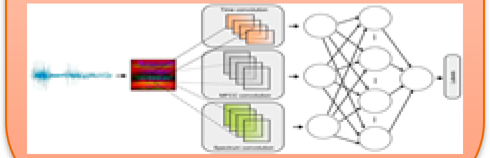
Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

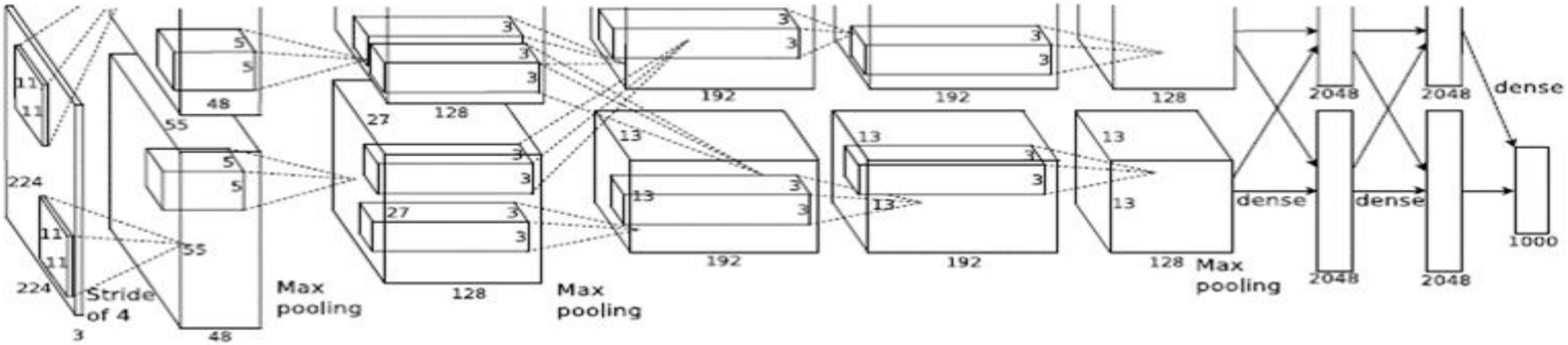
- ✓ Deep Learning Architectures
- ✓ CNN
- CNN in Images, Speech and Text



CNN in Images, Speech and Text

"CNN for ALL"

Recap: Turning Point: AlexNet



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

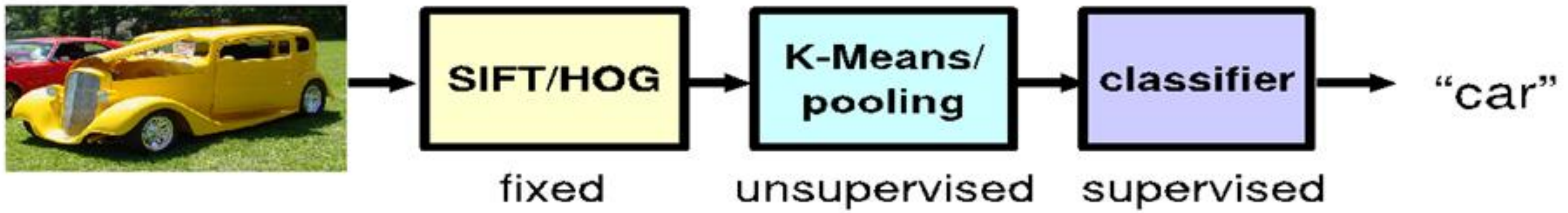
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

ImageNet Classification Task:

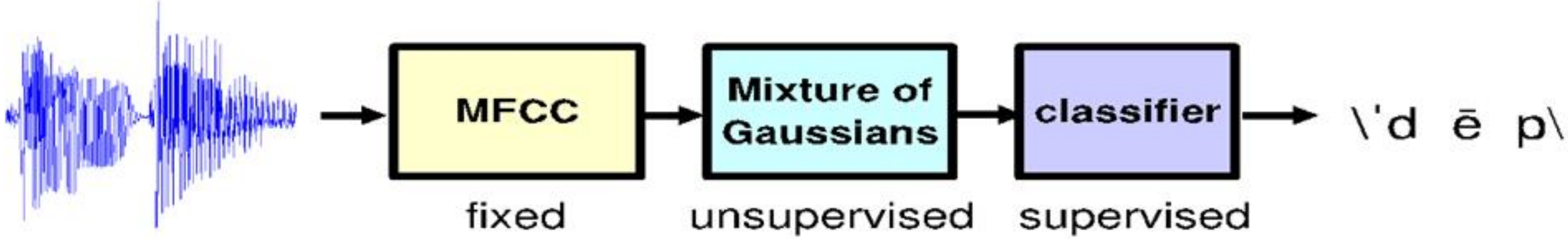
Previous Best : ~25% (CVPR-2011)
AlexNet : ~15 % (NIPS-2012)

Common Pipeline: Till Then

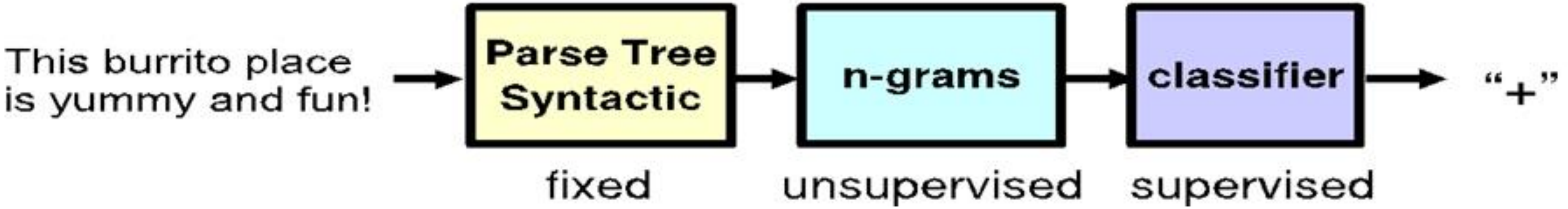
- VISION:



- SPEECH:



- NLP:

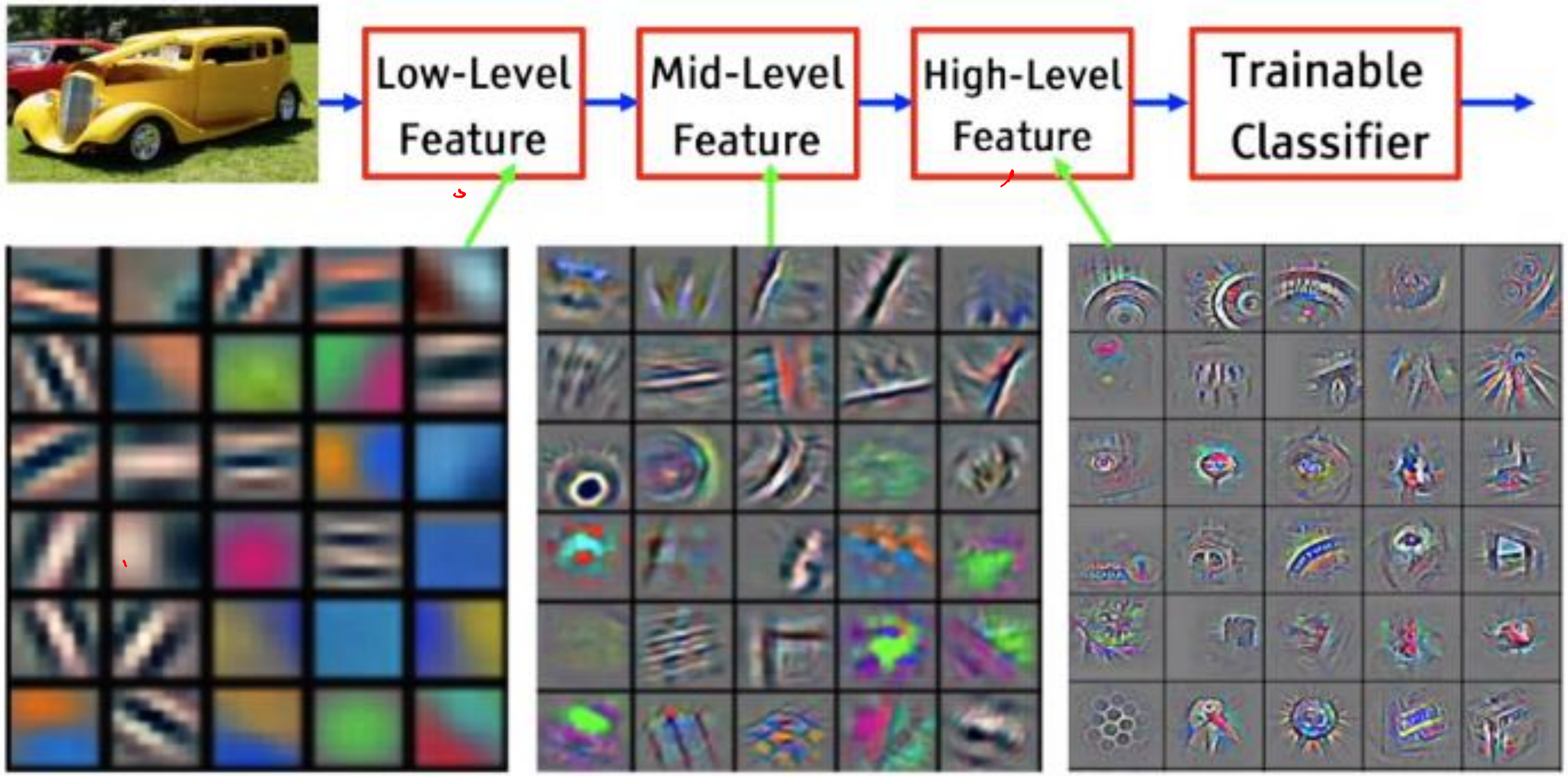


Learn the full pipeline

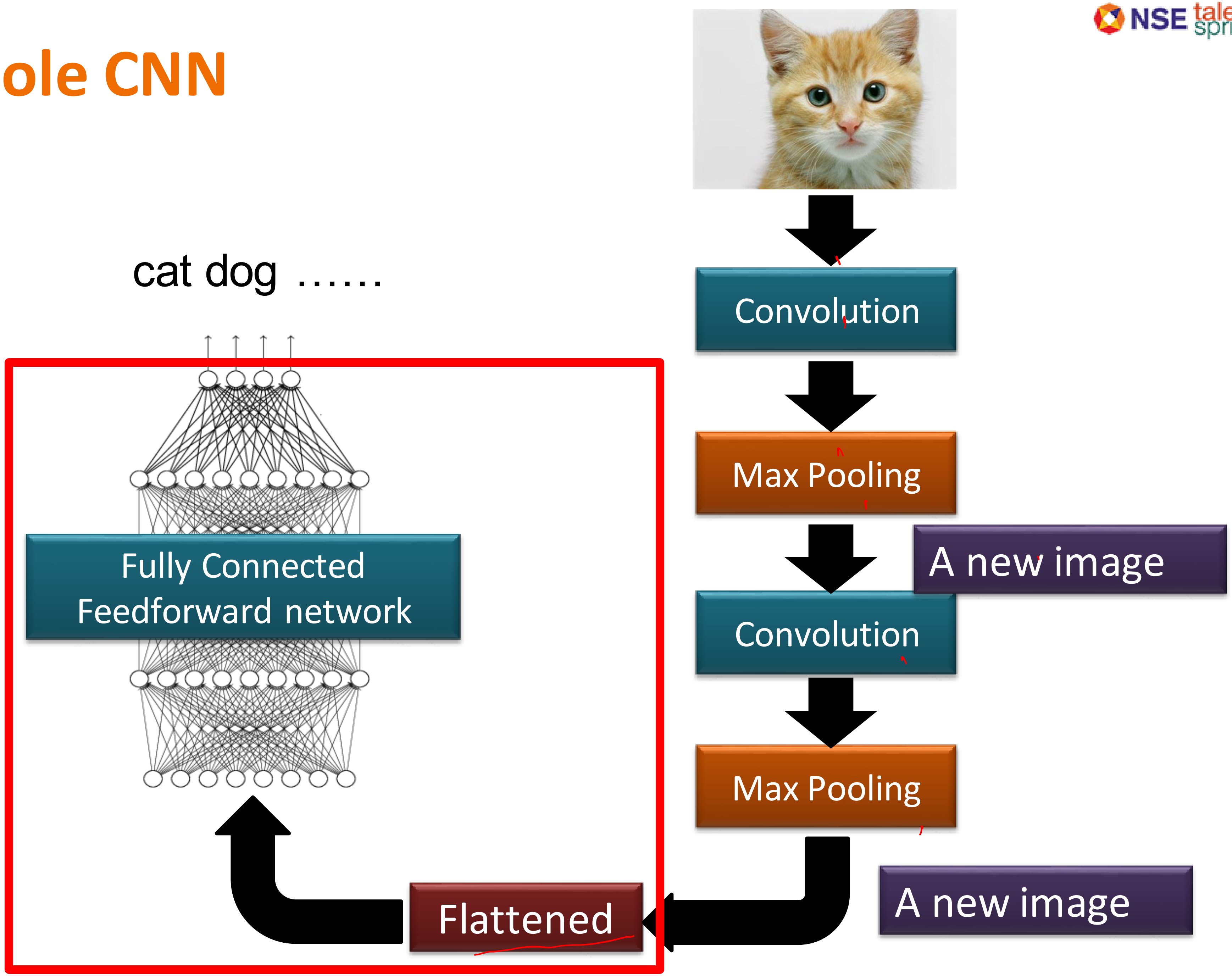
- VISION:
  – Pixels → edge → texton → motif → part → object
- SPEECH:
  – Sample → spectral → band → formant → motif → phone → word
- NLP:
  – Character → word → NP/VP/.. → Clause → sentence → story

Deep Learnt Features

- It's deep if it has more than one stage of non-linear feature transformation.

