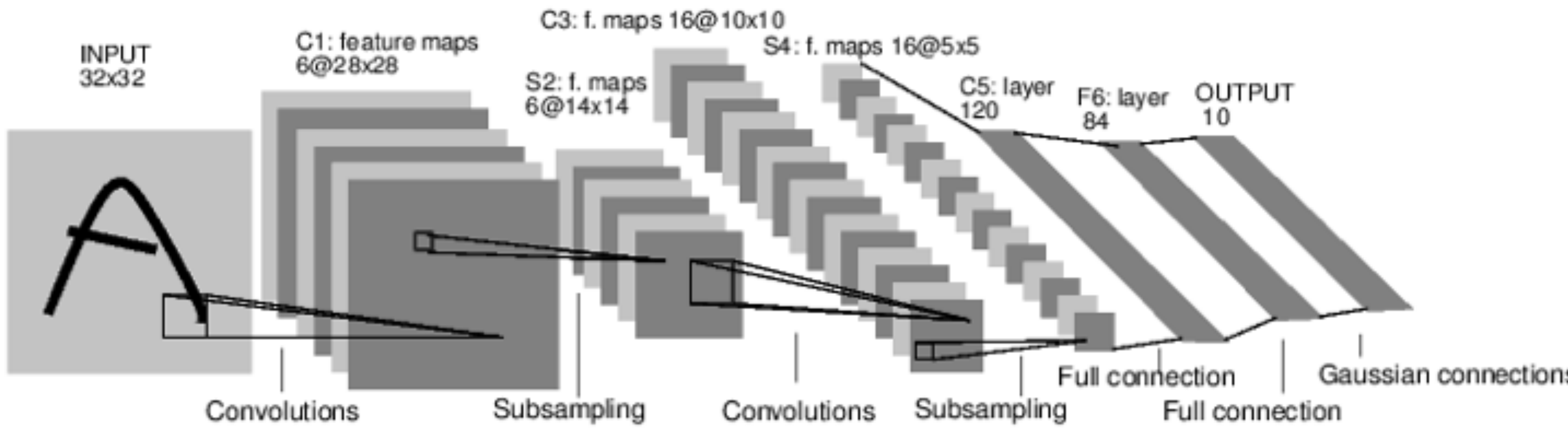
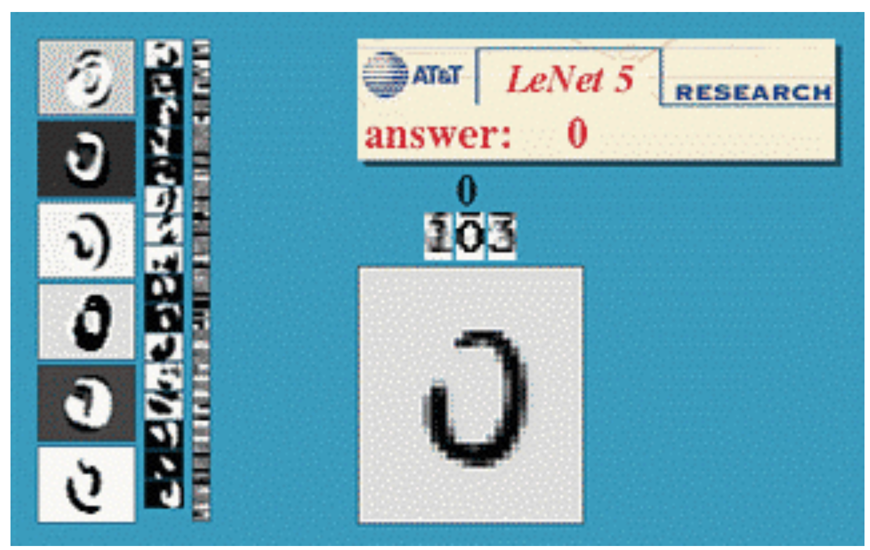
ConvNets & Multi-modal Log-bilinear Language Model

Renjie Liao

2017.Feb.14

*Some materials are credited to Jamie Kiros





LeNet5 (Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, november 1998.)

Motivation – ConvNets are everywhere!

(Krizhevsky et al, 2012)

