Classification of ncRNAs using machine learning methods

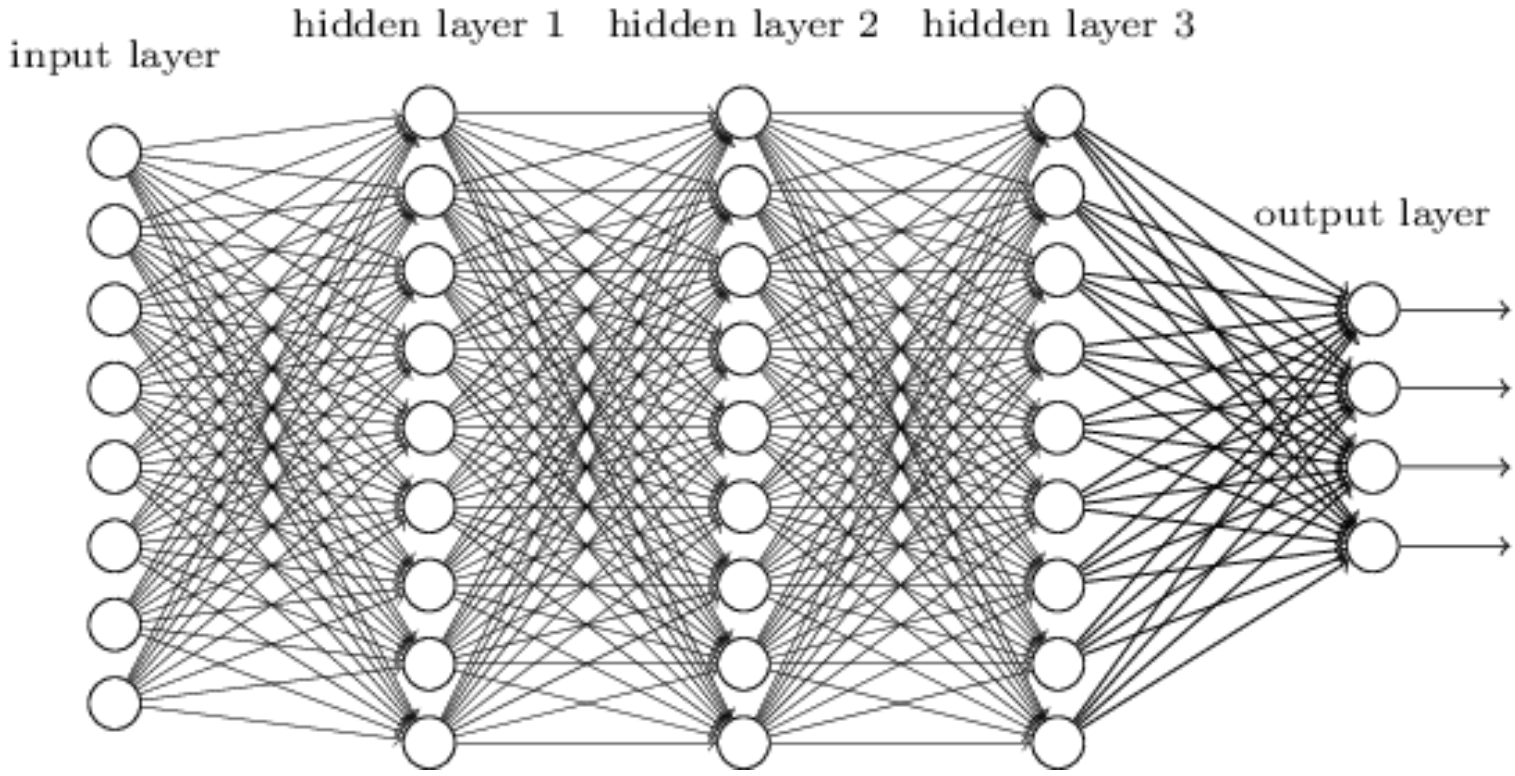
Stefan Simm

31. March 2023

CONVOLUTIONAL NEURAL NETWORK

ANN or CNN?

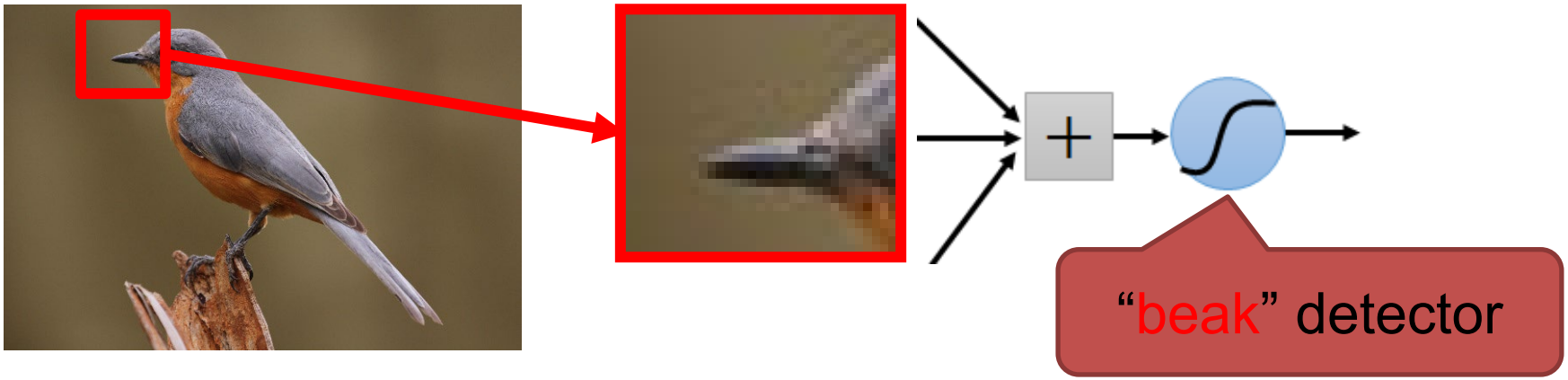
- We know it is better to train on small models if possible.
- From this fully connected model, do we really need all the edges?
- Can some of these be shared?



Consider learning from images:

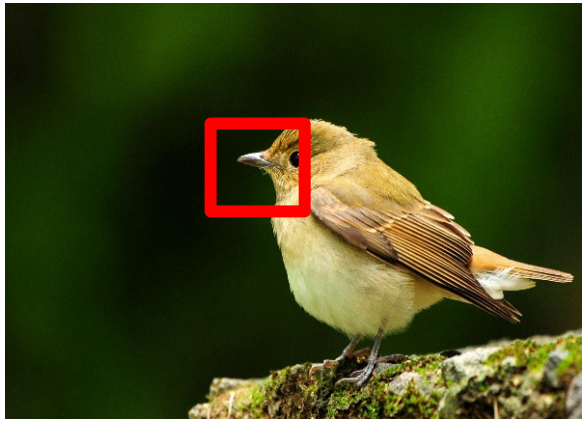
- Some patterns are much smaller than the whole image

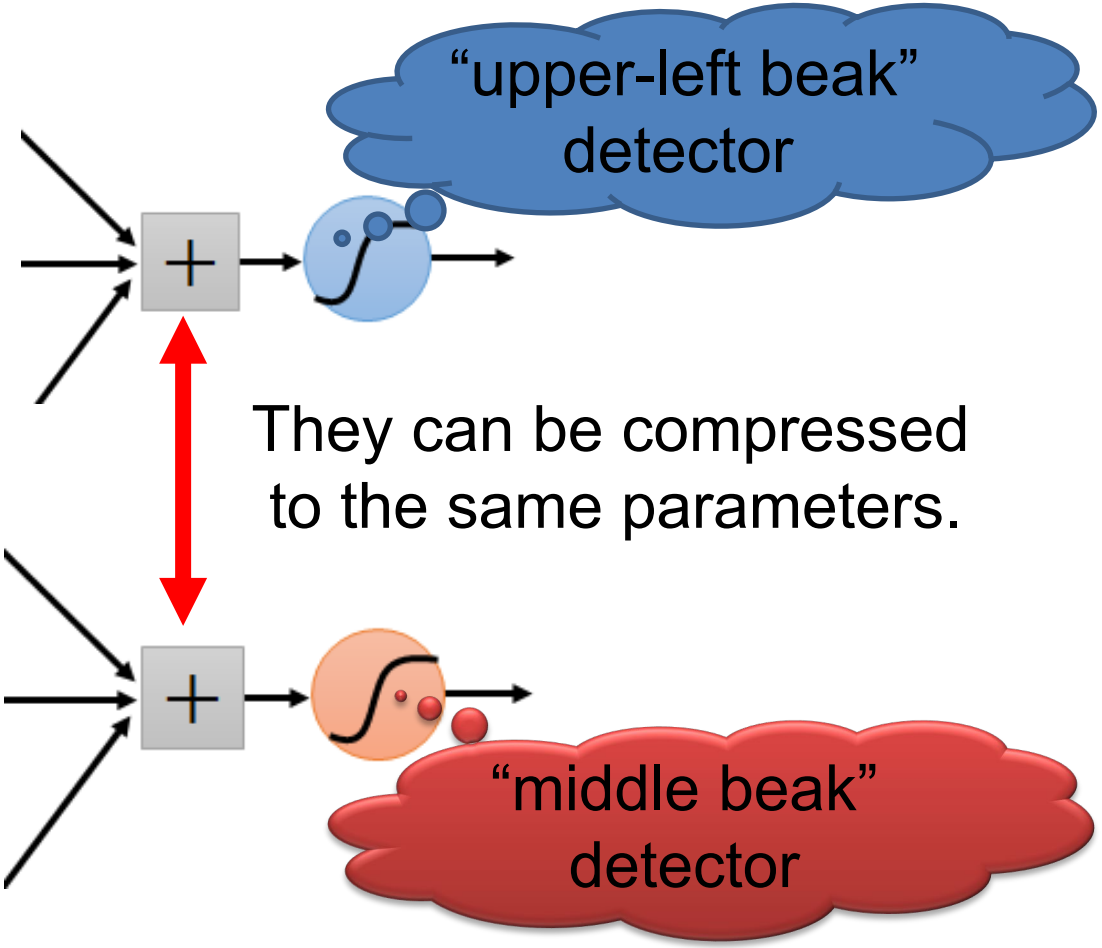
Can represent a small region with fewer parameters



Same pattern appears in different places: They can be compressed!

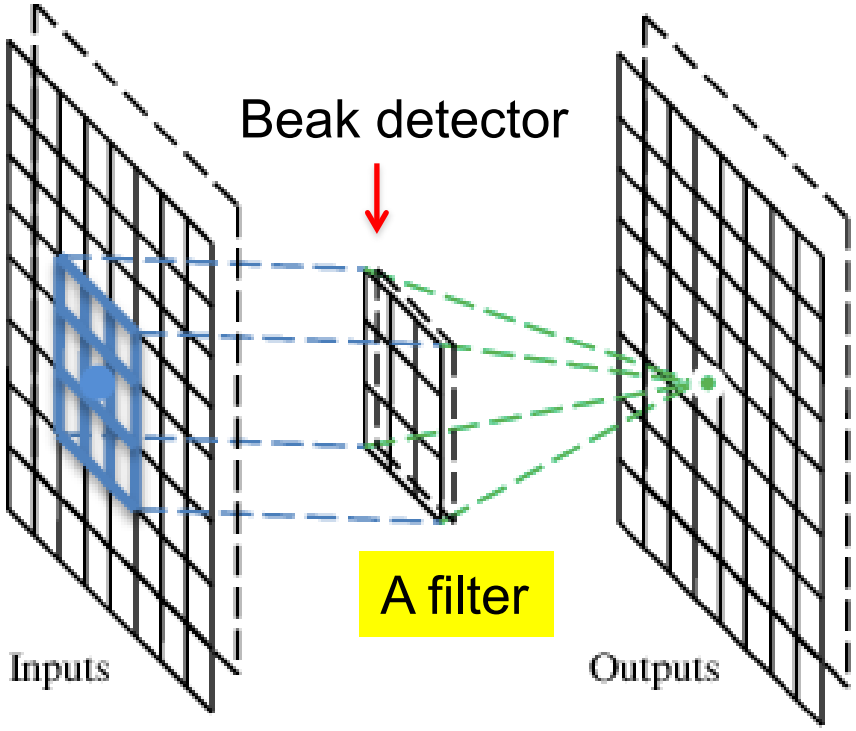
What about training a lot of such "small" detectors
and each detector must "move around".