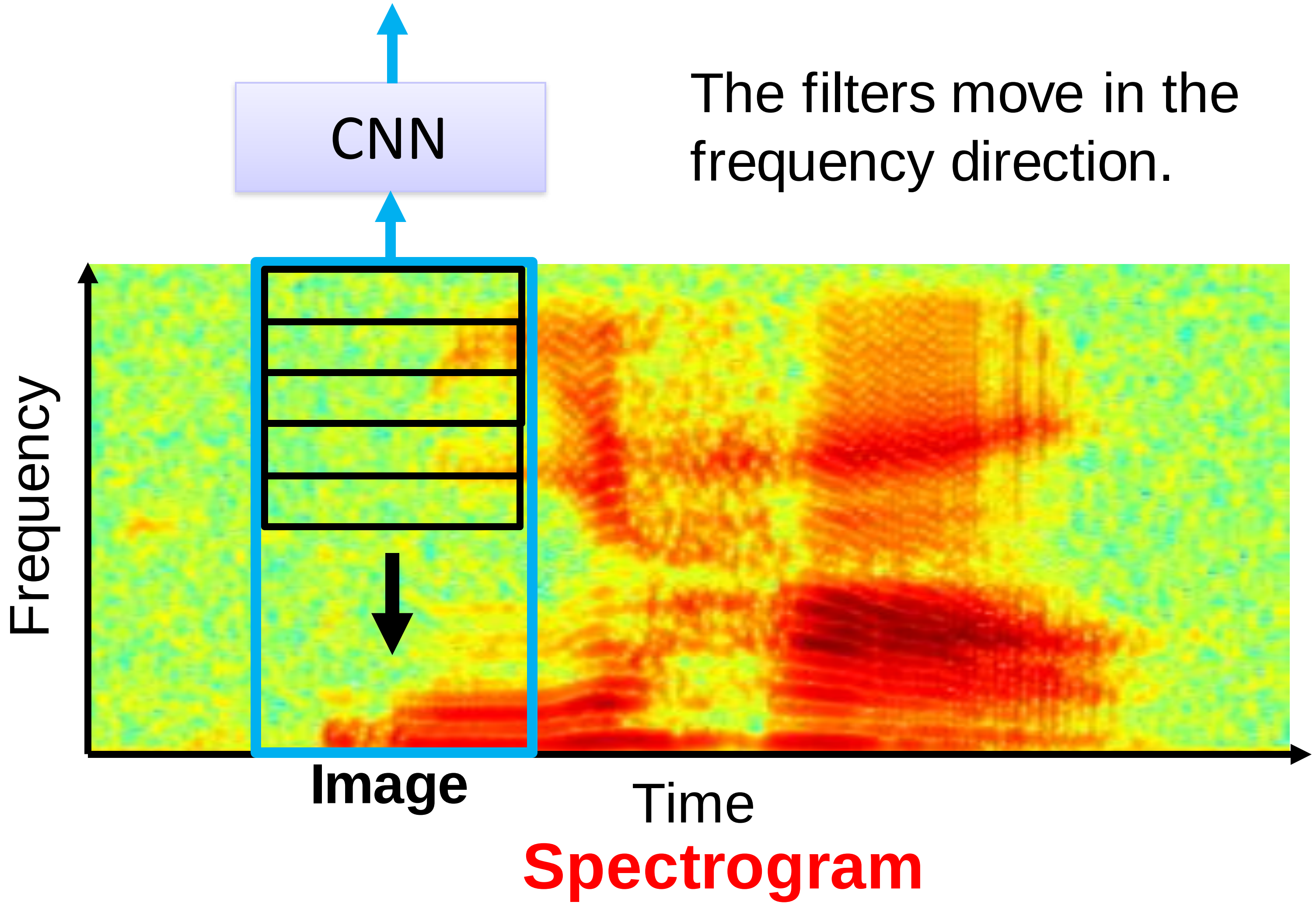


The whole CNN

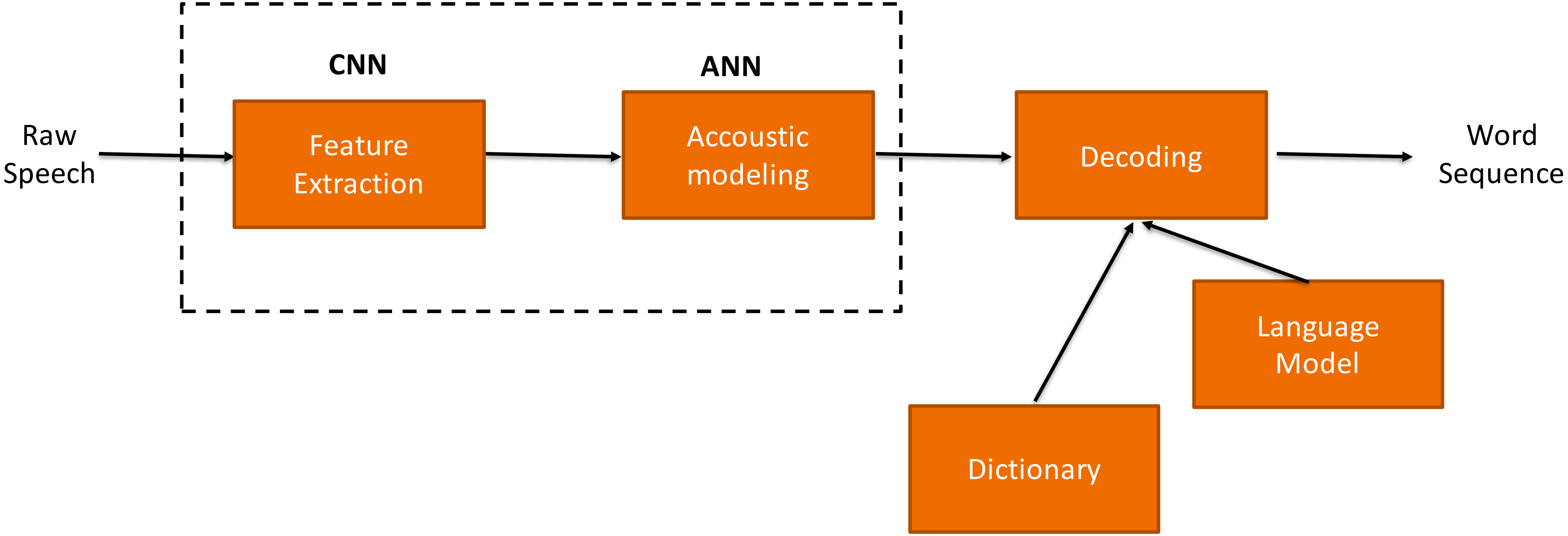


CNN in speech recognition

The filters move in the frequency direction.