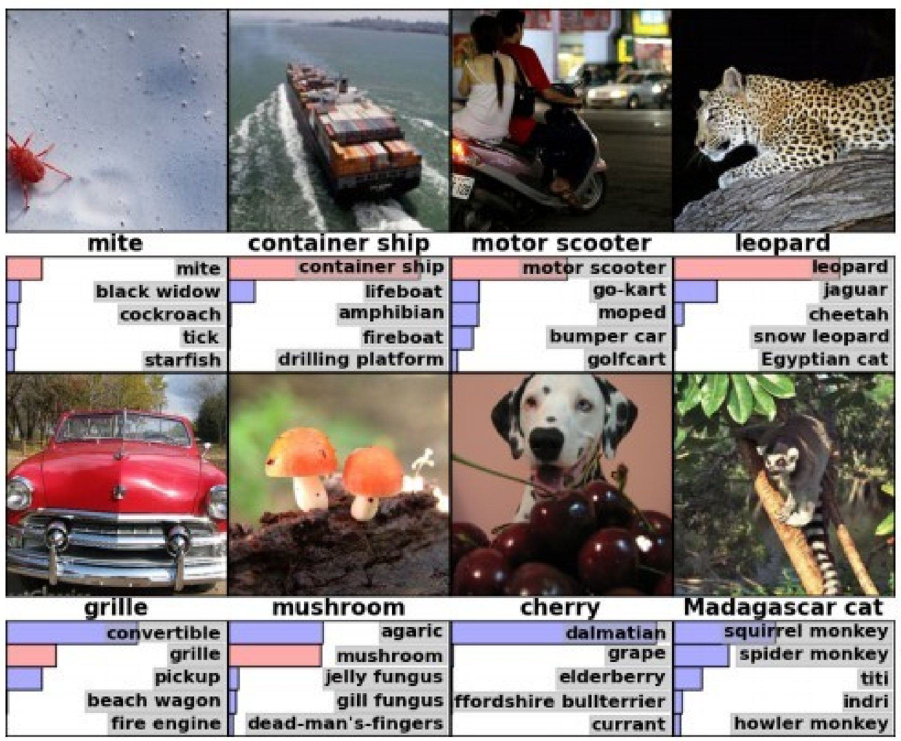


Image classification

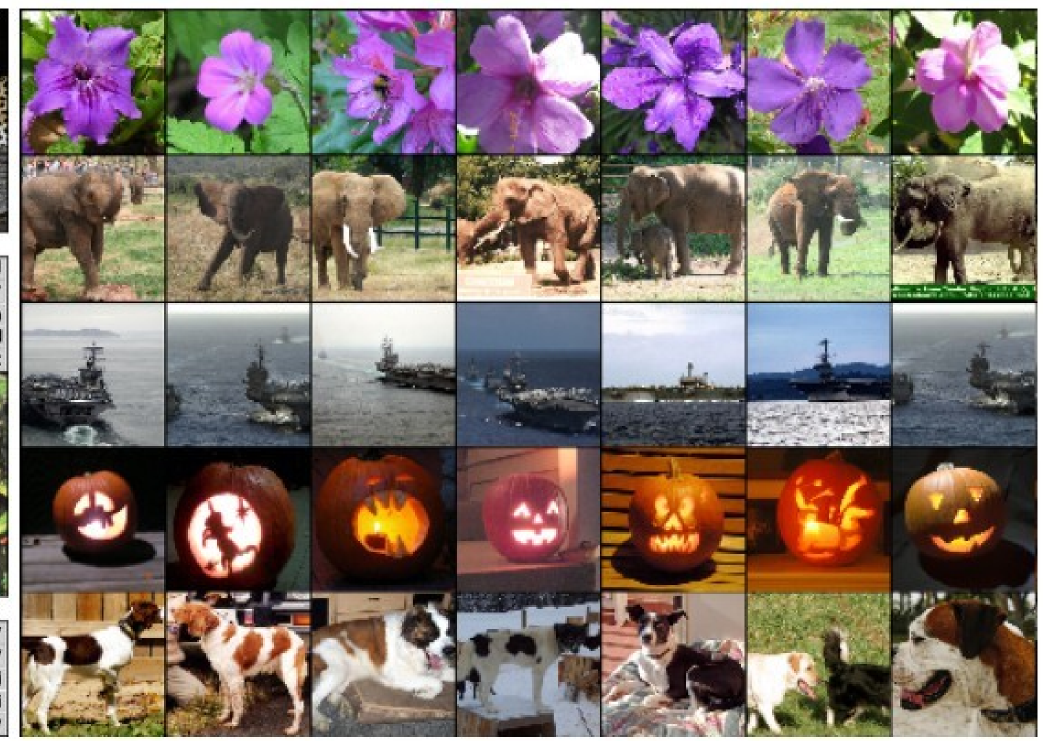
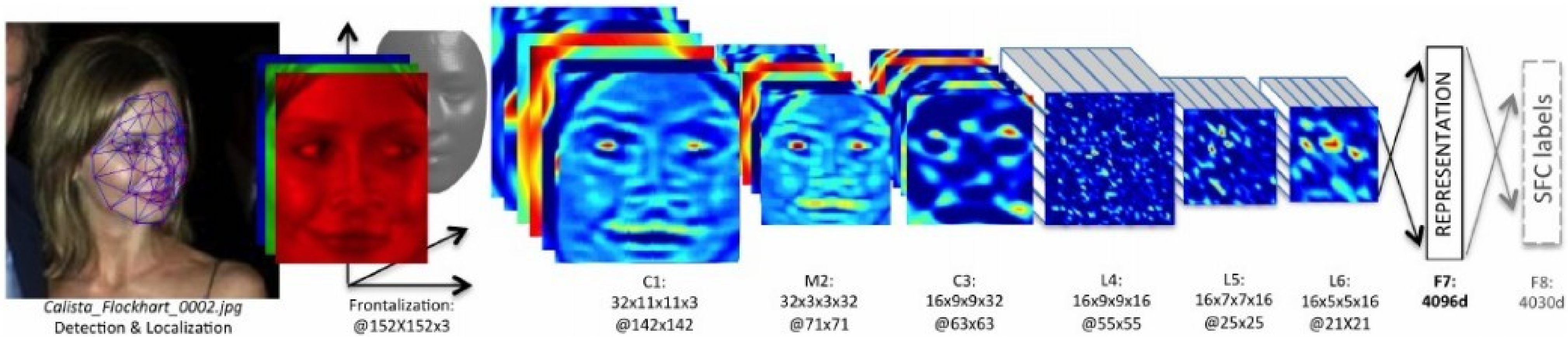
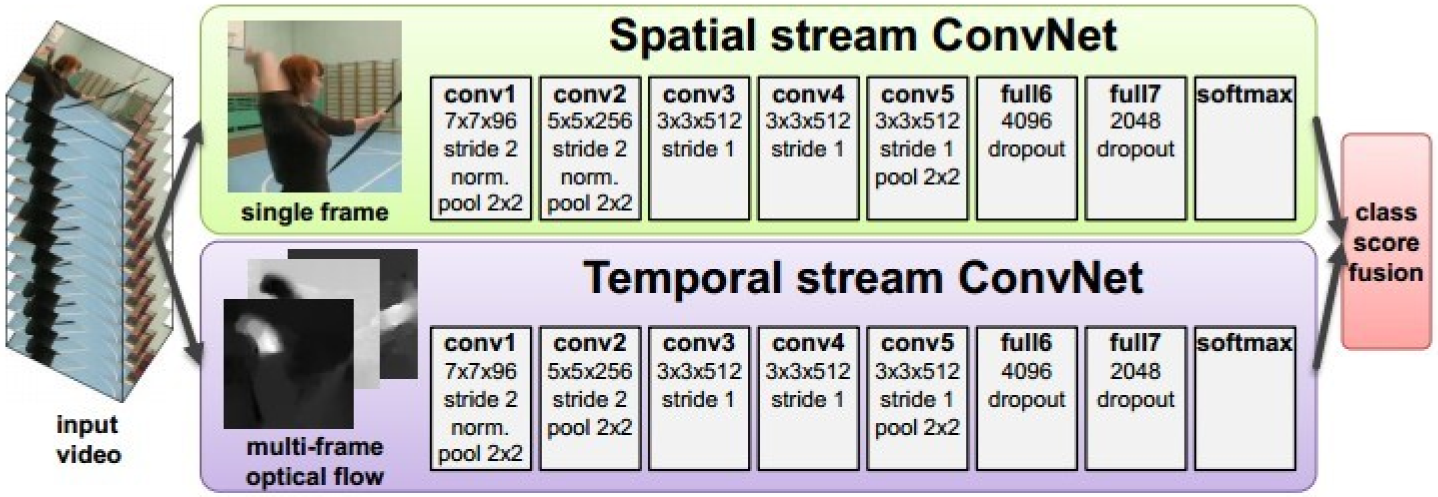