A convolutional layer

A CNN is a neural network with some convolutional layers (and some other layers). A convolutional layer has a number of filters that does convolutional operation.



Convolution

These are the network parameters to be learned.

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

-1	1	-1
-1	1	-1
-1	1	-1

Filter 2

⋮ ⋮

Each filter detects a small pattern (3 x 3).

Convolution

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

stride=1

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

Dot product

3 -1

6 x 6 image

.Convolution

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

If stride=2

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

3 -3

Convolution

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

stride=1

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

3	-1	-3	-1
-3	1	0	-3
-3	-3	0	1
3	-2	-2	-1

Convolution

-1	1	-1
-1	1	-1
-1	1	-1

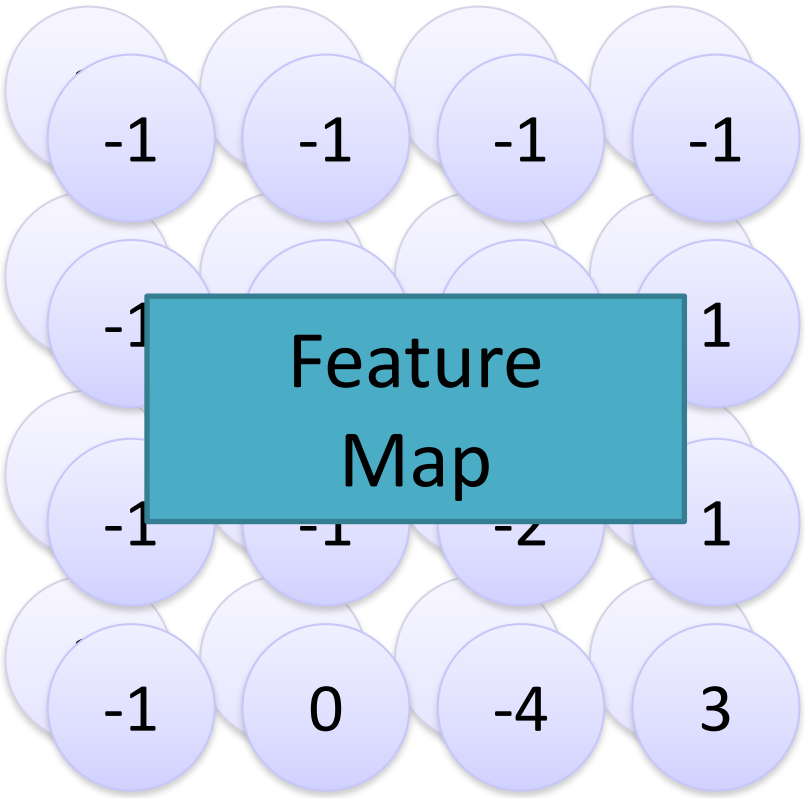
Filter 2

stride=1

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

Repeat this for each filter



Two 4 x 4 images
Forming 2 x 4 x 4 matrix

Color image: RGB 3 channels

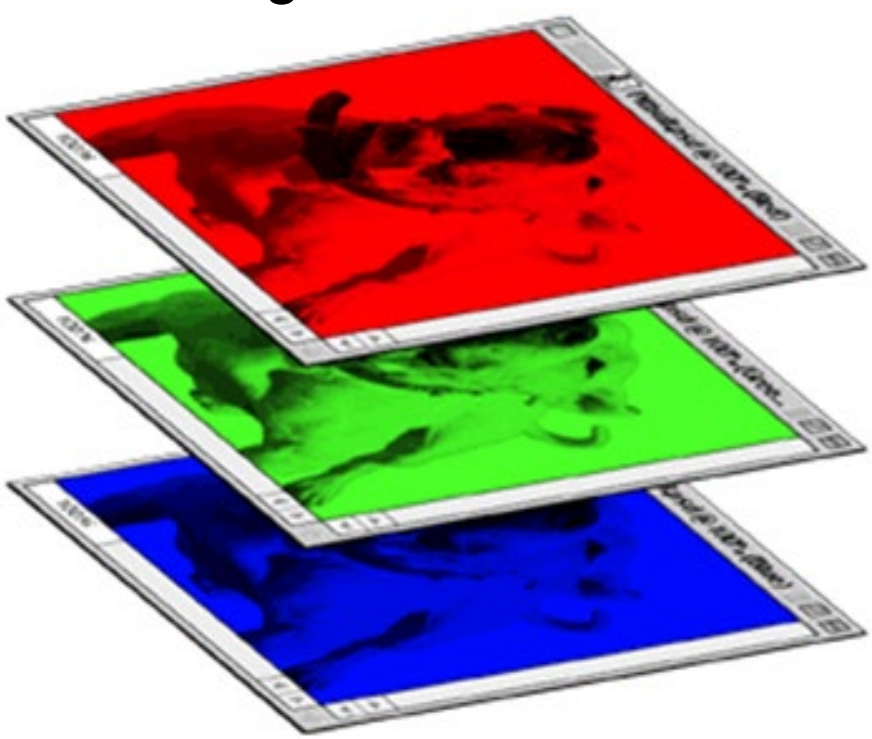
1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

-1	1	-1
-1	1	-1
-1	1	-1

Filter 2

Color image



1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

Convolution vs. Fully Connected

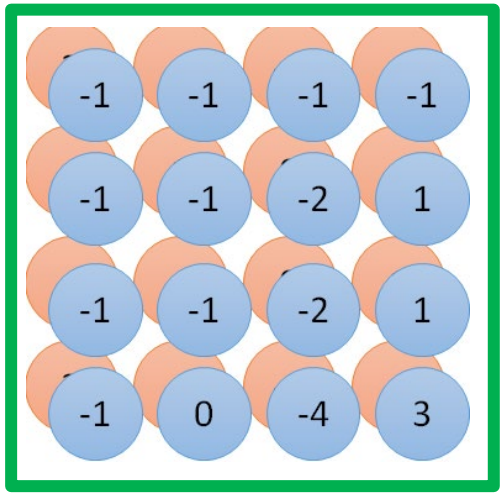
1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