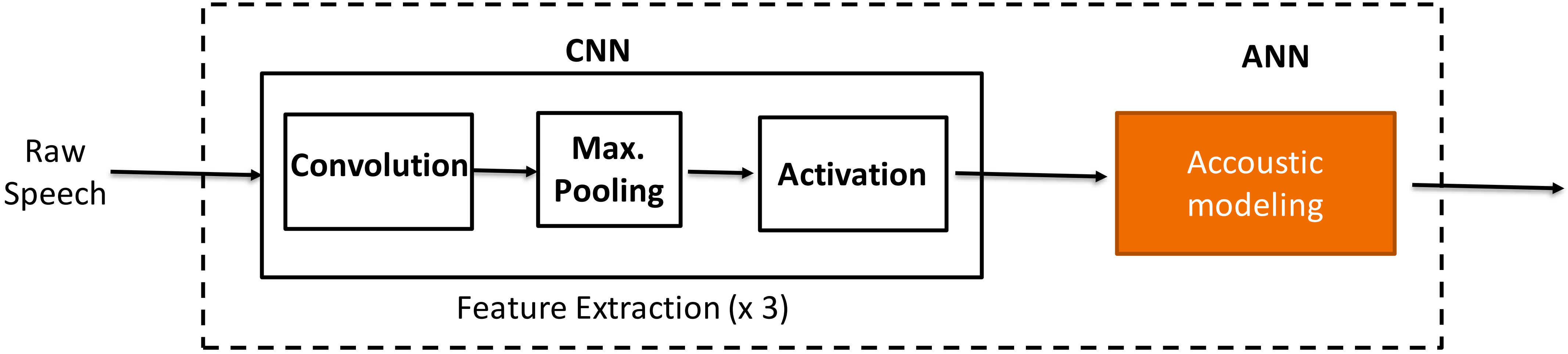


Speech Recognition model

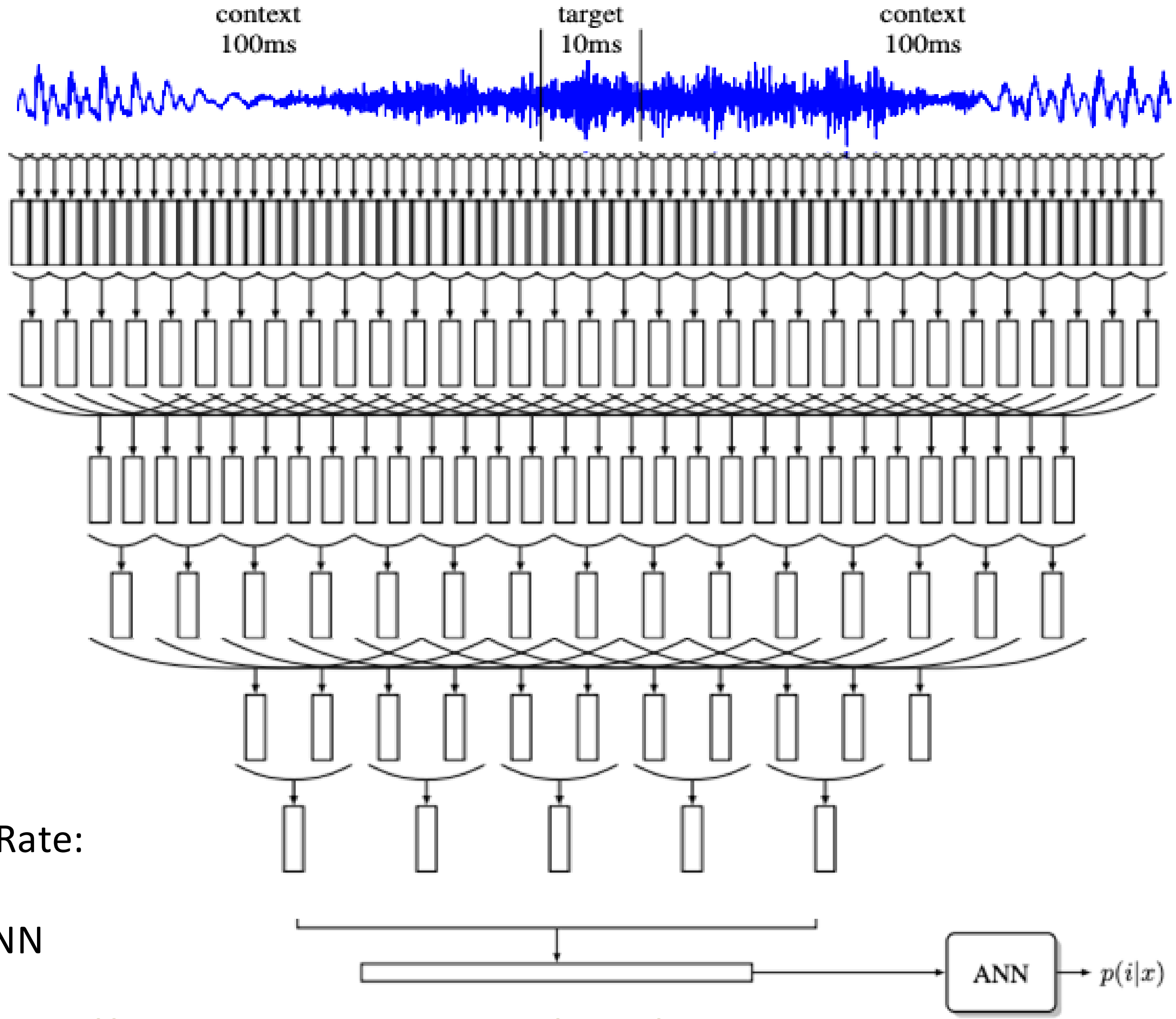


Dimitri Palaz, Mathew Magimai-Doss and Ronan Collobert, “Convolutional Neural Networks-based Continuous Speech Recognition using Raw Speech Signal”, ICASSP 2015, pp.4295-4299.

Speech Recognition model



Detailed View



Convolution

Max pooling

Convolution

Max pooling

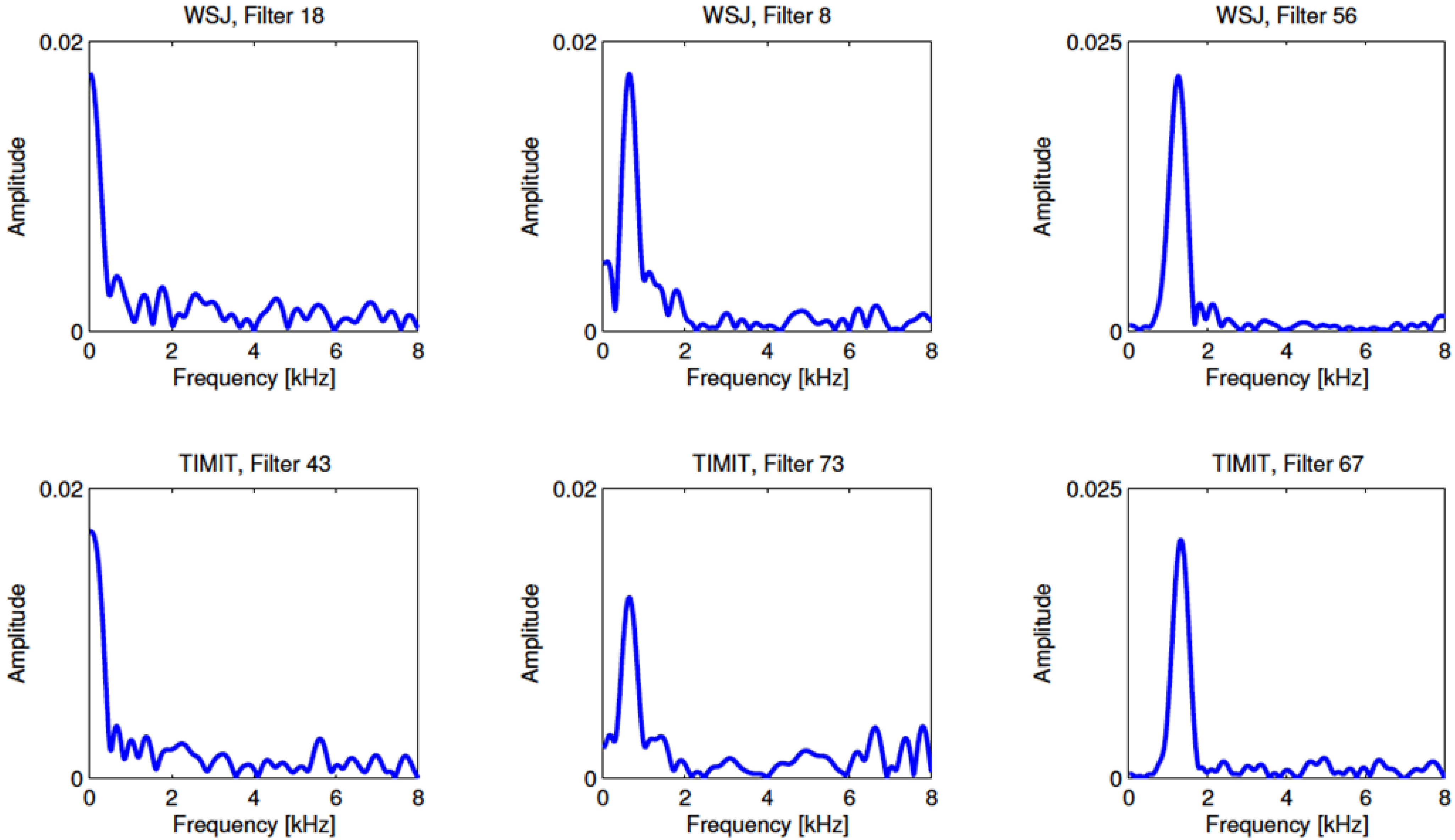
Convolution

Max pooling

WSJ Word Error Rate:
64.% vs 5.6% for
MFCC+ANN vs CNN

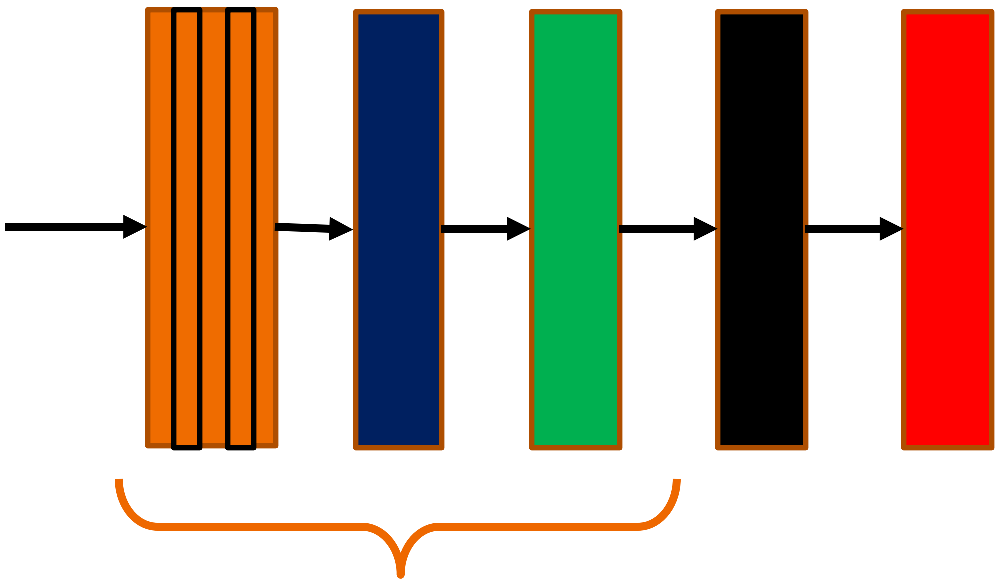
https://pdfs.semanticscholar.org/d224/6d41fe4ca1cb1f13eaa7bec0b86b95fdd8ac.pdf

Comparing Learnt Filters: WSJ vs. TIMIT



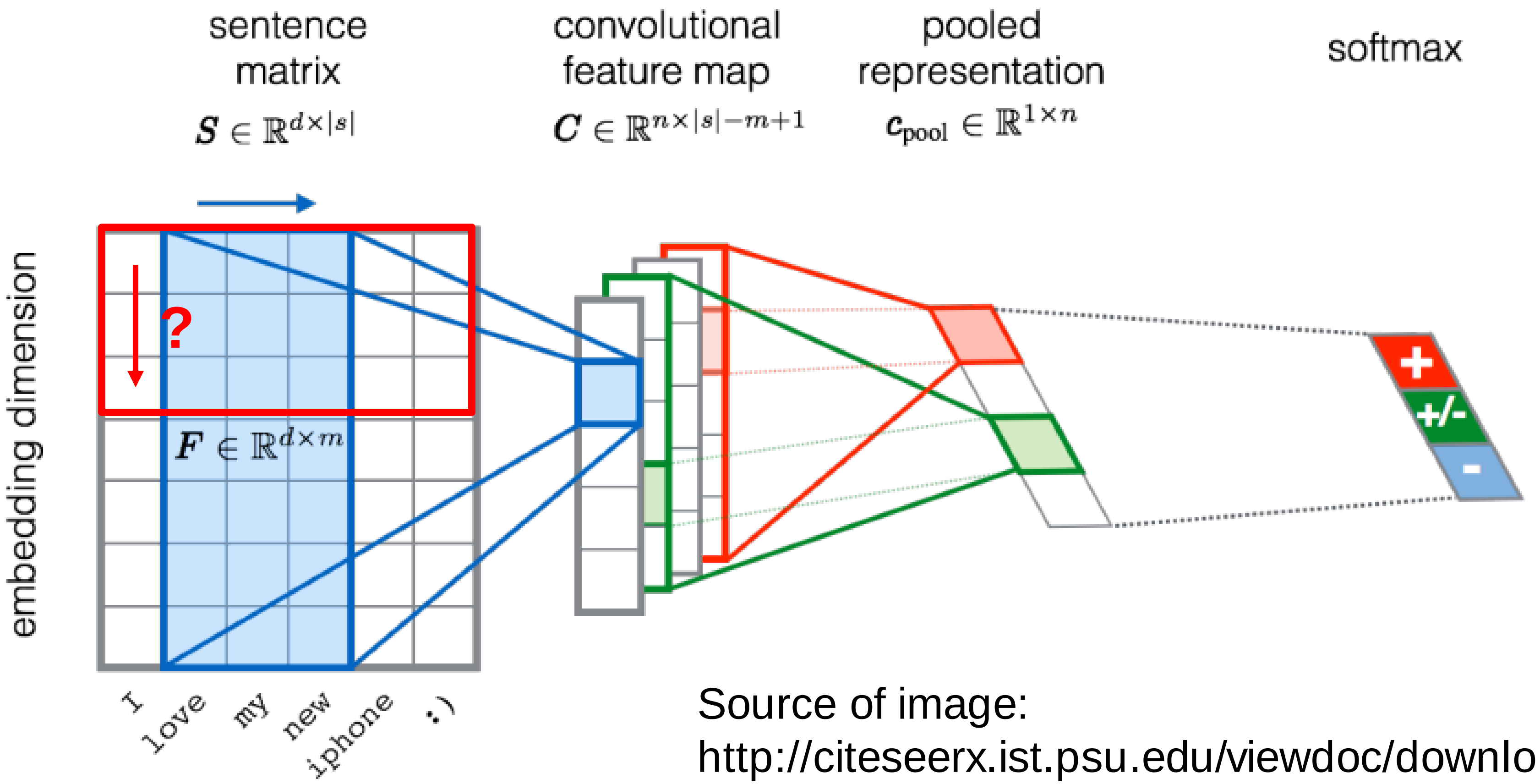
Summary: Layers of a Neural Network

- Based on the connection pattern and operations, we can think of a layer in a Neural Network as:
  - Convolutional
    - A Layer can have multiple Channels
  - Non-Linear (often not drawn)
  - Max-Pooling
  - Fully Connected
  - Soft Max