Image retrieval

Motivation – ConvNets are everywhere!



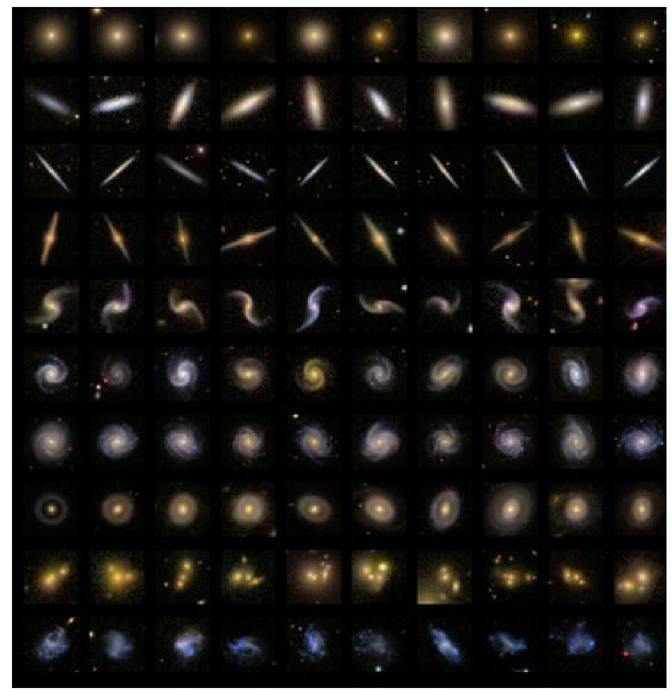
Face recognition (Taigman et al, 2014)



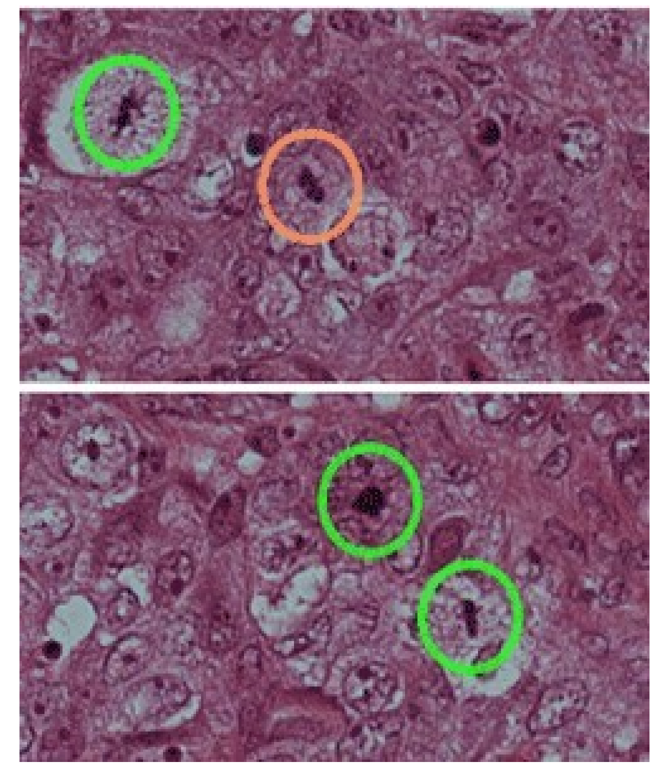
Action recognition from video (Simonyan et al, 2014)

Motivation – ConvNets are everywhere!

Street sign recognition
(Sermanet et al, 2011)



Galaxy classification
(Dieleman et al, 2014)



Mitosis detection
(Ciresan et al, 2013)

Motivation – ConvNets are everywhere!

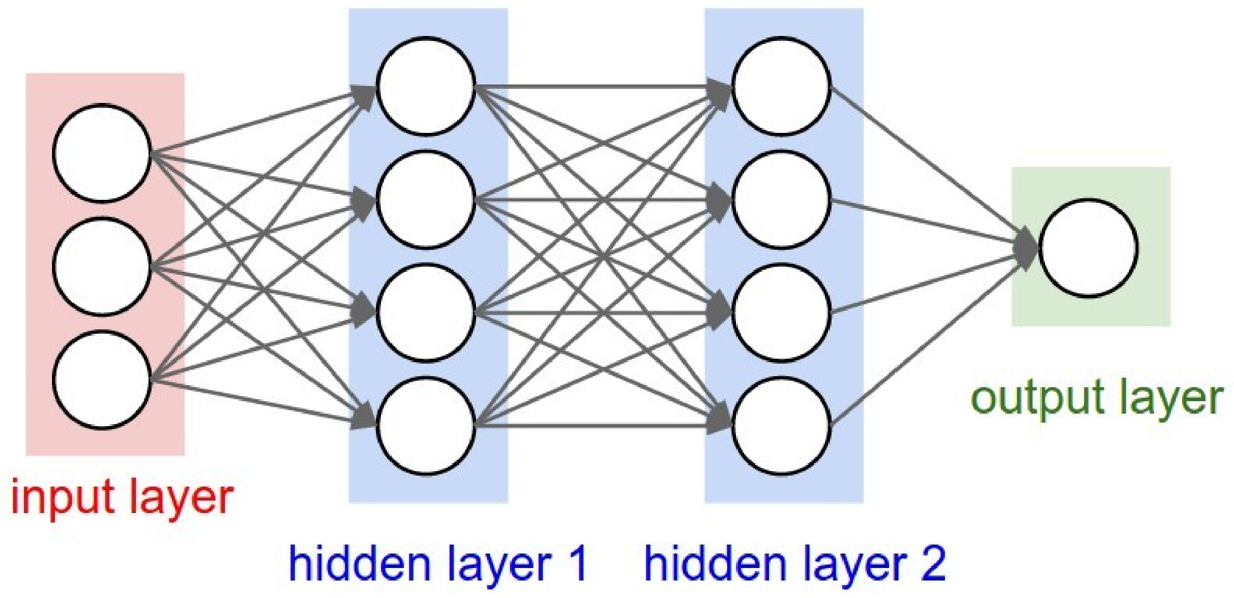
Playing Atari games (Mnih et al, 2013)