image

1	-1	-1
-1	1	-1
-1	-1	1

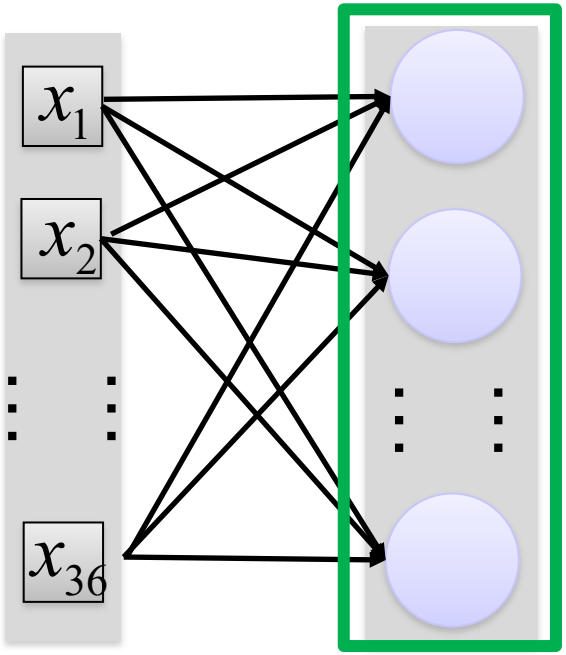
-1	1	-1
-1	1	-1
-1	1	-1

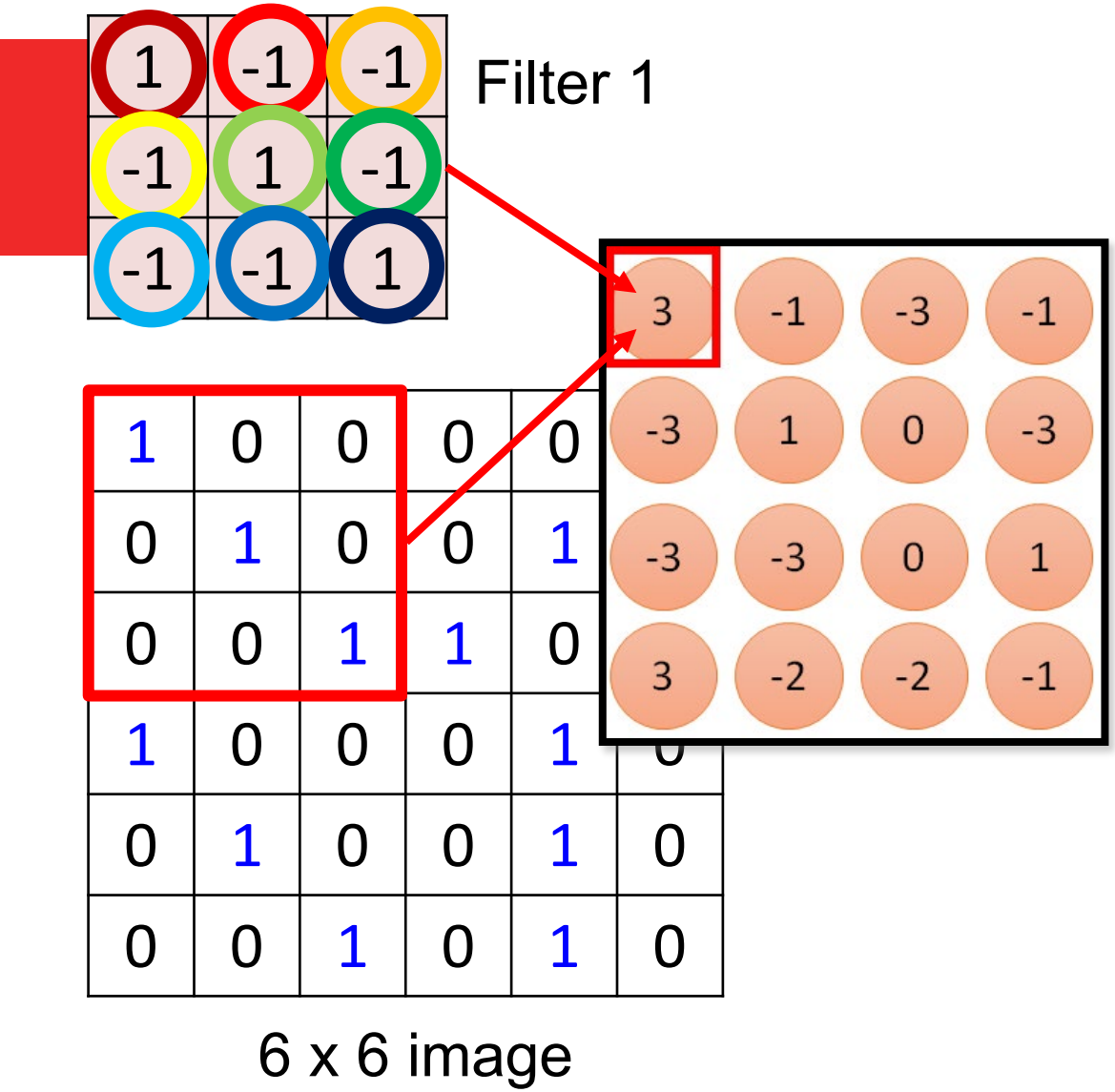
convolution



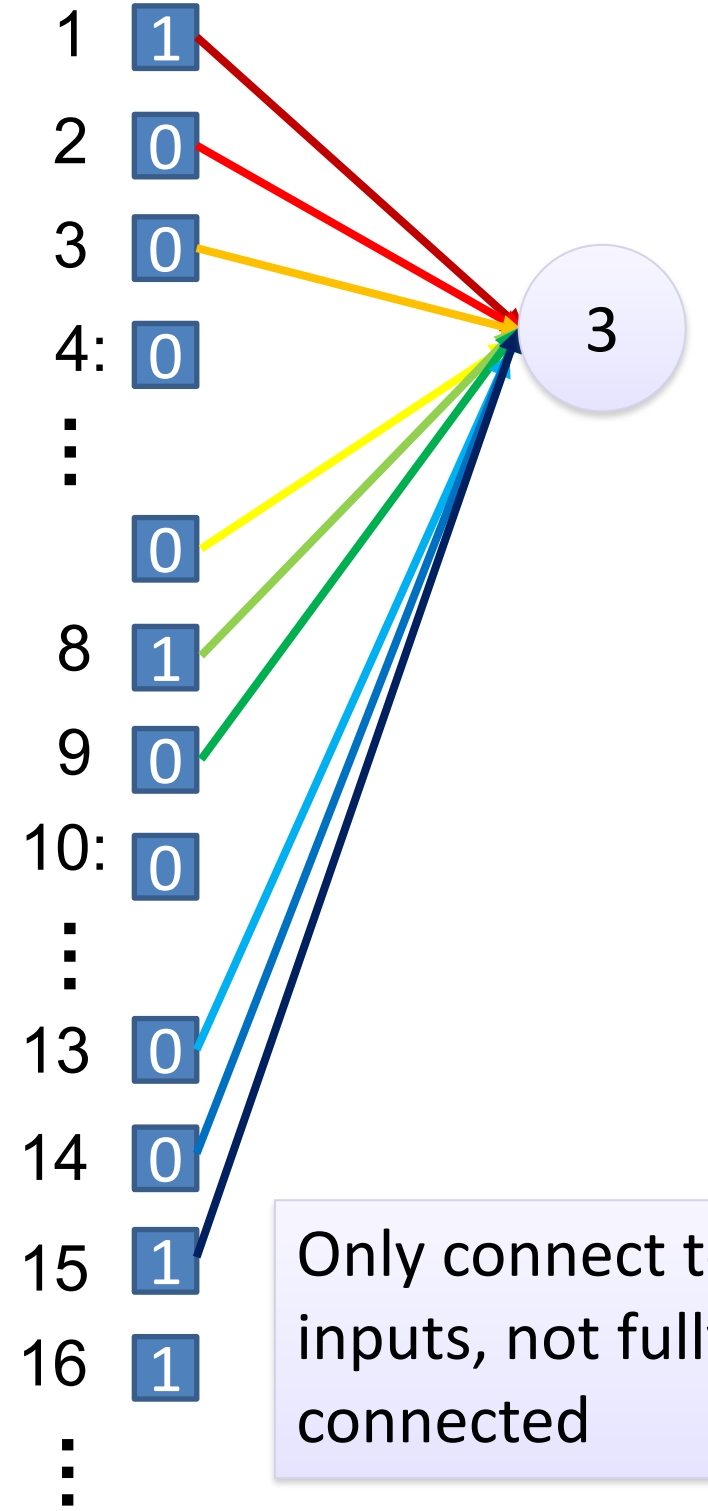
Fully-connected

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

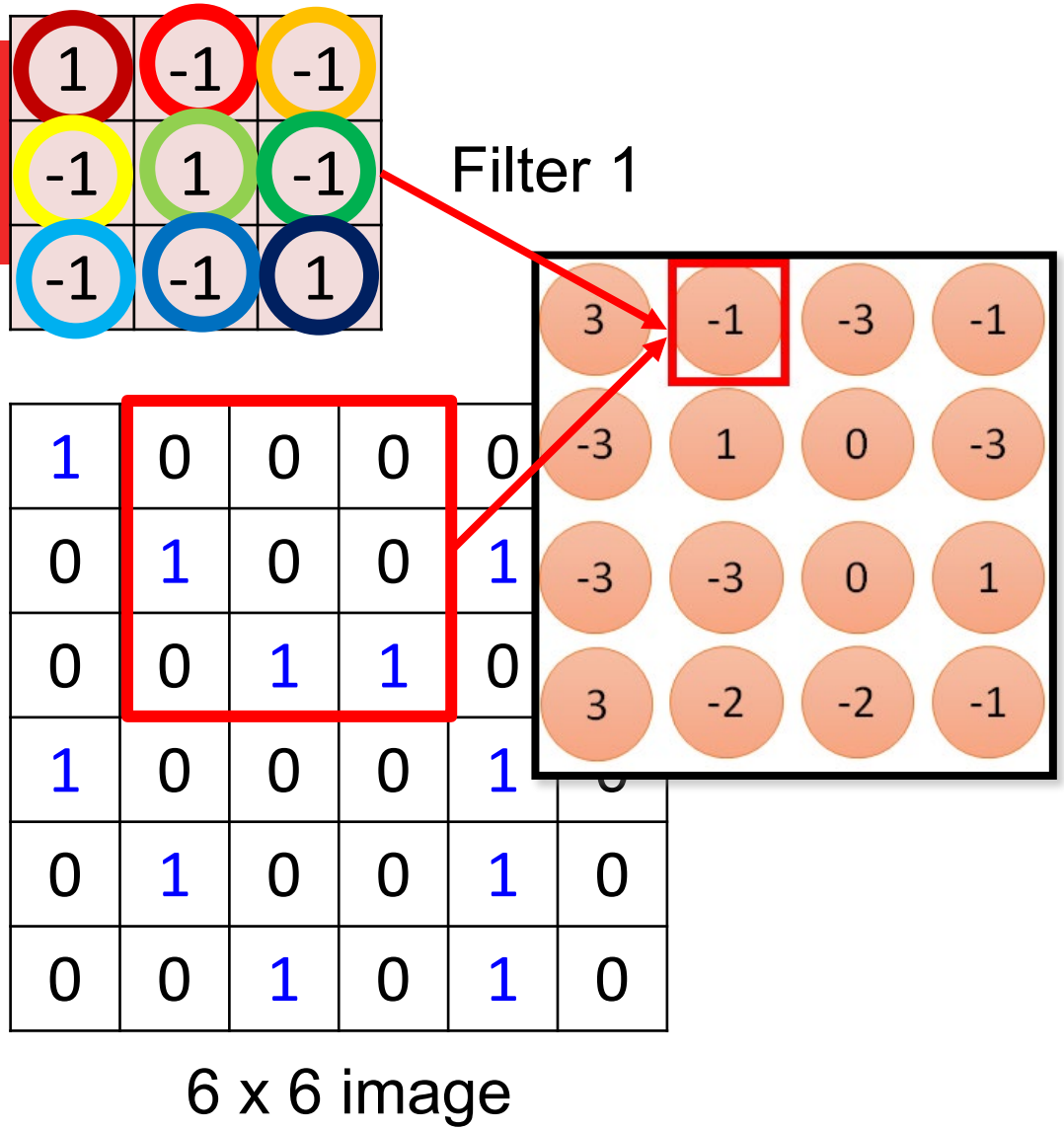


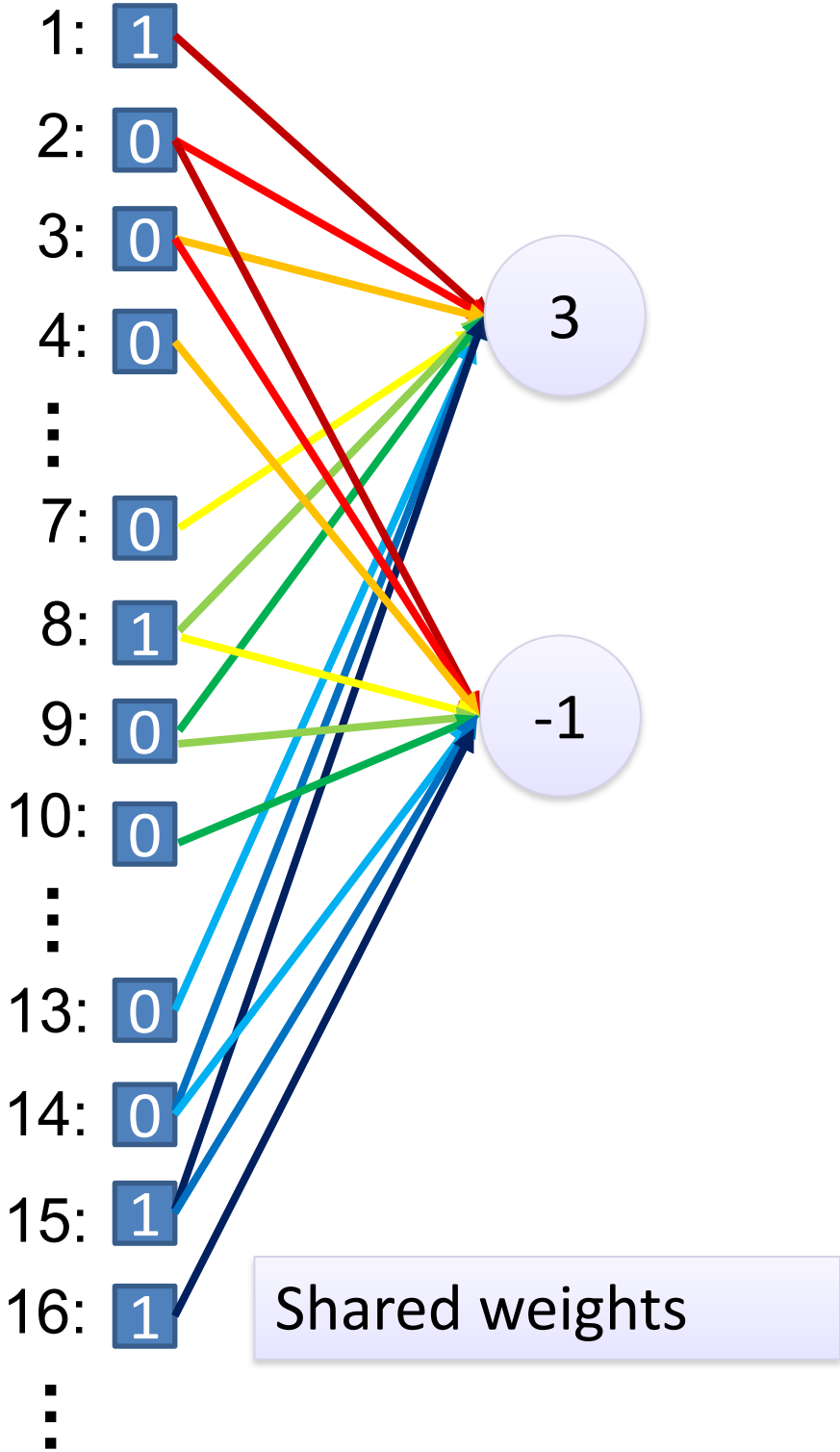


fewer parameters!