This is often repeated multiple times

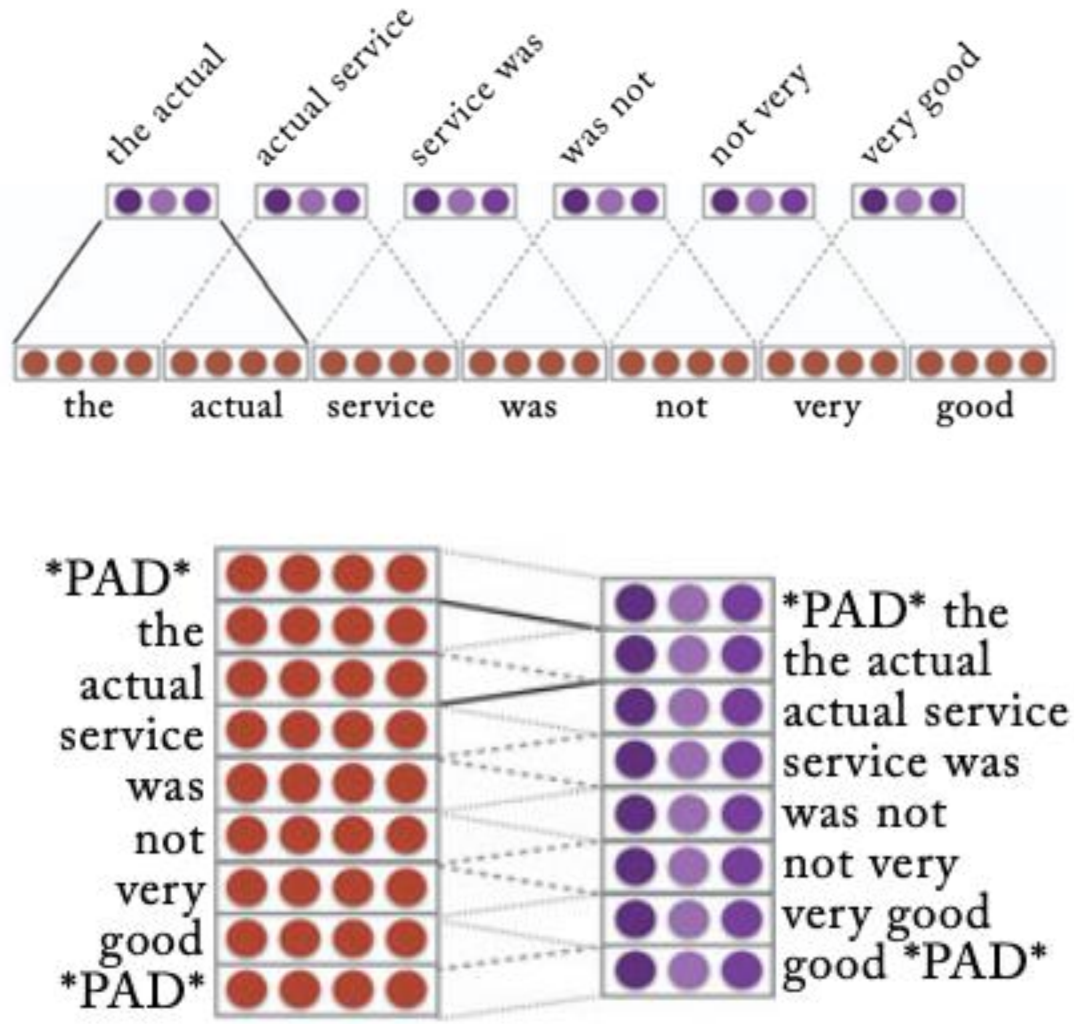
CNN in text classification



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

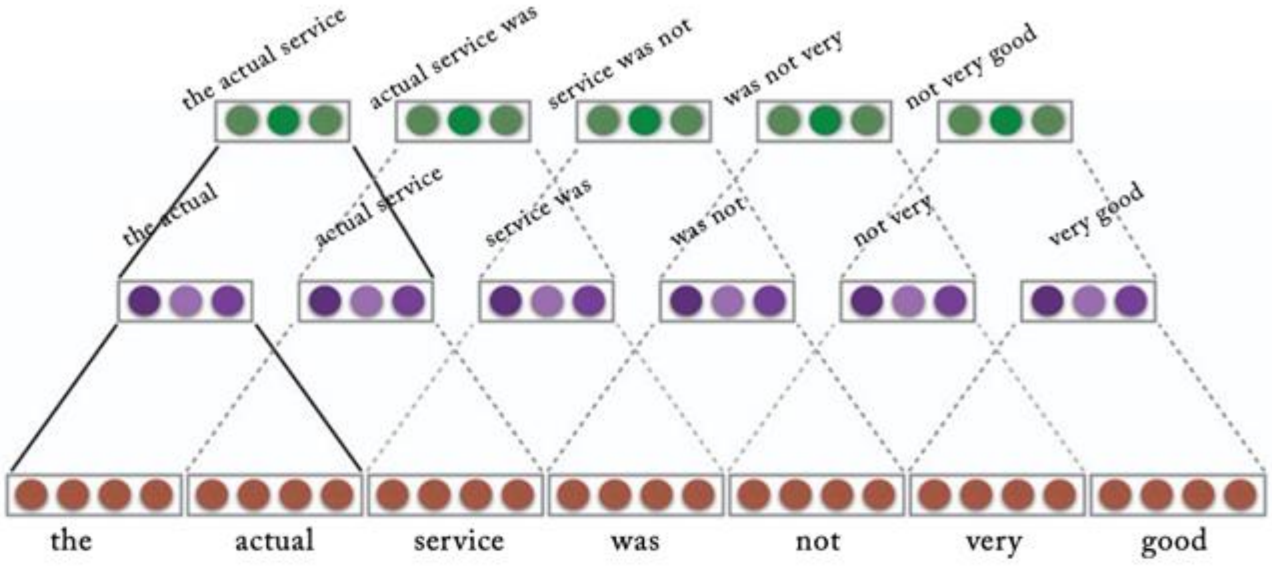
Convolution

- n-word sliding window over the sentence
- Learning to identify indicative ngrams in the input
- Filter transforms a window of k words into a scalar value
- Several filters applied to capture properties of the words in the window



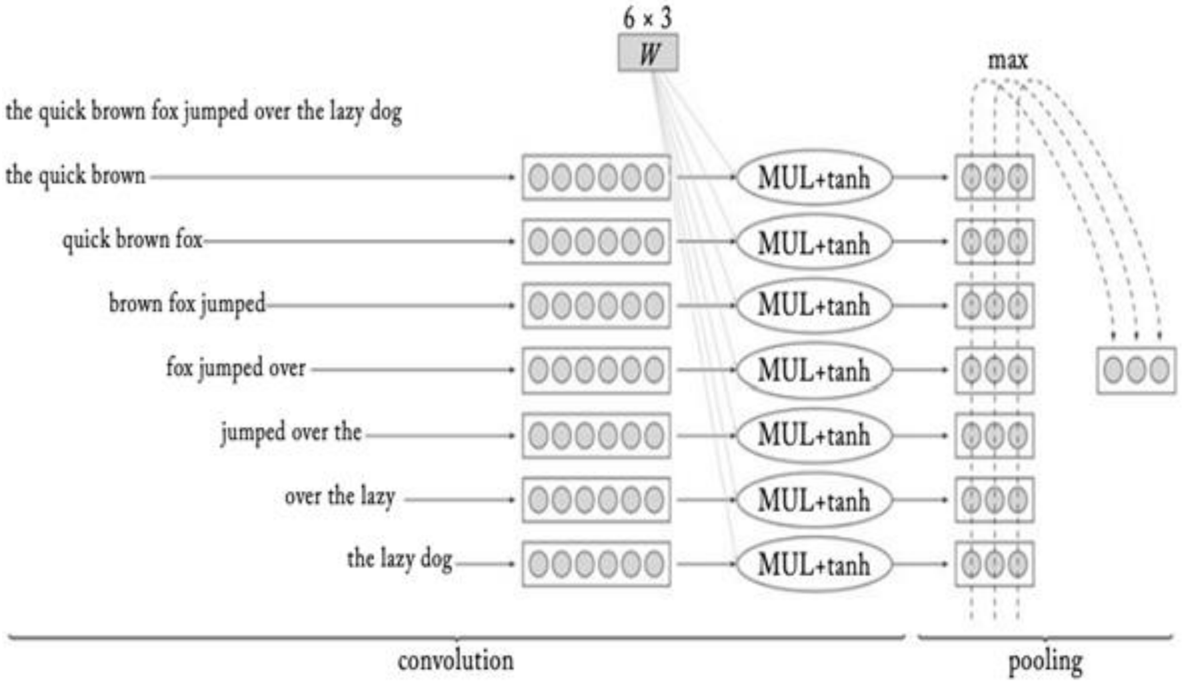
HIERARCHICAL CONVOLUTIONS

- This approach can be extended to hierarchy of convolutional layers
- Increase in depth of CNN leads to capture increasingly larger effective windows for a sentence
- Dilated convolutions help in learning the relationship between Non-Adjacent words

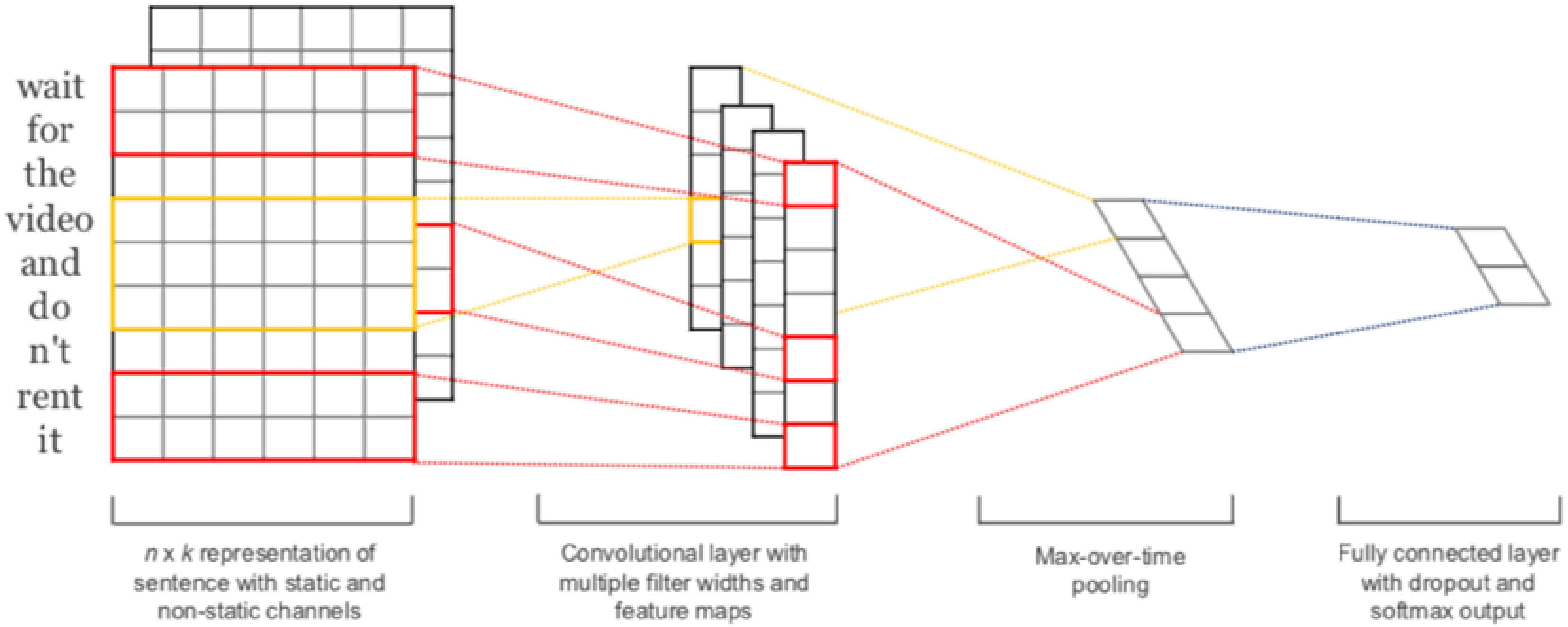


Pooling

- Pooling operation combines the vectors from the different windows into a single dimensional vector
- By taking the max or the average value
- The intention is to focus on the most important features
- Each filter extracts a different indicator from the window
- Pooling operation zooms in on the important indicators