- Many, many more applications (and not only vision):
  - Object detection
  - Image segmentation
  - Pose estimation
  - Image captioning
  - Pedestrian detection
  - Semantic image search
  - Extractive summarization
  - Sentiment analysis of text

A brief review

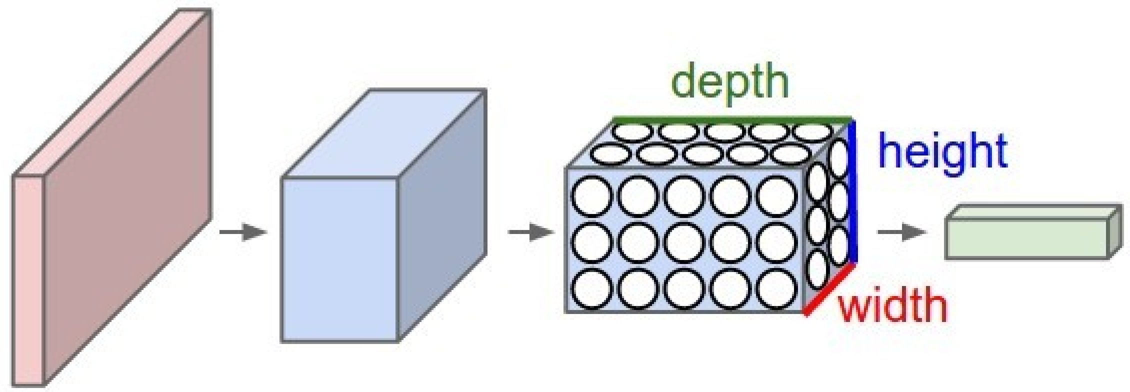
Fully connected:
(unique weights across all pairs of neurons)

Main operation:
Matrix Multiply

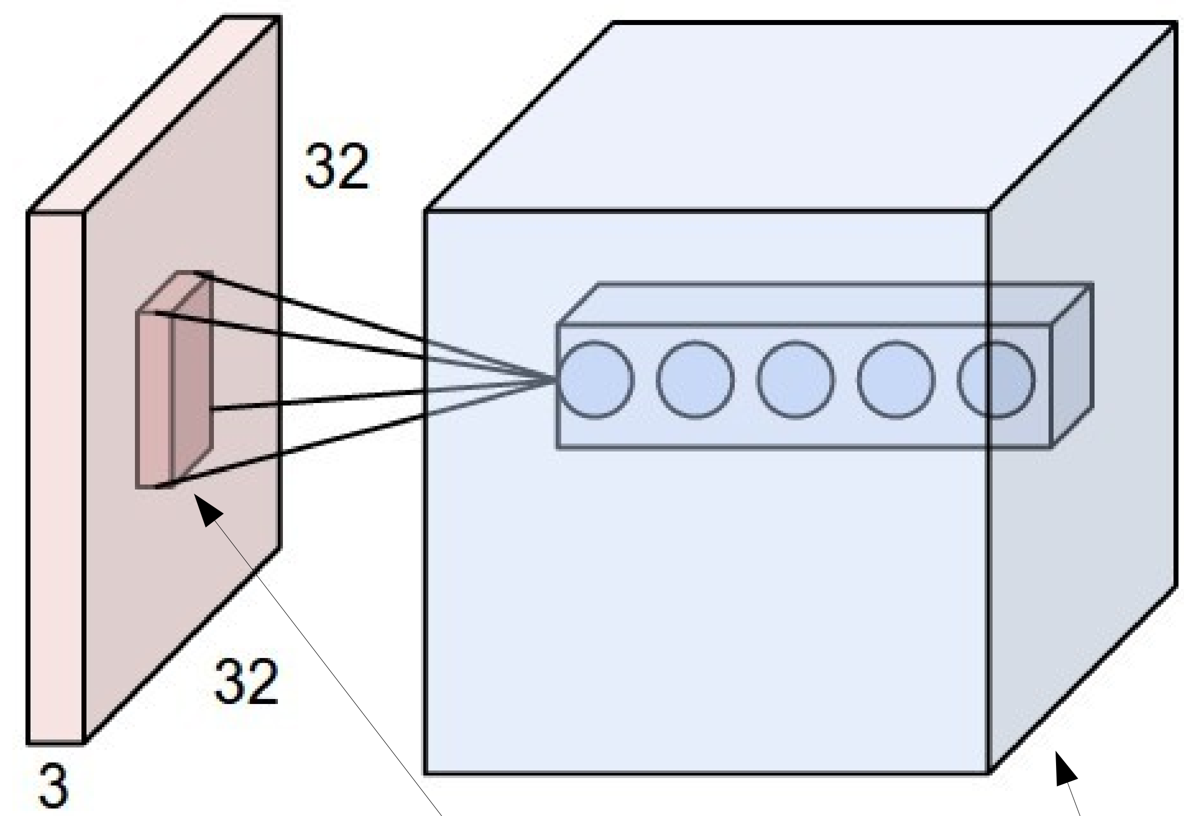


Convnet:
(neurons are volumes,
weights are shared)