Only connect to 9 inputs, not fully connected

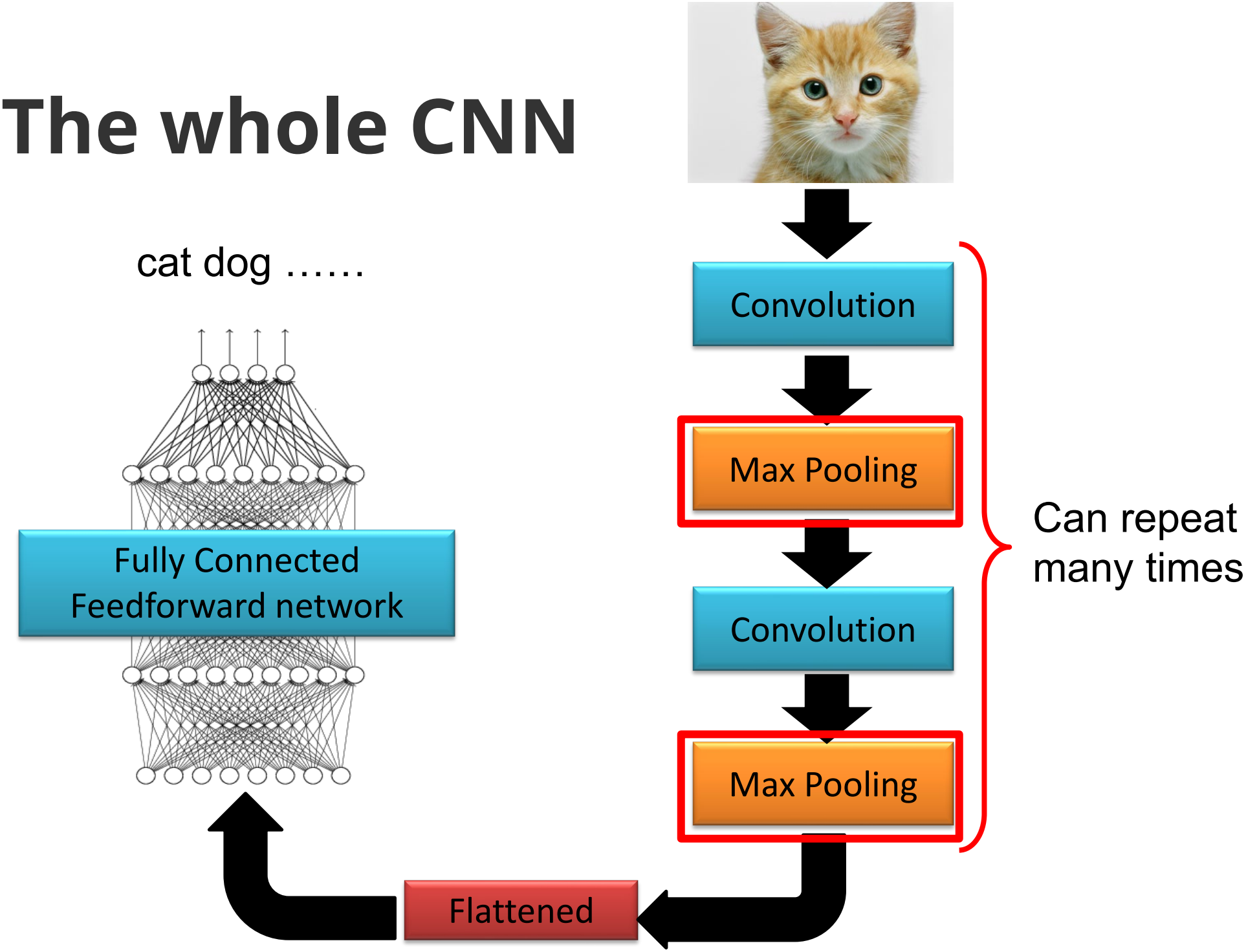




Fewer parameters

Even fewer parameters

The whole CNN



Max Pooling

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

-1	1	-1
-1	1	-1
-1	1	-1

Filter 2

3	-1
-3	1

-3	-1
0	-3

-3	-3
3	-2

0	1
-2	-1

-1	-1
-1	-1

-1	-1
-2	1

-1	-1
-1	0

-2	1
-4	3

Why Pooling

- Subsampling pixels will not change the object

bird

Subsampling

bird

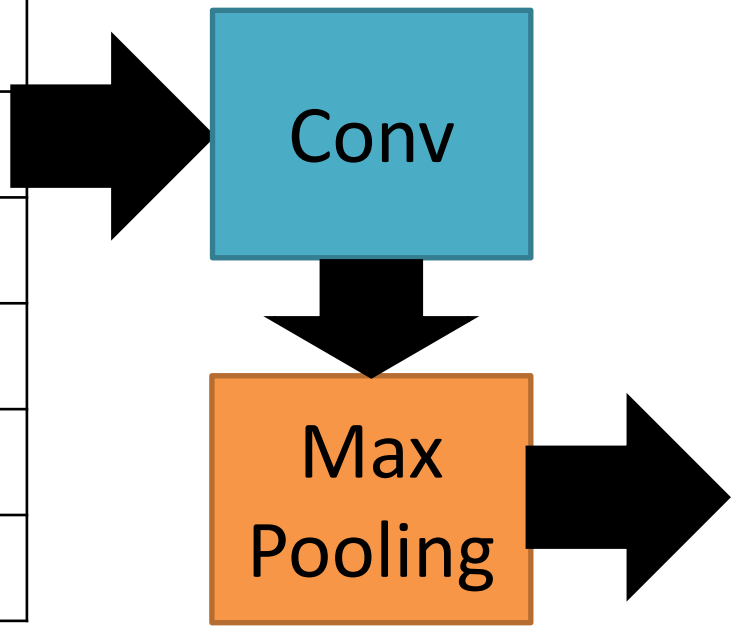
We can subsample the pixels to make image smaller