Convolutional Neural Networks for Sentence Classification



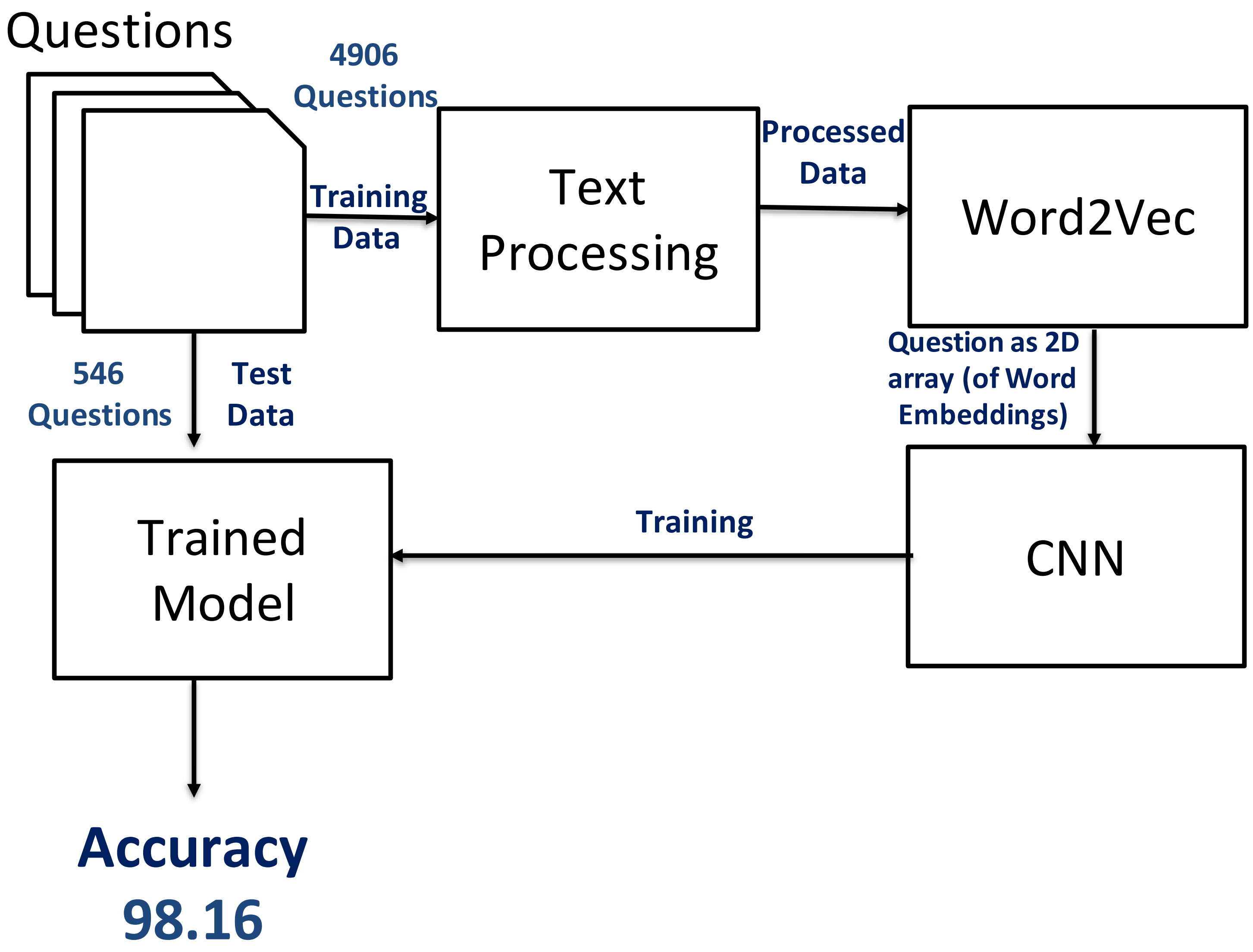
Convolutional Neural Networks for Sentence Classification, Yoon Kim

Case Study

Text Classification with CNNs

NLP with CNN

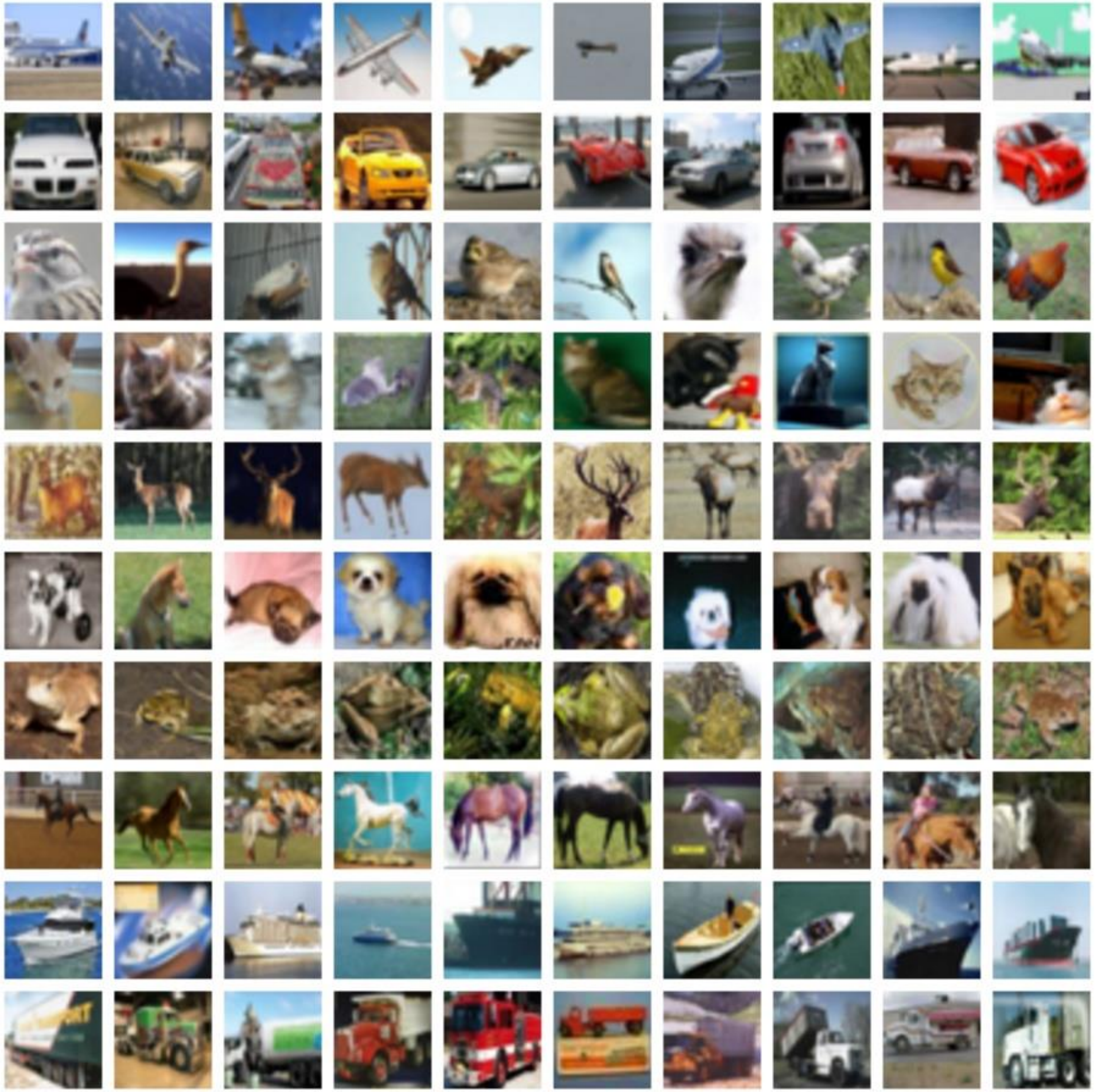
- Problem: Classification of Questions
  - Data: Various questions labeled based on the type
  - Labels: Abbreviation, Entity, Description, Human, Location and Numeric value.
- Process the data: Remove punctuations etc.
- Represent the Sentence as a 2D array with word embeddings
- Train the CNN
- Use the trained model for Testing



http://aclweb.org/anthology/C02-1150

Is there an inherent hierarchy in natural images?

Can you identify any structure or common elements in these images?



CIFAR10 dataset
Image source: [CIFAR-10 image classification with Keras ConvNet - Giuseppe Bonaccorso]

Hierarchy of visual elements

Parts of images: one step down the hierarchy

Even lower down the hierarchy

Image source: Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks. *ECCV 2014*,

Even lower down the hierarchy:

Textures and colours

Image source: Zeiler, Matthew D., and Rob Fergus. 2014. “Visualizing and Understanding Convolutional Networks. *ECCV 2014*,

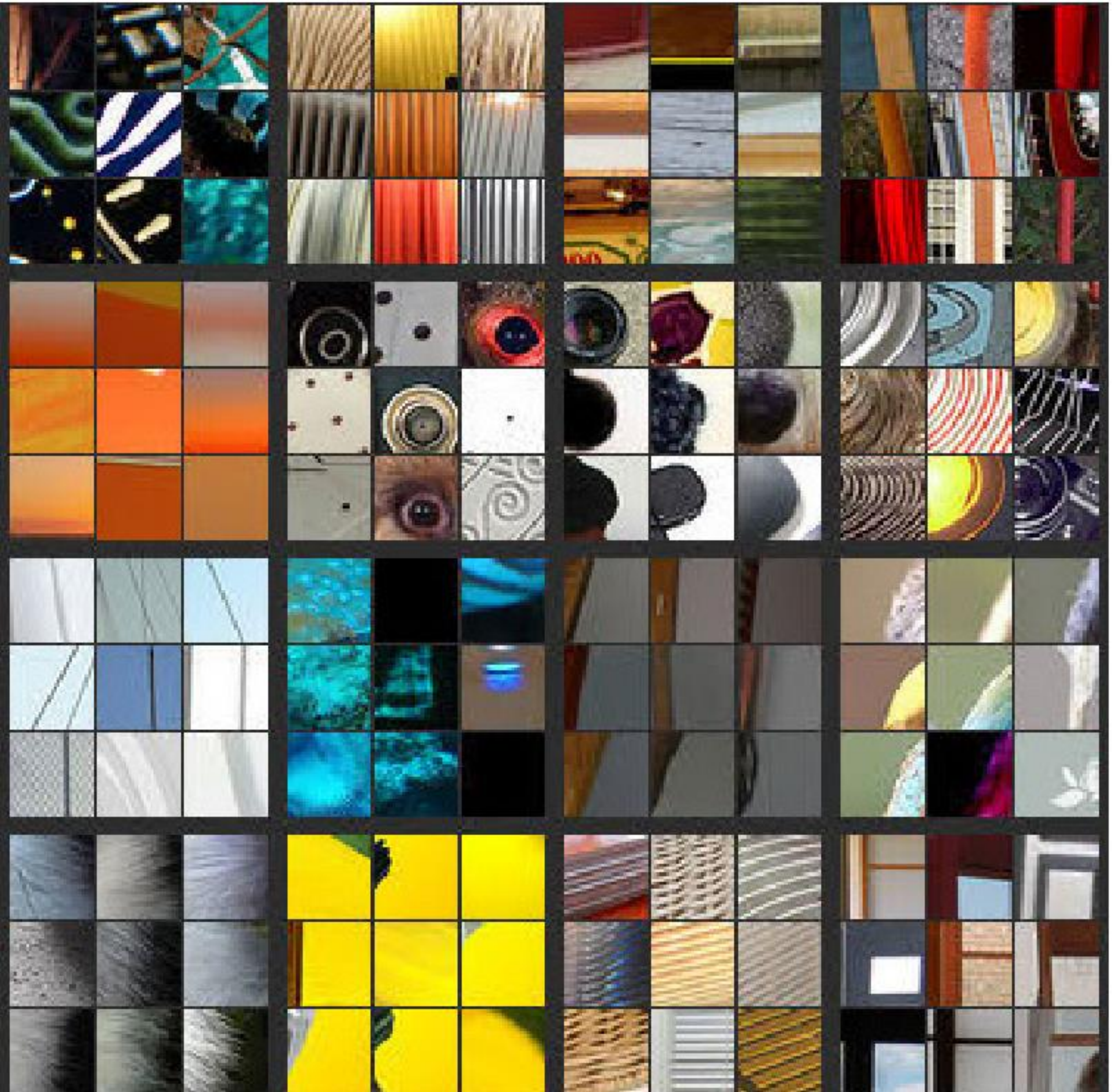


Image source:
Zeiler, Matthew D., and Rob Fergus. 2014. “Visualizing and Understanding Convolutional

Hierarchy in face images?