Main operation:
Convolution



Some terminology



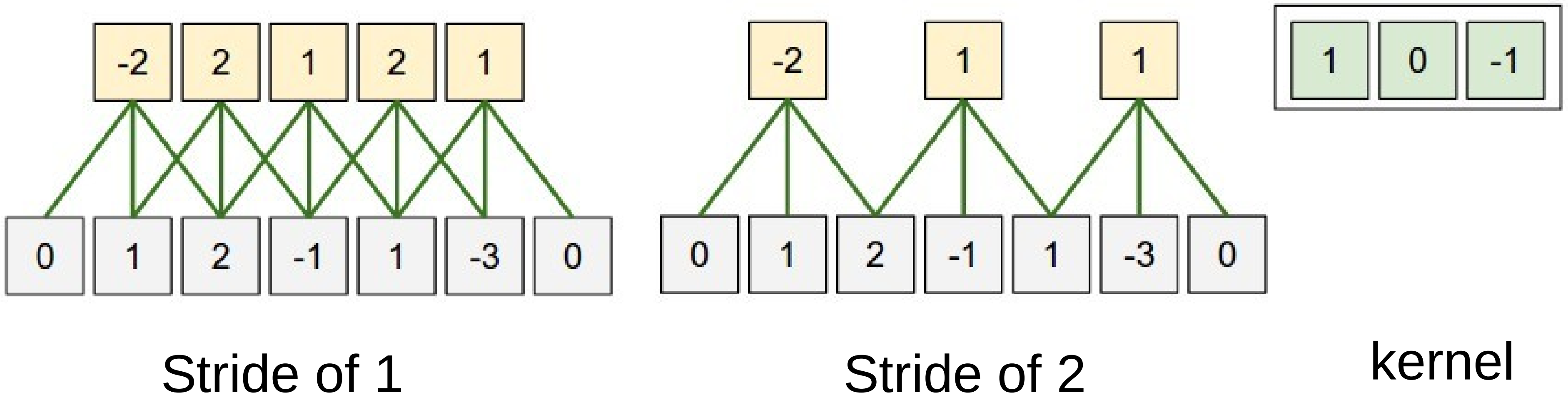
(think of this just like an image, but with 5 channels instead)

Channels
(e.g. 3 for RGB image)

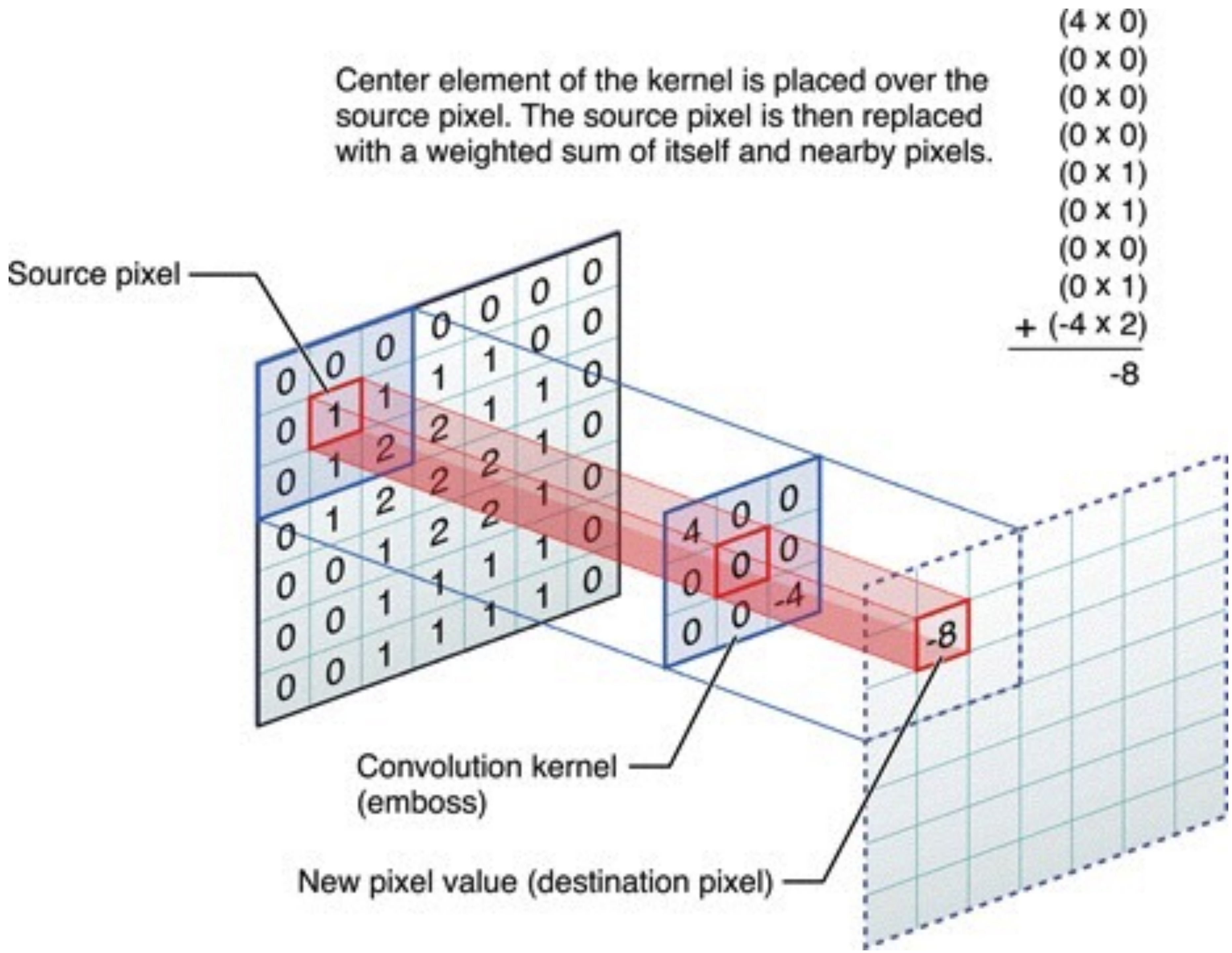
Kernel (or filter)
5 in this example

Each “slice” across depth
Is a feature map

1D forward pass, strides, padding



- Weight sharing: the kernel is scanned across the input (as opposed to fully connected networks)
- Larger strides reduce computation cost, but usually at the expense of accuracy
- In this example, each side is "padded" with an extra 0