→ fewer parameters to characterize the image

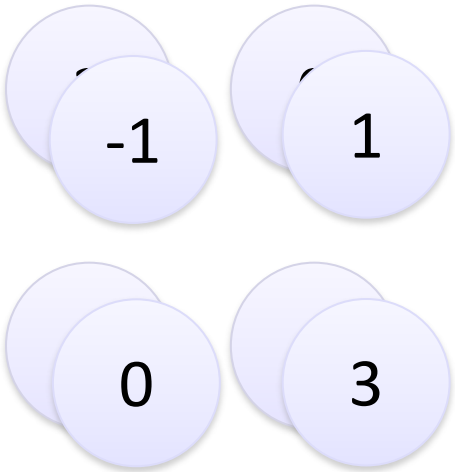
Max Pooling

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image



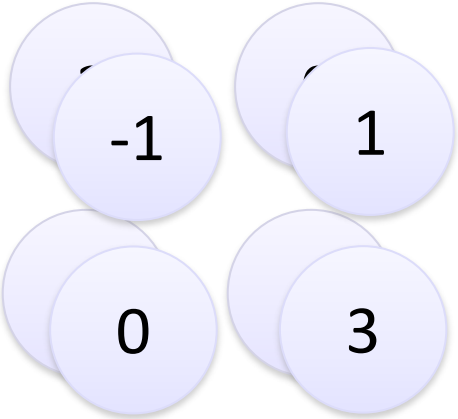
New image but smaller



2 x 2 image

Each filter is a channel

The whole CNN



A new image

Convolution

Max Pooling

Convolution

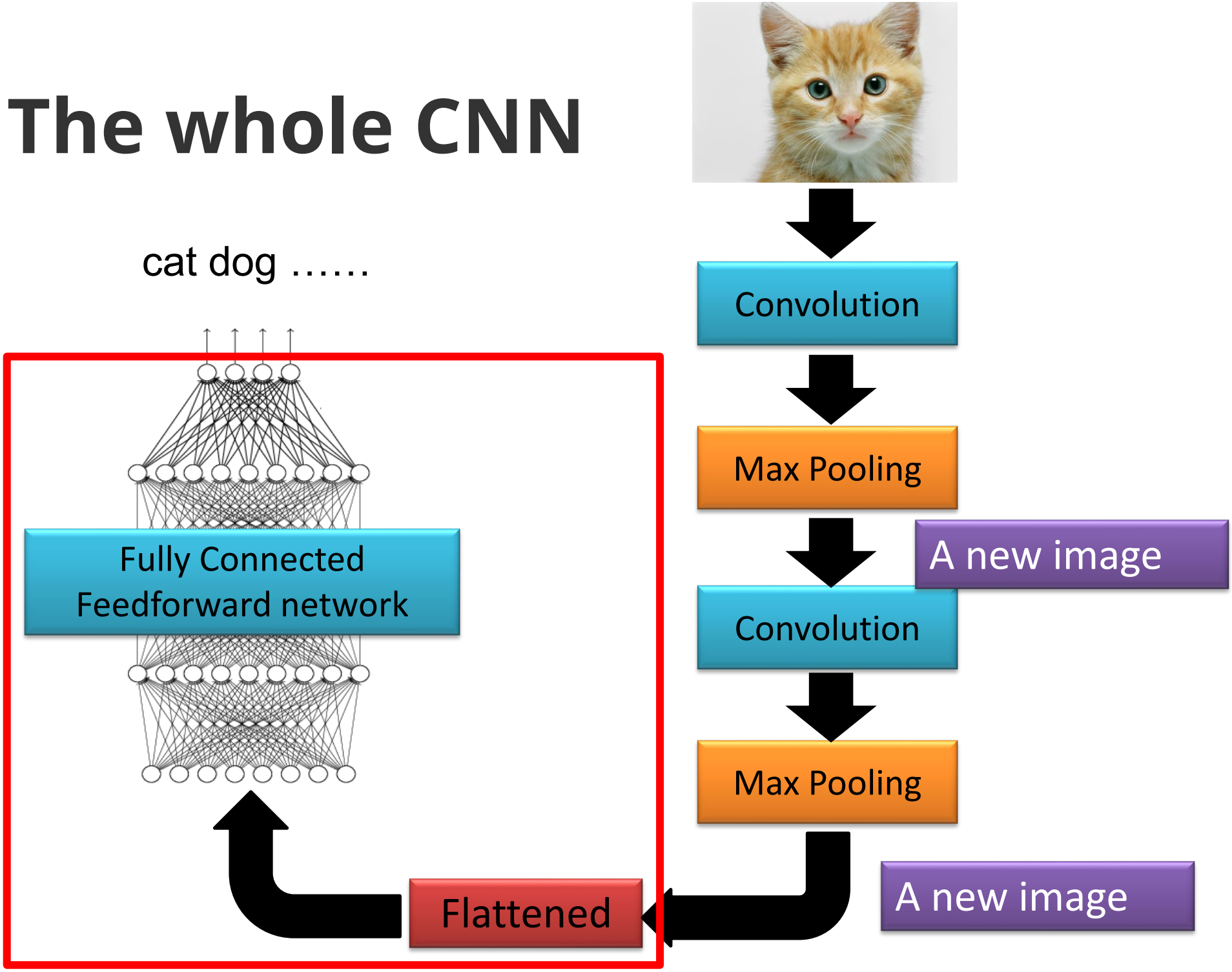
Max Pooling

Can repeat many times

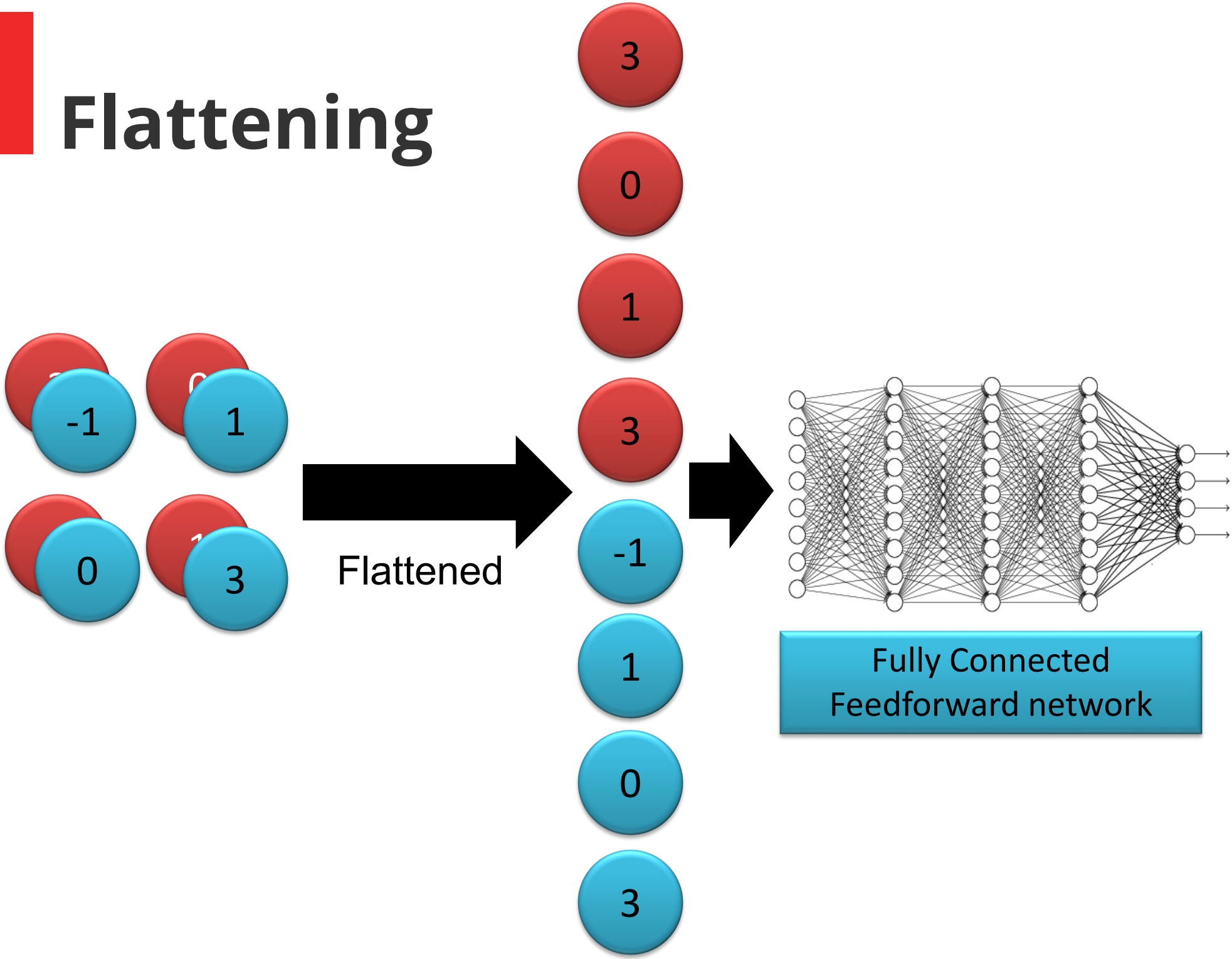
Smaller than the original image

The number of channels is the number of filters

The whole CNN



Flattening

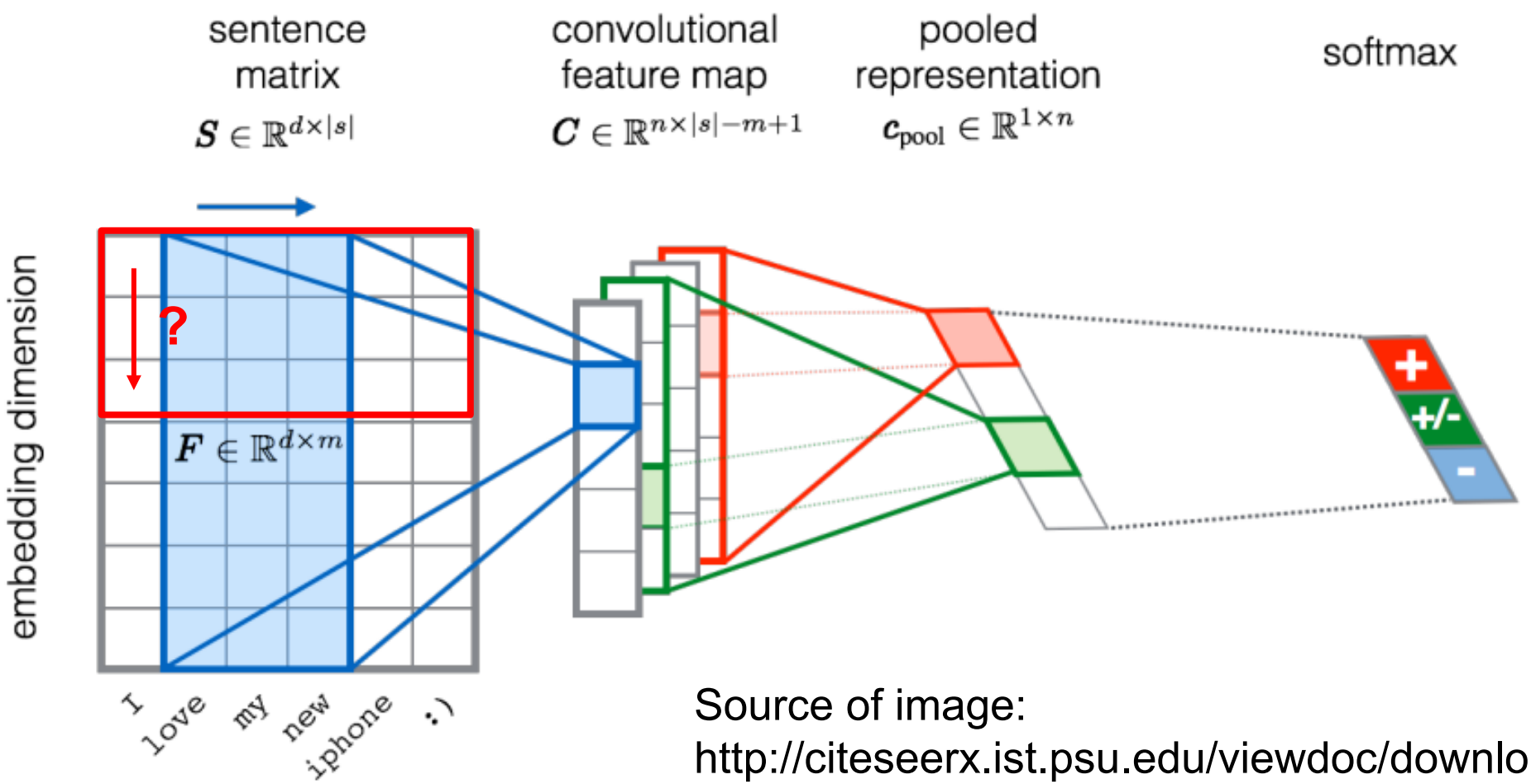

-1
1
0
3
Flattened
3
0
1
3
-1
1
0
3
Fully Connected
Feedforward network

Transfer on other data input

- How to deal with Text as input?
  - Transform to image?
  - Is the input parameter independent?

- Same requirements fitting?
  - Letters transform to integers?
  - Is the sequence order important?
  - Why smaller portions?

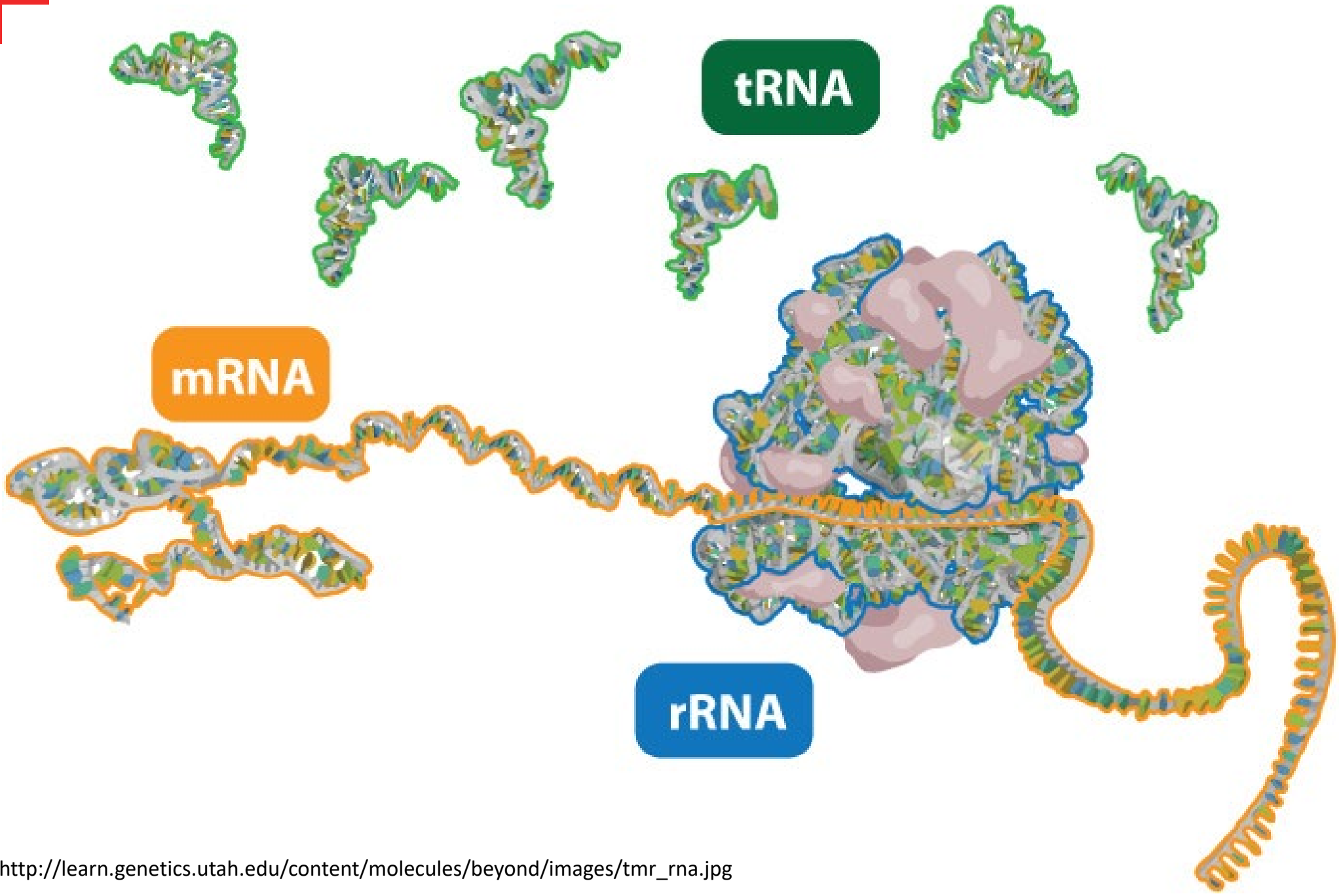
CNN in text classification



Source of image: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.703.6858&rep=rep1&type=pdf