Highest level: whole faces, identities

Same Person still Looks Different

Image source: [Face clustering with Python - PyImageSearch]

Face parts

Much more uniform than whole faces

Different Person But Looks Different

Still Smaller: Textures? Even more uniform

Do CNNs work in a similar way?

Convolution as Part Search

Higher level filters

- If we create a filter for each pattern we want to recognize, there will be too many variations!
- Utilize image hierarchy

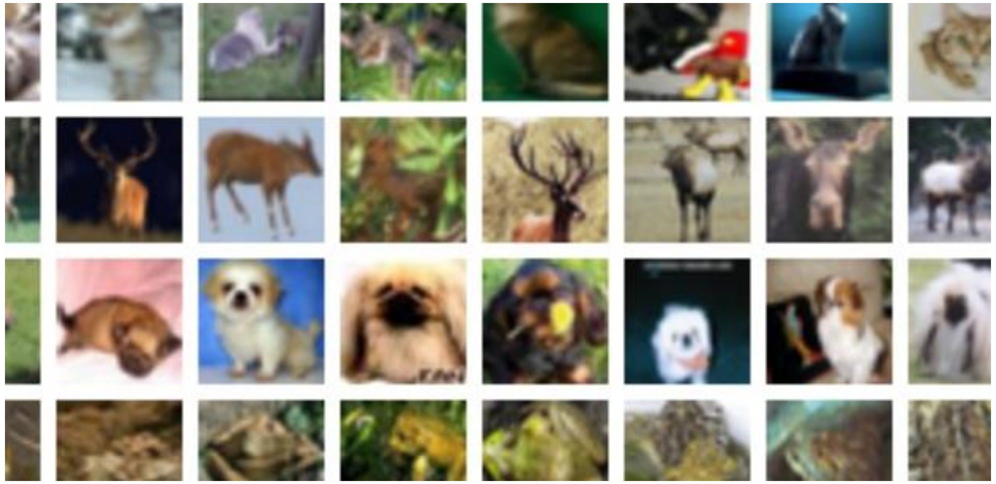


Image-level variation is high

Part-level variation is low

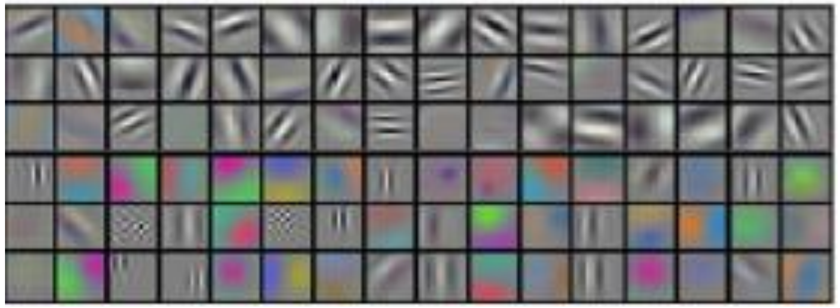
Image sources:

1. CIFAR 10 [CIFAR-10 image classification with Keras ConvNet - Giuseppe Bonaccorso]
2. Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks. *ECCV 2014*

Visualizing CNNs

- CNNs are cool ☺ but some of the below questions need answers before we move forward :-
- How do I interpret the learned filters?
- What is it that stimulates/excites a neuron?
- How do I decide the architecture or improve existing ones?

Source: Krizhevsky et.al. NIPS’12

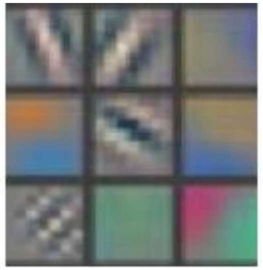


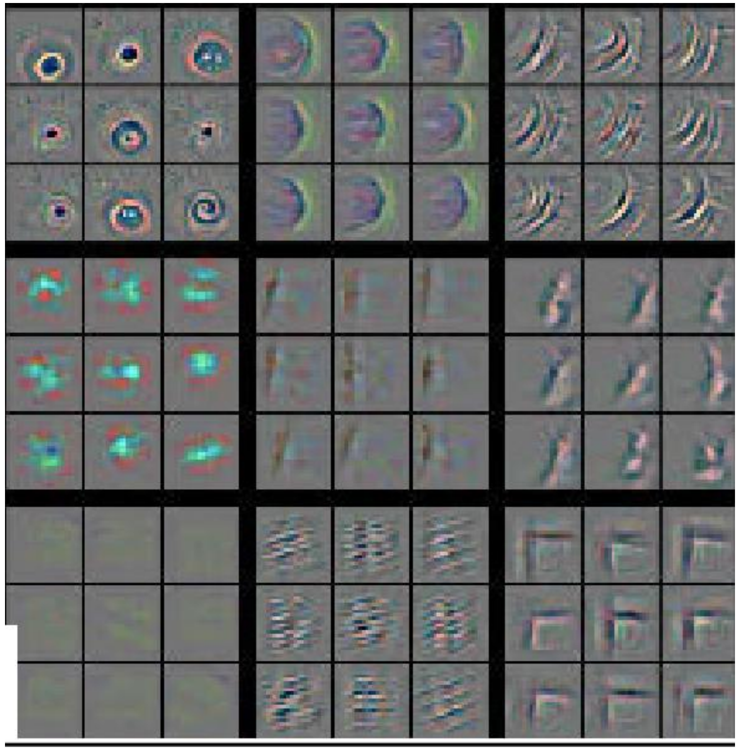
Visualizing the first conv. layer is possible but how about the later layers.

?

Zeiler and Fergus , Visualizing and Understanding Convolutional Networks,. ECCV 2014

Composition of filters





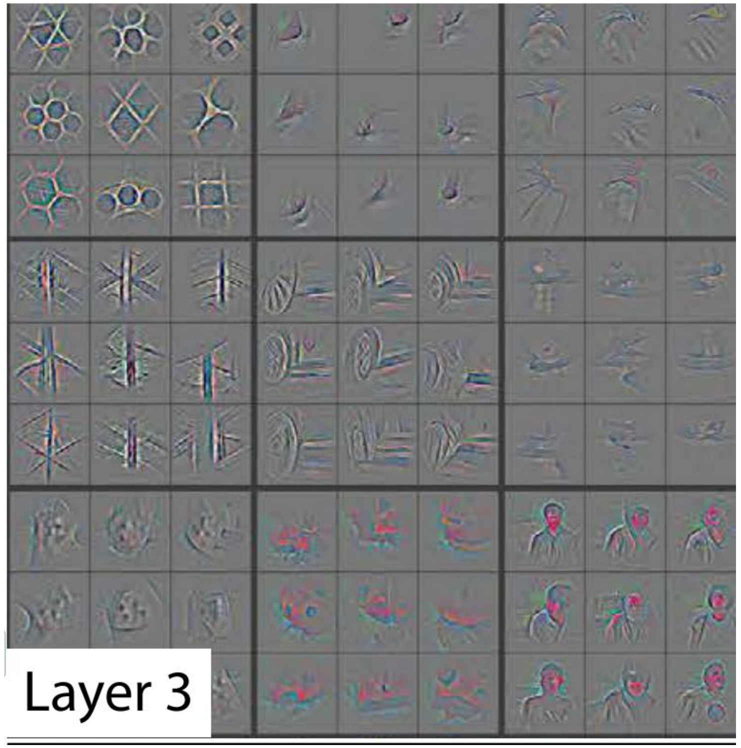
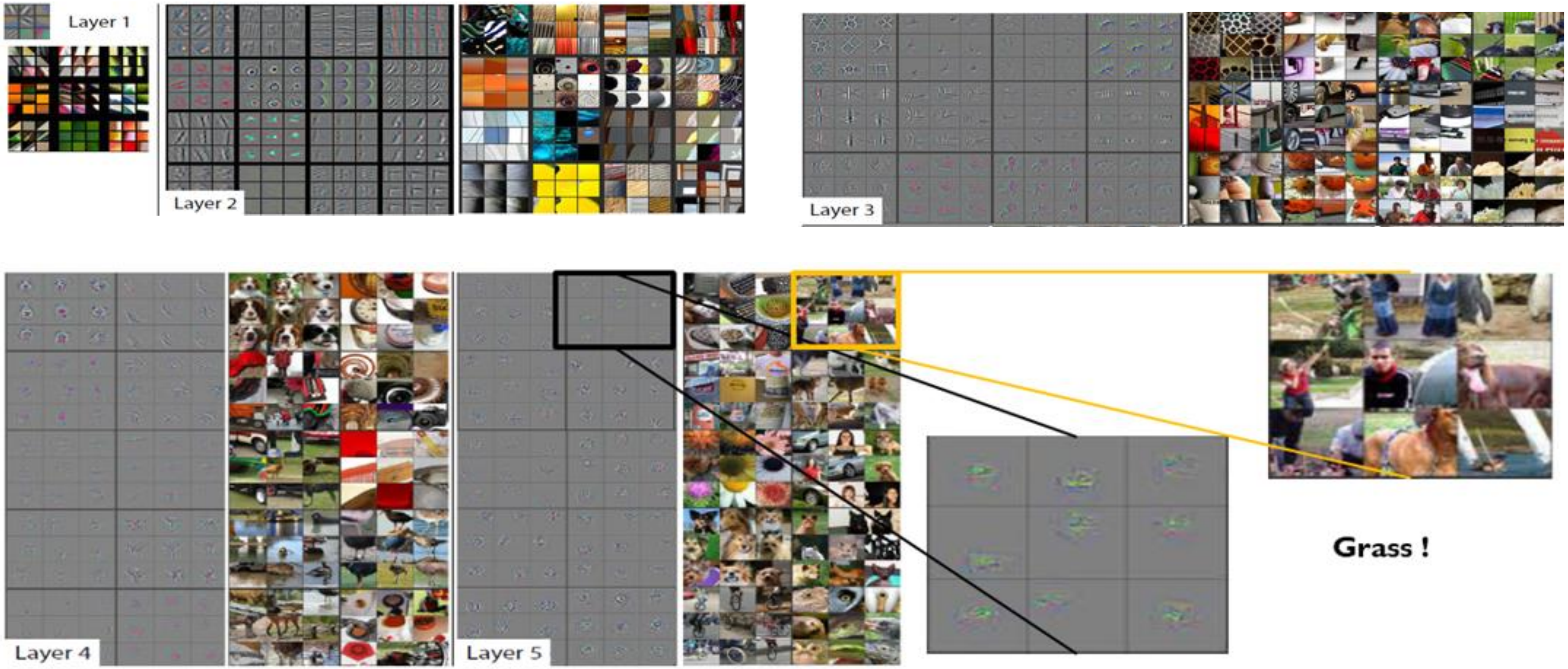


Image source:
Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks. *ECCV 2014*
These are top activations projected down to pixel space using deconvolution architecture

Visualizing CNNs

A. How do I interpret the learned filters?



Source: Zeiler e.t. al. ECCV'14

Zeiler and Fergus , Visualizing and Understanding Convolutional Networks,. ECCV 2014

Early Layers Converge Faster

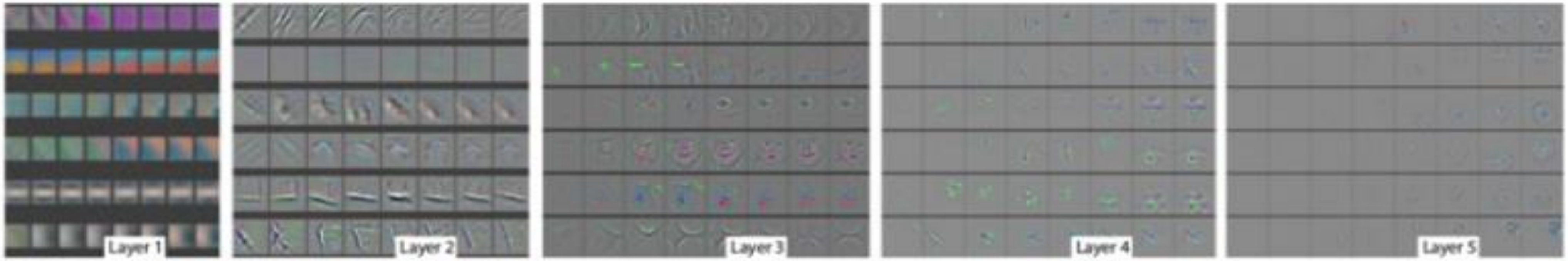


Figure: Evolution of randomly chosen subset of model features generated using deconvnet through training at epoch 1, 2, 5, 10, 20, 30, 40, 64.