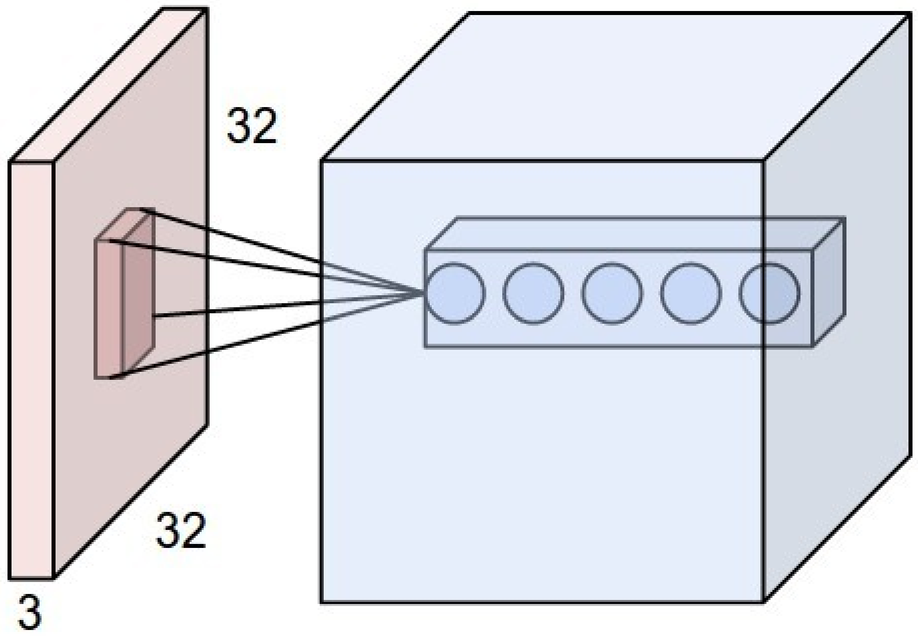

Center element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.
(4 x 0)
(0 x 0)
(0 x 0)
(0 x 0)
(0 x 1)
(0 x 1)
(0 x 0)
(0 x 1)
+ (-4 x 2)
-8
Source pixel
Convolution kernel (emboss)
New pixel value (destination pixel)

2D Convolution Example

$$
* \quad \begin{array}{|c|c|c|} \hline 1 & 0 & -1 \\ \hline 2 & 0 & -2 \\ \hline 1 & 0 & -1 \\ \hline \end{array}
$$

Example #1

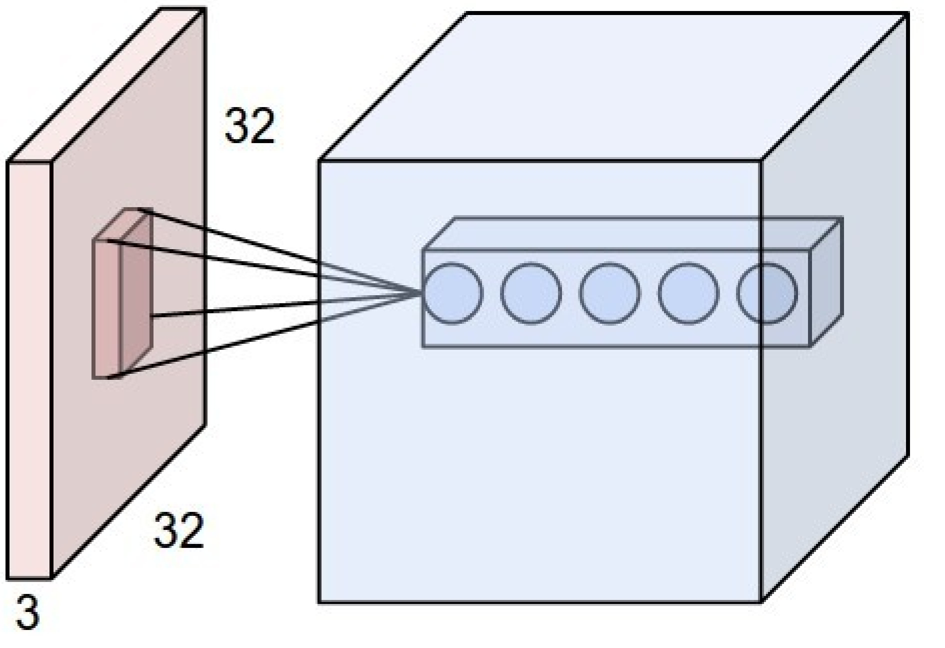
- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 1
- No padding



- What is the output volume?
- How many parameters are there?