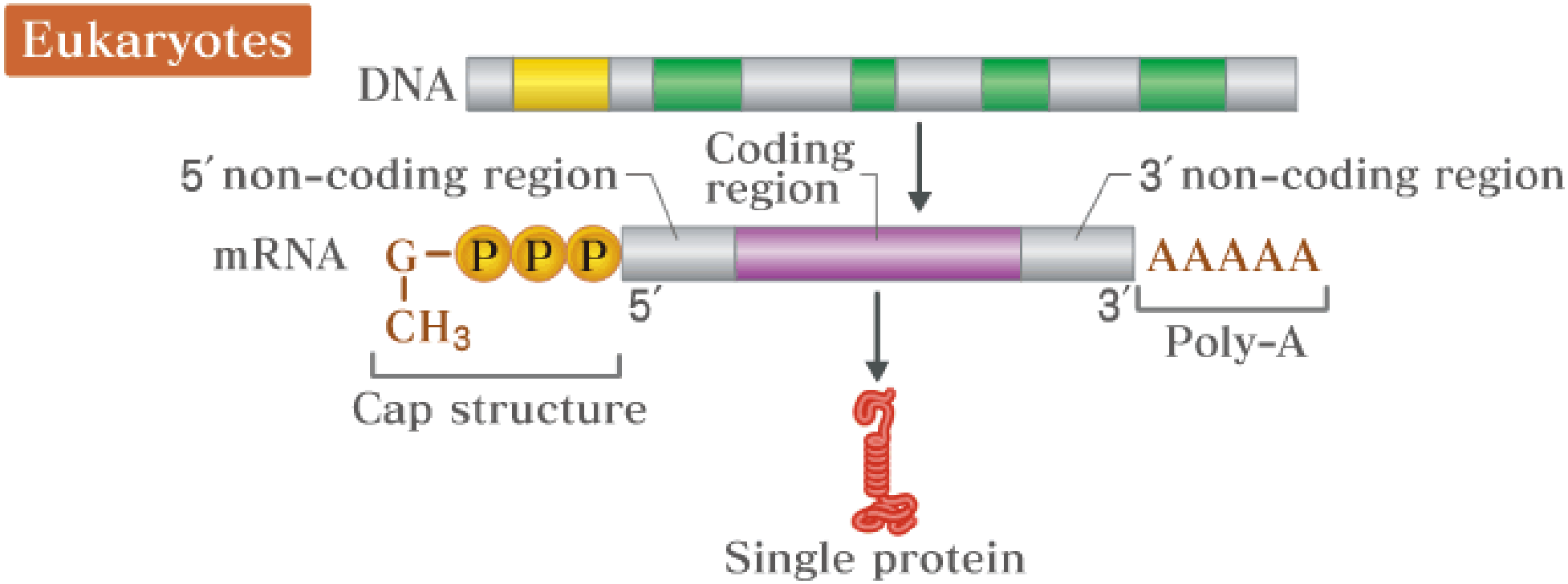
TYPES OF RNA AND THEIR SPECIFIC STRUCTURES

Types of RNA



http://learn.genetics.utah.edu/content/molecules/beyond/images/tmr_rna.jpg

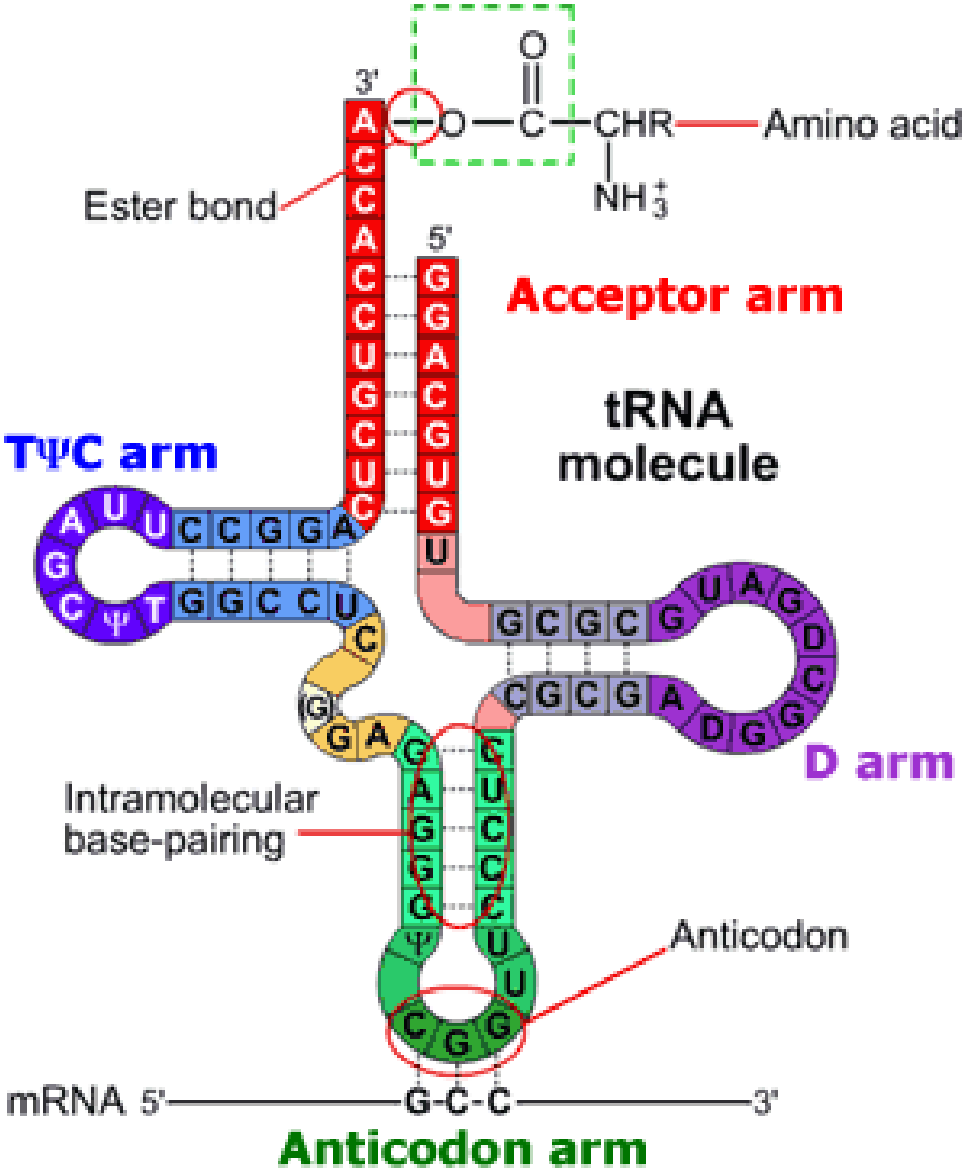
mRNA features / structure



- Comprises only 5% of the RNA in the cell
- Most heterogeneous in size and base sequence
- All members of the class function as messengers carrying the information in a gene to the protein synthesizing machinery

tRNA features / structure

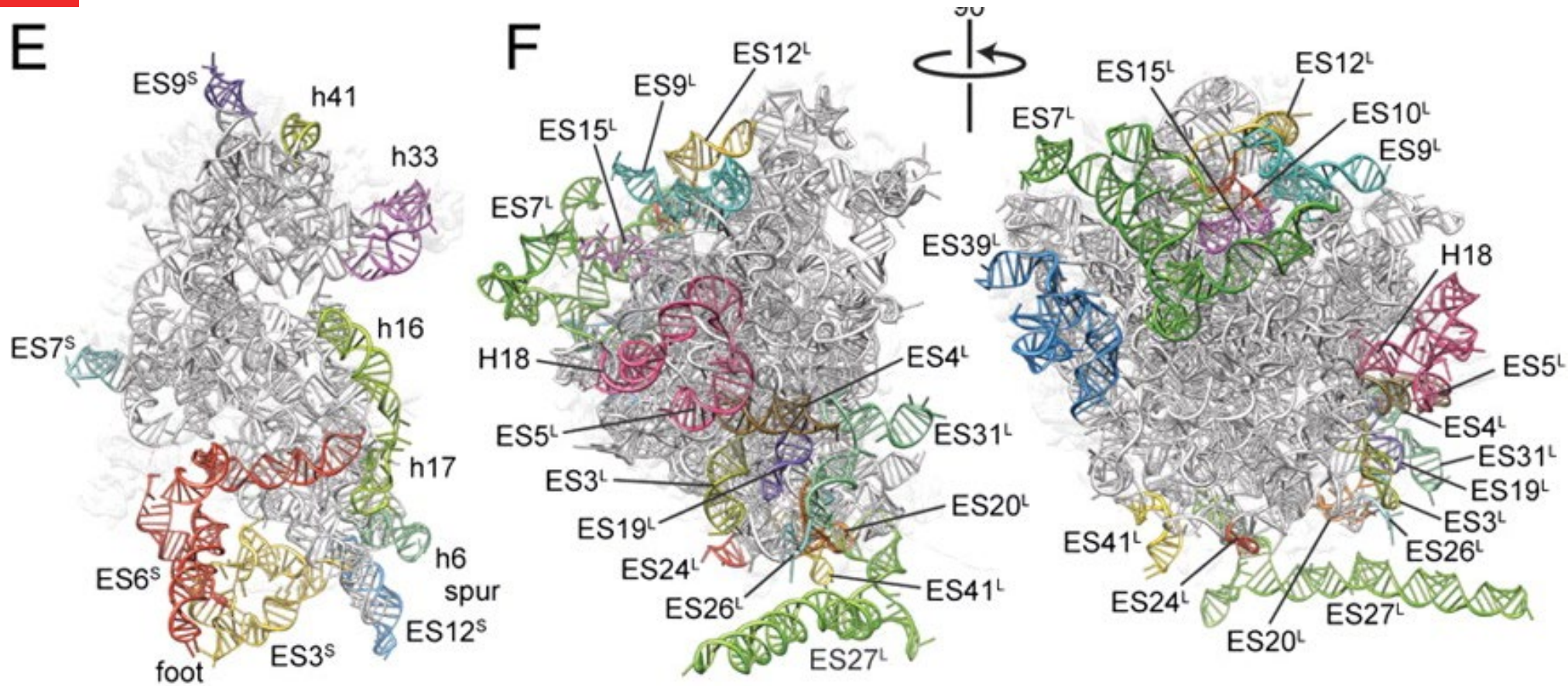
http://www.alamy.com/search-results.asp?qt=Friedrich+Saurer+Einstein&npg=7



http://higheredbcs.wiley.com/legacy/college/boyer/0471661791/structure/tRNA/trna_diagram_small.gif

- Small RNA species (74 – 95 nucleotides)
- Transfer amino acid to translation machinery (easily soluble RNA)
- At least 20 species for 20 amino acids required