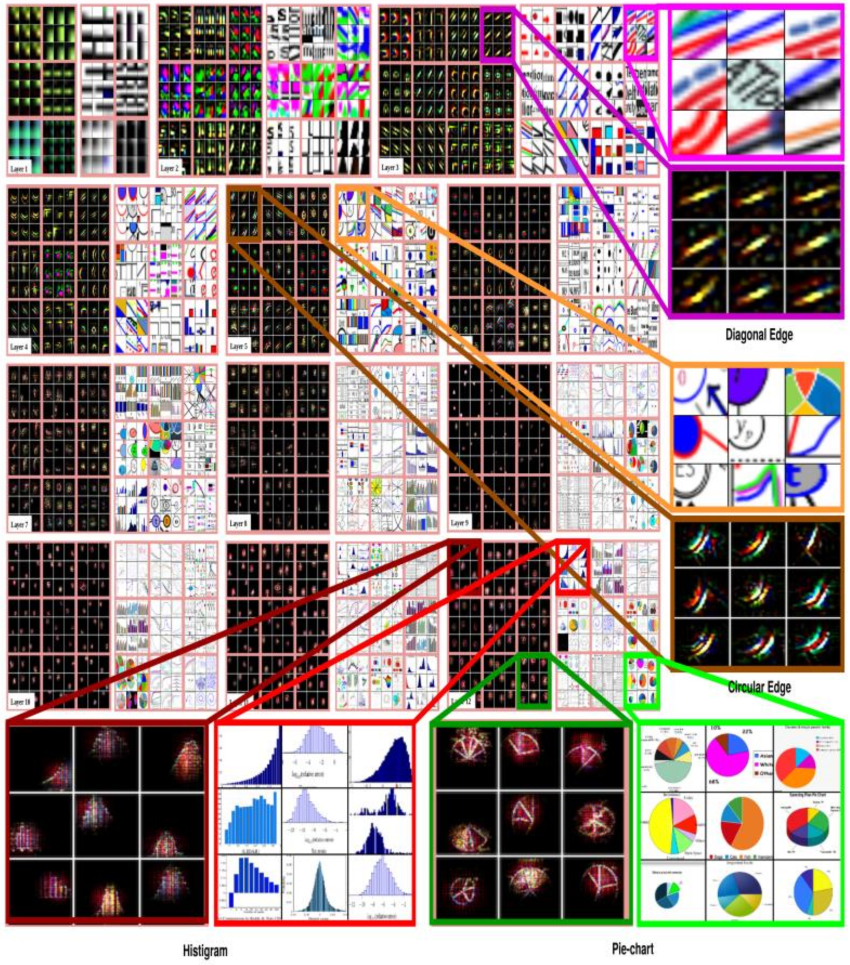
Example: Classification of Documents

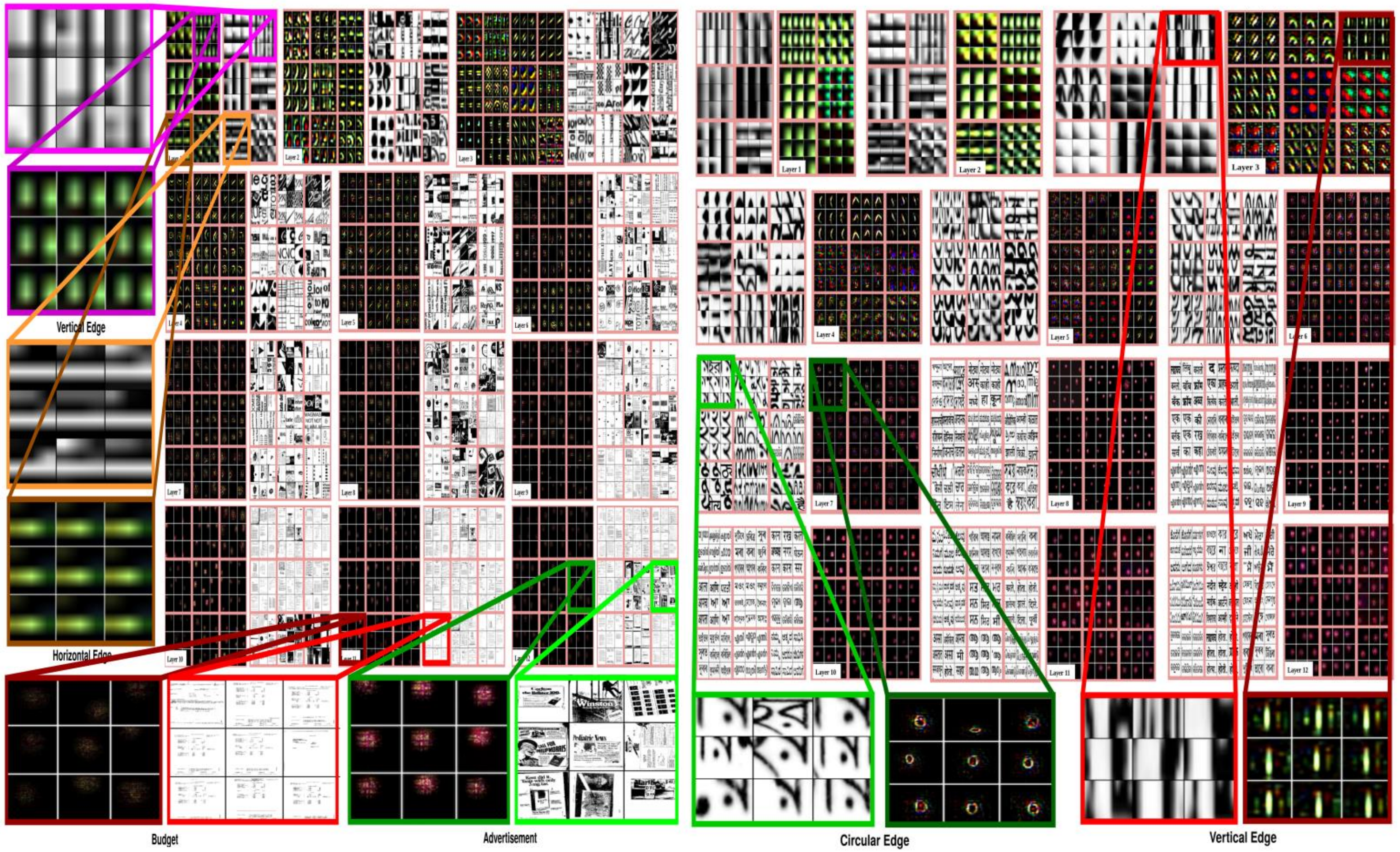
Task 1

Task 2

Task 3

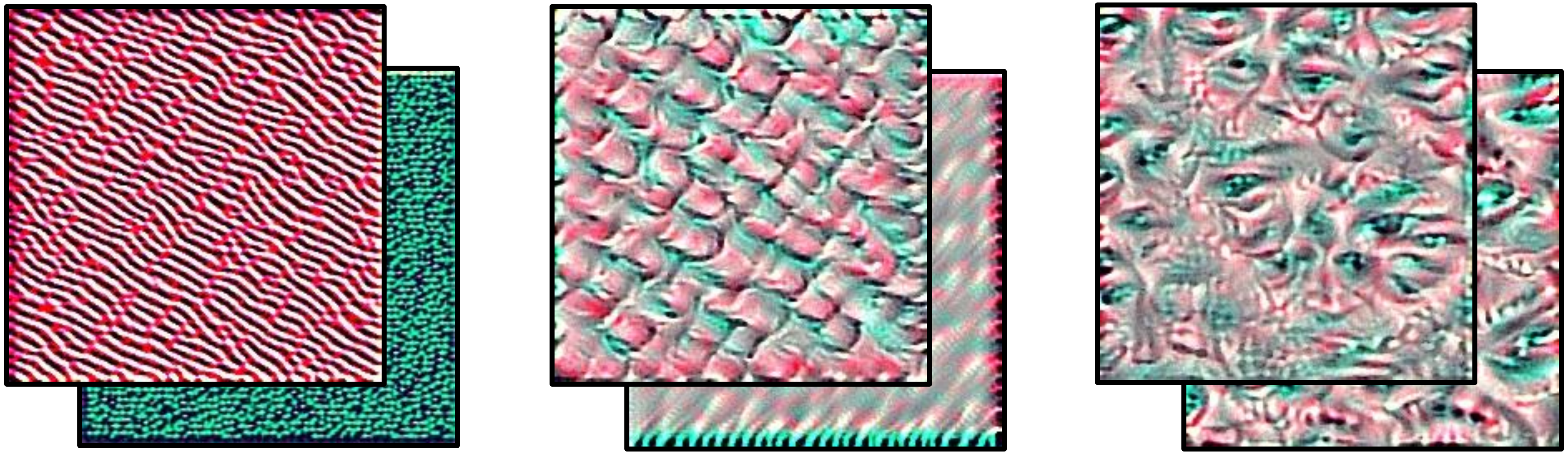
Examples


Diagonal Edge
Circular Edge
Histigram
Pie-chart


Vertical Edge
Horizontal Edge
Budget
Advertisement
Circular Edge
Vertical Edge

What does it look for faces?

- What type of image causes a convolutional filter to give a high activation?