Example #1

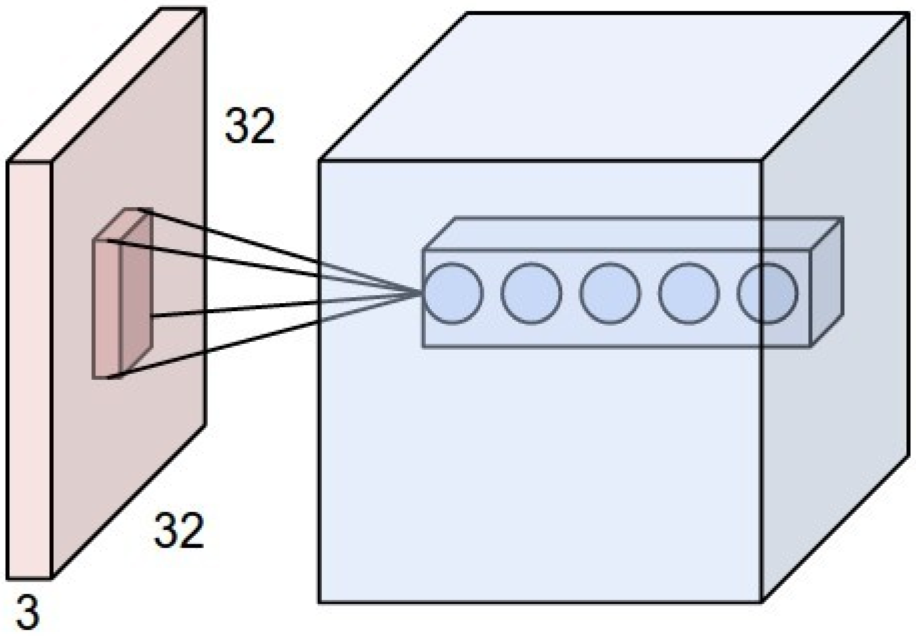
- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 1
- No padding



- What is the output volume? 28 x 28 x 5
- How many parameters are there? ((5 x 5) x 3) x 5 = 375

Example #2

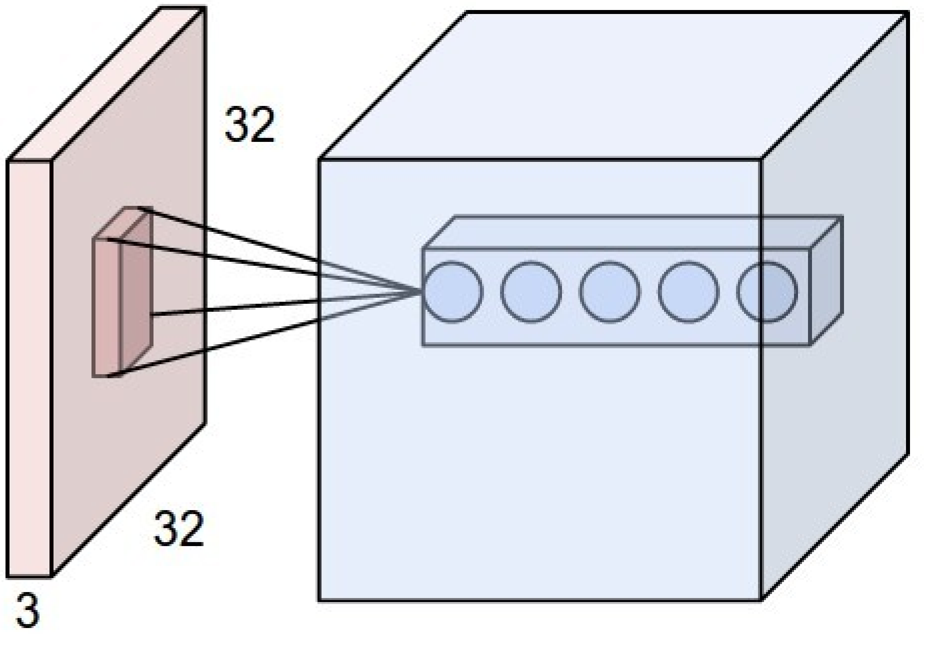
- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 3
- No padding



- What is the output volume?
- How many parameters are there?

Example #2

- Input: 32 x 32 x 3 image
- 5 Filters, each 5 x 5
- Stride of 3
- No padding



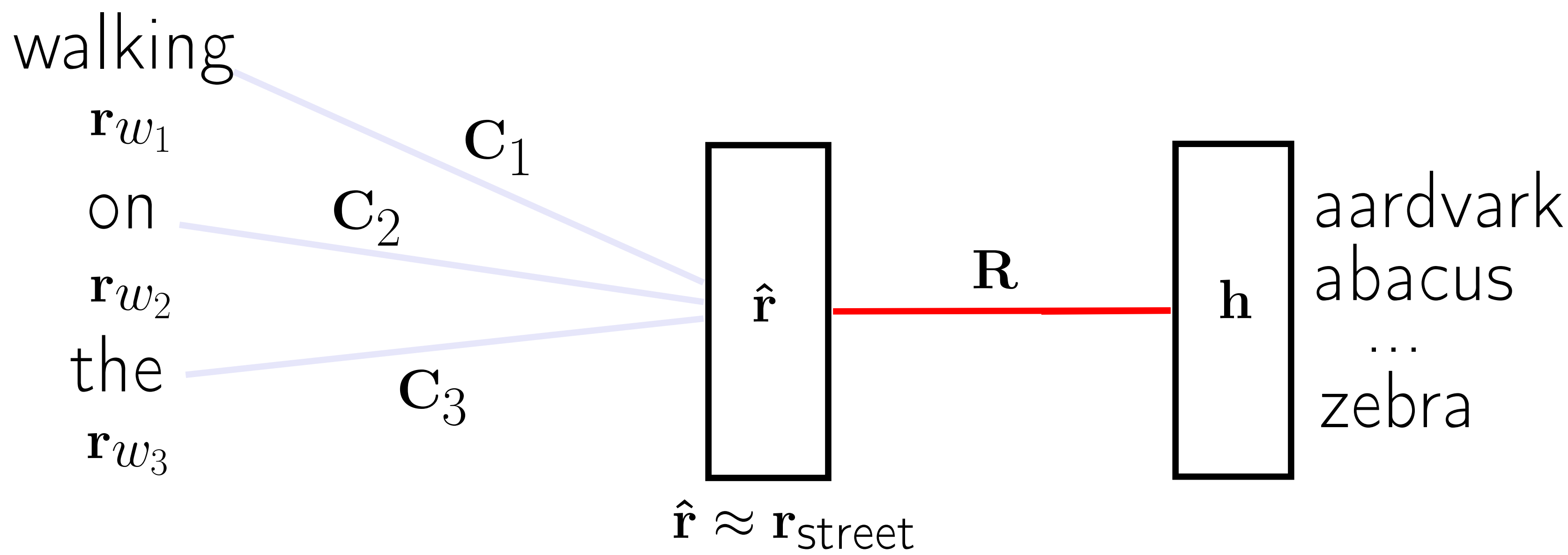
- What is the output volume? 10 x 10 x 5
- How many parameters are there? ((5 x 5) x 3) x 5 = 375