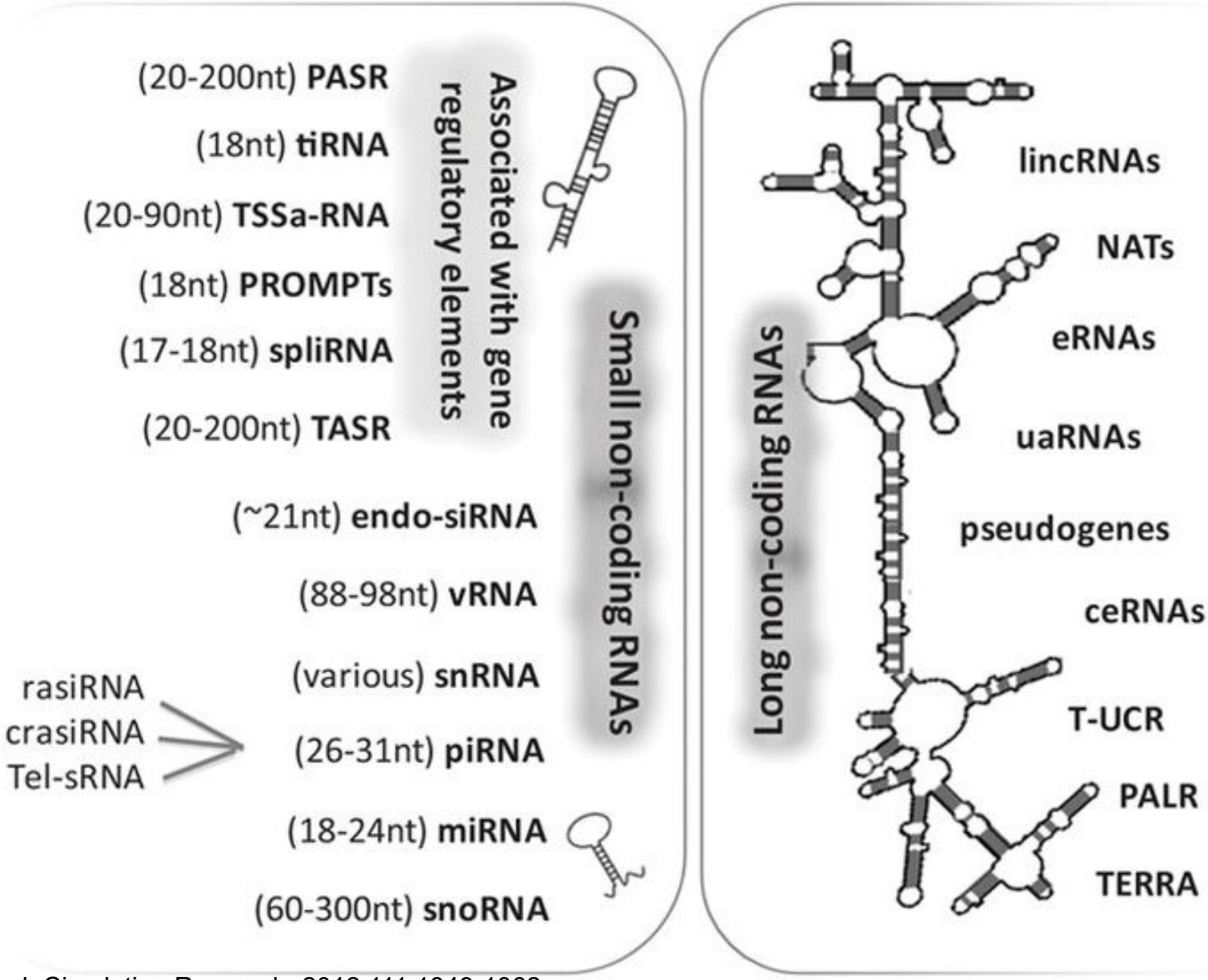
rRNA features / structure



http://www.pnas.org/content/107/46/19748/F3.large.jpg

- Large and small ribosomal subunit:
- 60S → 5S, 5.8S and 28S rRNA
- 40S → 18S rRNA
- 5S is independently transcribed
- All other comes from 45S precursor

non-coding (nc) RNAs



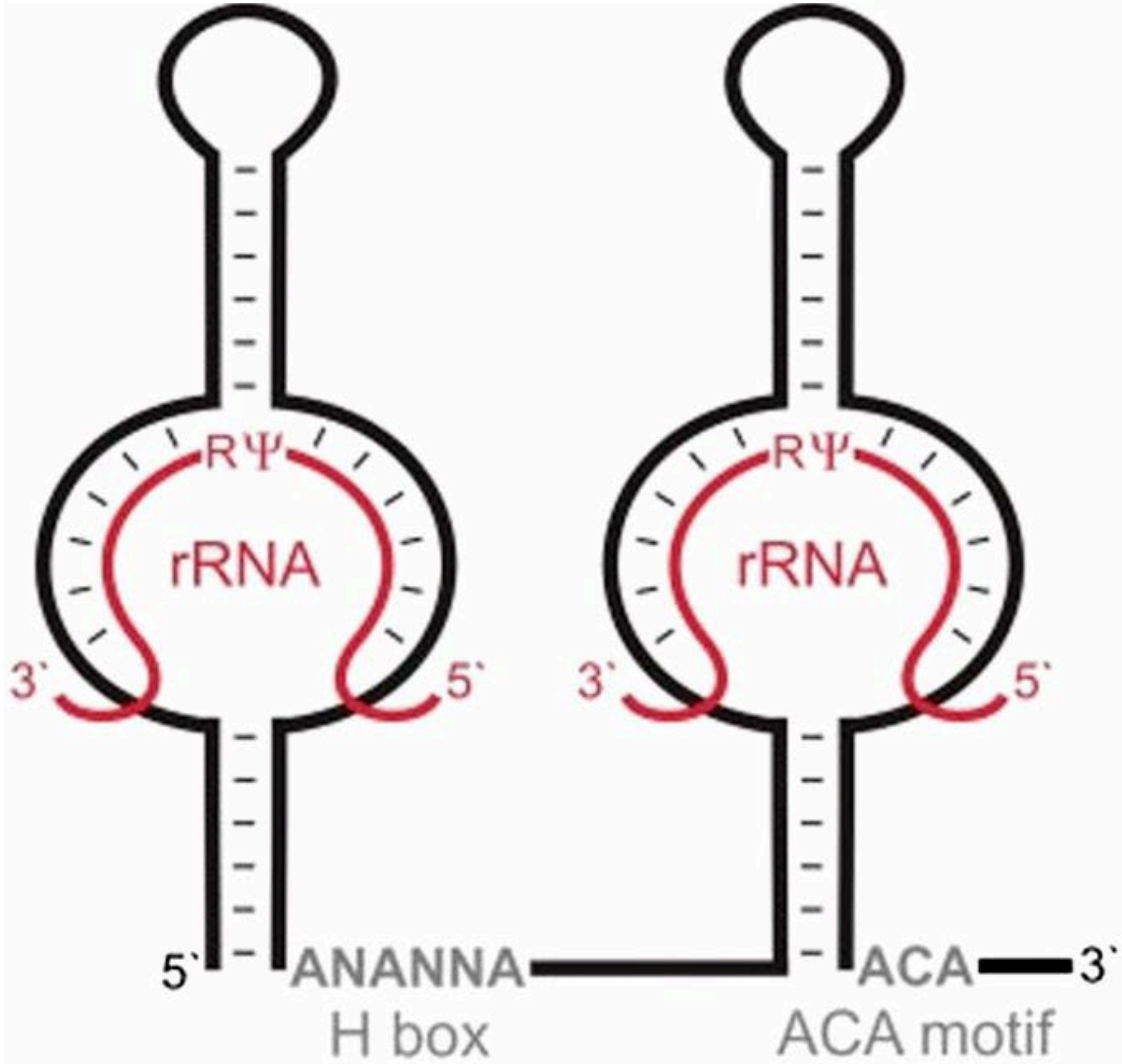
Schonrock N et al. Circulation Research. 2012;111:1349-1362