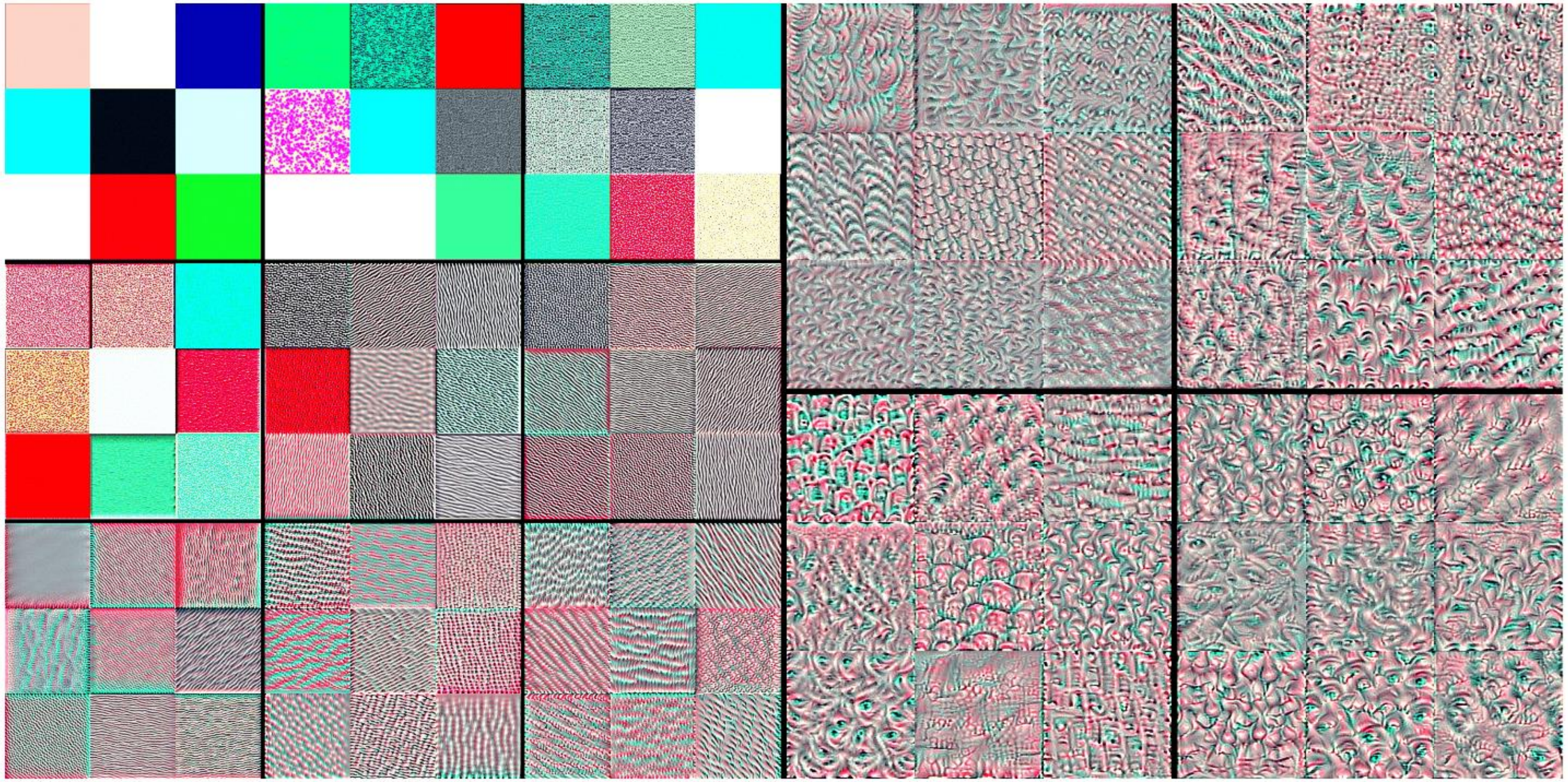


Low-level features

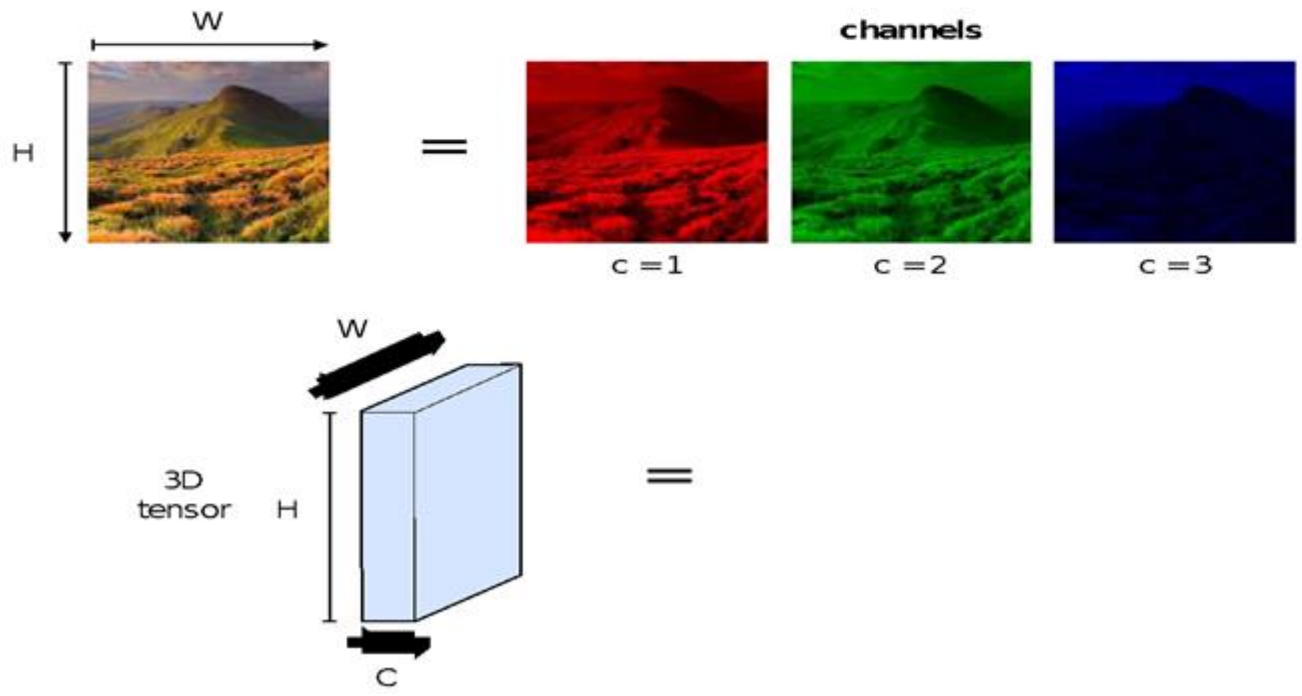
Mid-level features

High-level features

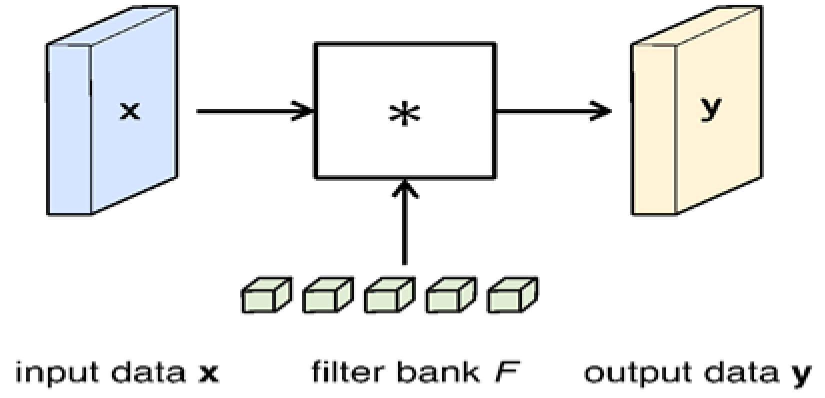
Results: face recognition

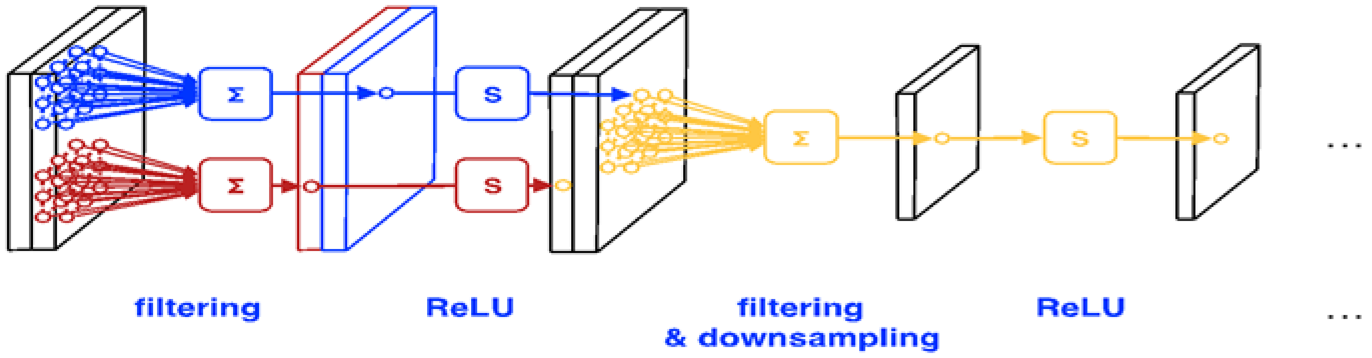


CNNs: Summary



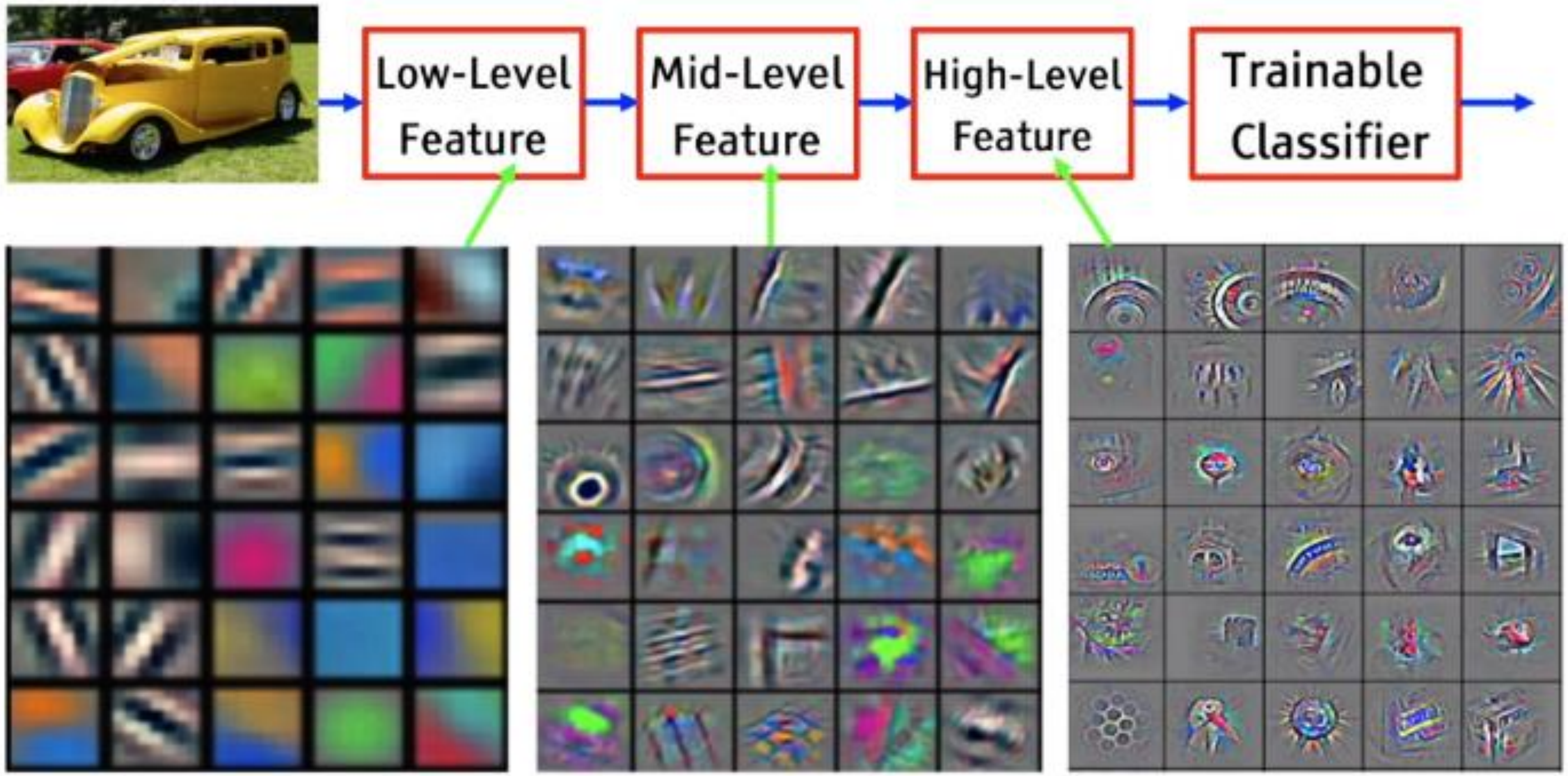
$$\mathbf{y} = F * \mathbf{x} + b$$





Deep Learnt Features

- It's deep if it has more than one stage of non-linear feature transformation.



Summary

Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- ✓ Deep Learning Architectures
- ✓ CNN
- ✓ CNN in Images, Speech and Text

Thanks!!

Questions?