Generating text conditioned on images

in this picture there is another grey pavement on the right ; three grey clouds and a blue sky in the background ; the houses and on the left before it ; a dark green , wooded slopes behind it ; grey clouds in a light blue sky in the background ; snow covered mountains

this product contains a slip resistant and mesh upper is fully designed for breathable durability . the detachable leather footbed is the high , they feature a lady - like footbed that light sophistication and flirty tear silhouette to glam up your feet , style to help your thing . with traditional support .

The Log-Bilinear Language Model (LBL)

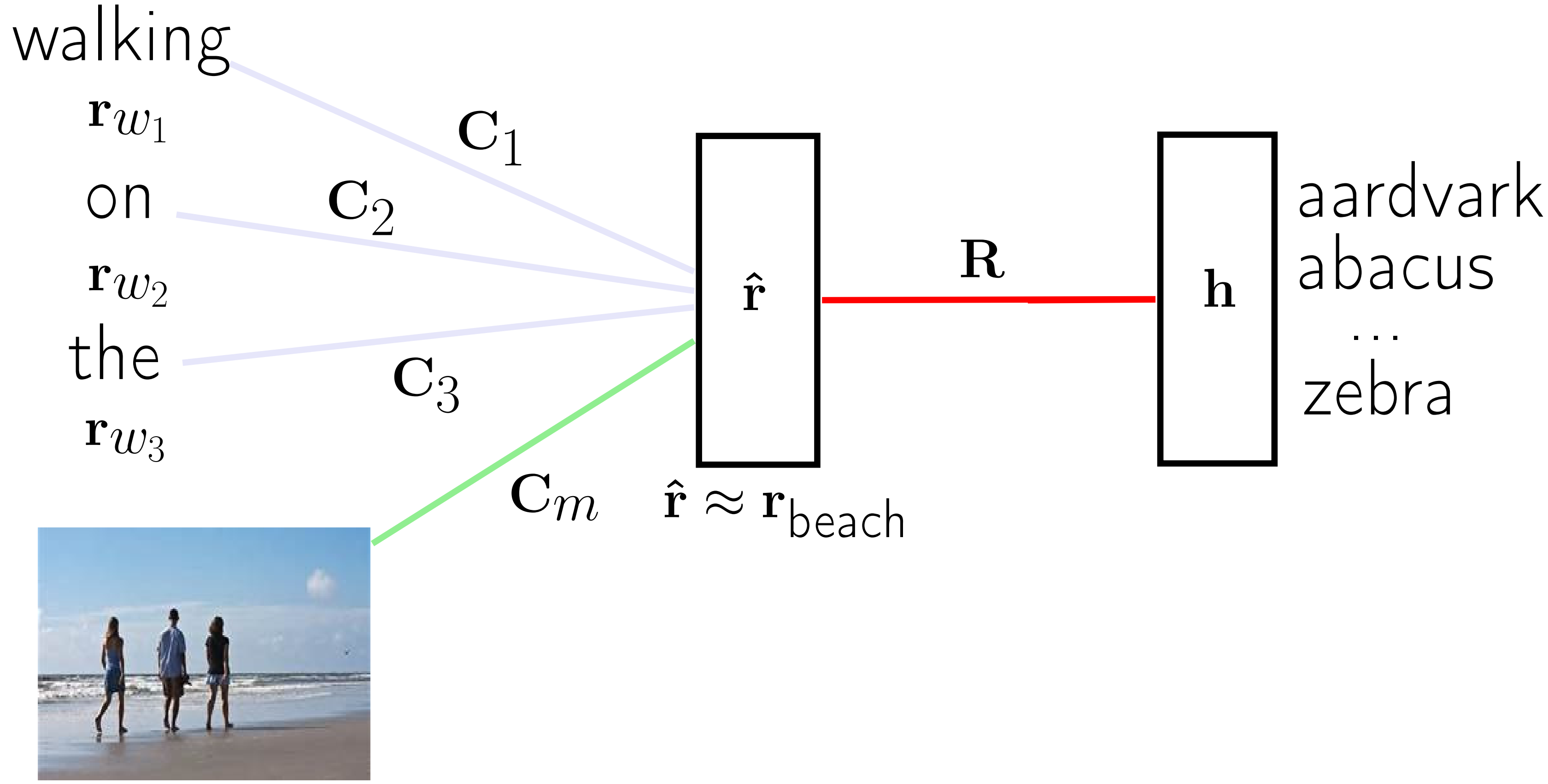


- Word representations $\mathbf{r}_{w_i}$, context matrices C_i
- Predicted next word representation $\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}_i \mathbf{r}_{w_i}$
- $\mathbf{R}$: matrix where each row is a word feature from the vocabulary
- Score $\hat{\mathbf{r}}$ with each word and normalize:

$$P(w_n = w | w_{1:n-1}) = \frac{\exp(\hat{\mathbf{r}}^T \mathbf{r}_w + b_w)}{\sum_j \exp(\hat{\mathbf{r}}^T \mathbf{r}_j + b_j)}$$

- Backprop through both parameters and word embeddings

ADDITIVE MODALITY BIASING (MLBL-B)



- Suppose we have image features $\mathbf{x}$
- Simplest approach: Bias the predicted next word representation:

$$\hat{\mathbf{r}} = \left(\sum_{i=1}^{n-1} \mathbf{C}_i \mathbf{r}_{w_i} \right) + \mathbf{C}_m \mathbf{x}$$

- This turns out to be a surprisingly effective model (given good image features)