Aim for todays lesson

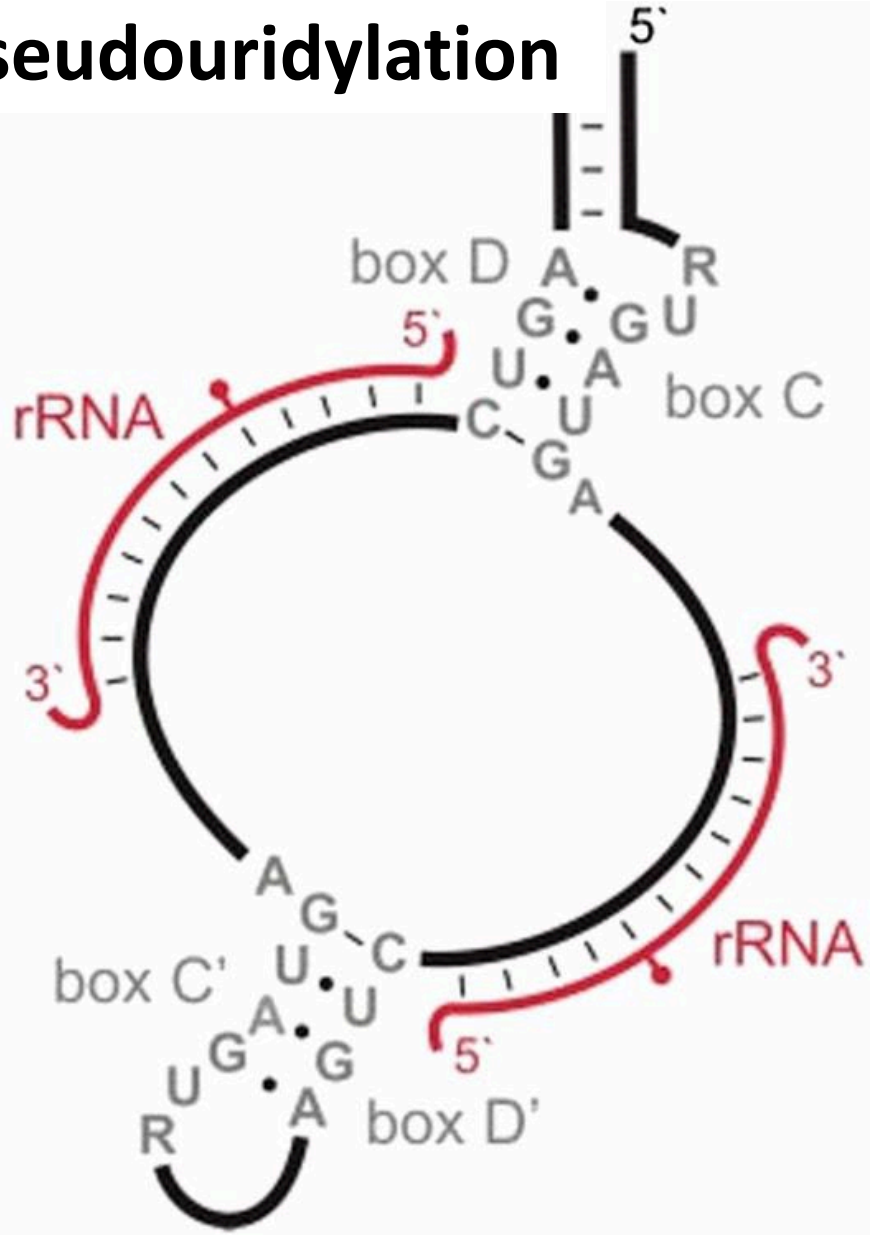
CNN FOR SEQUENCE CLASSIFICATION

snoRNAs

2′-*O*-ribose methylation

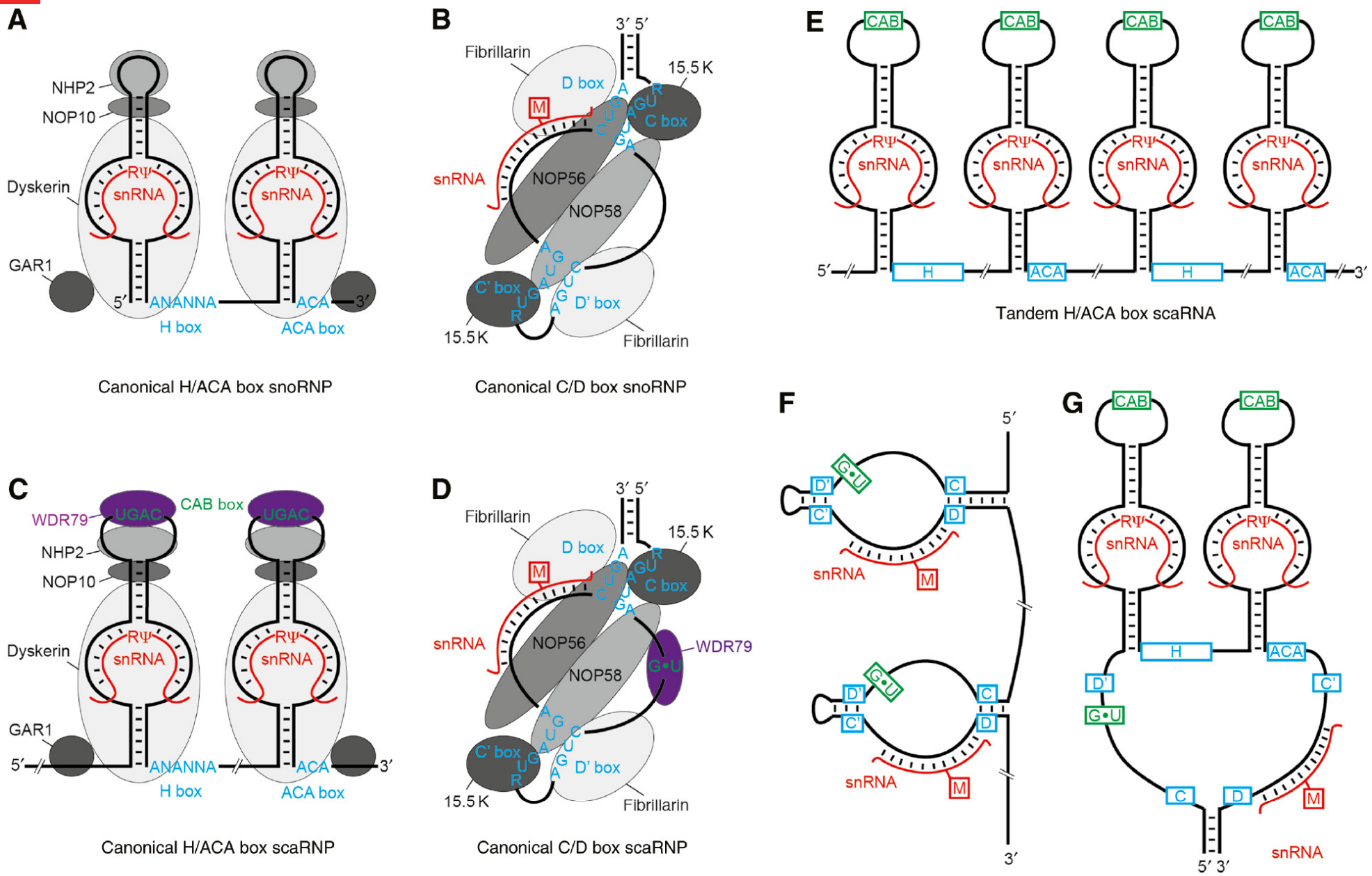


Pseudouridylation



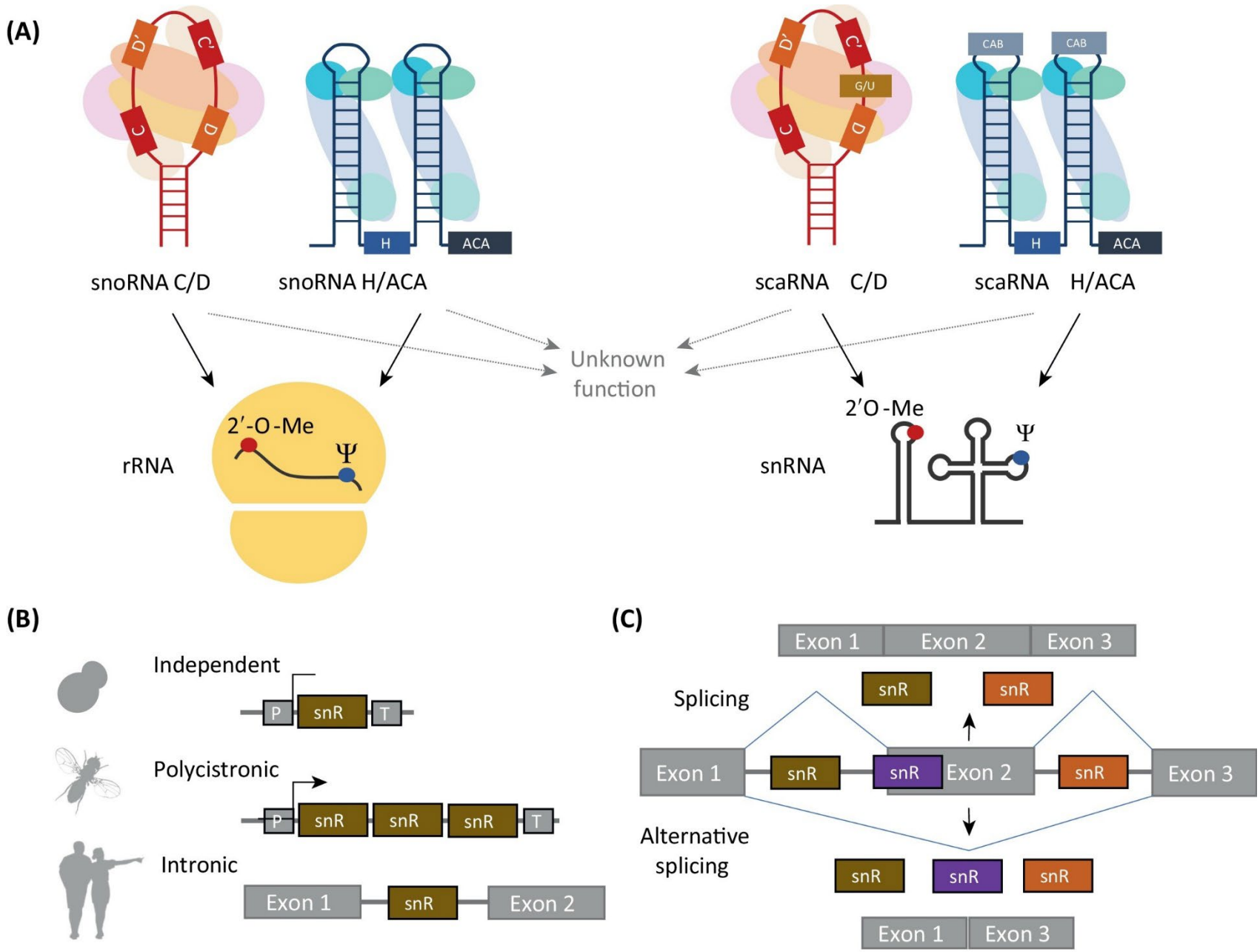
http://www.genetics.org/content/genetics/195/3/643/F6.large.jpg

scaRNAs



https://www.degruyter.com/document/doi/10.1515/hsz-2018-0205/html?lang=de

scaRNAs vs snoRNAs



https://www.cell.com/trends/genetics/fulltext/S0168-9525%2818%2930203-8

Trends in Genetics

The dataset

- **RNAcentral**
  - snoRNA (C/D-box, H/ACA-box), scaRNA
  - No hypotheticals, no partials

	snoRNA (H/ACA)	snoRNA (C/D)	scaRNA
Overall	89,522	152,235	6,747
Mammalia	57,640	34,286	4,094
Insecta	976	2,388	47
Fish (Actinopterygii)	7,320	11,530	752
Plants (Viridiplantae)	5,808	76,445	28
Fungi (Ascomycota)	5,375	7,835	1

- **Preprocessing step**
  - Clustered via CD-HIT to a max. of 0.9 (90%) similarity
  - Balanced to 1913 sequences per class

Workflow

- Get data
- Inspect data
- Split data into Train / Test sets
- Create model
- Train model
- Validate model
- Run 10-fold cross validation
- Extend to different input features
- Compare accuracy

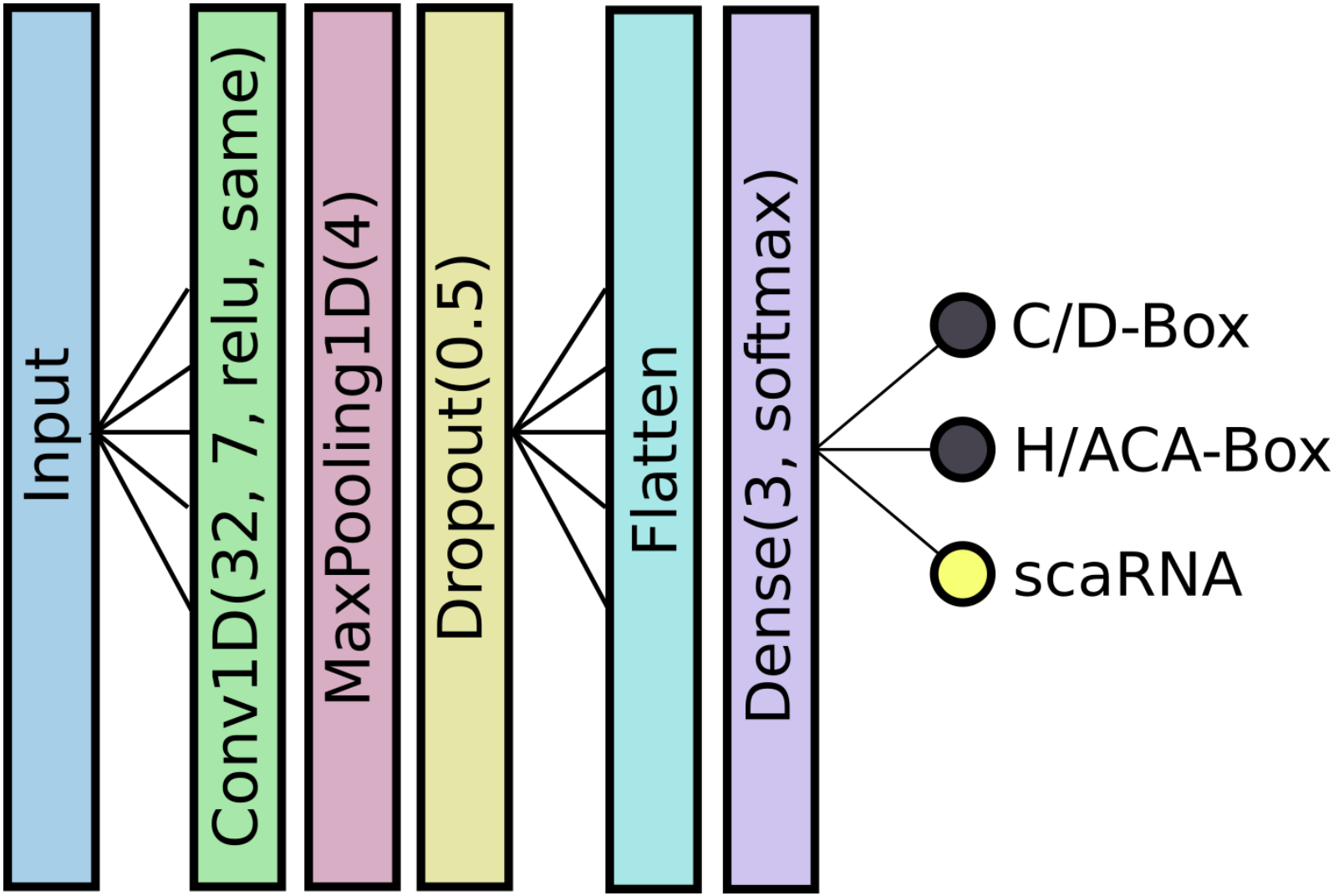
Repeat for optimiziation

First Notebook

Second Notebook

Third Notebook

CNN our Standard Architecture



Let’s Classify