Deep learning and Astrophysics

R. Moradi

ICRANet and ICRA–SAPIENZA University of Rome

F. RastegarNia

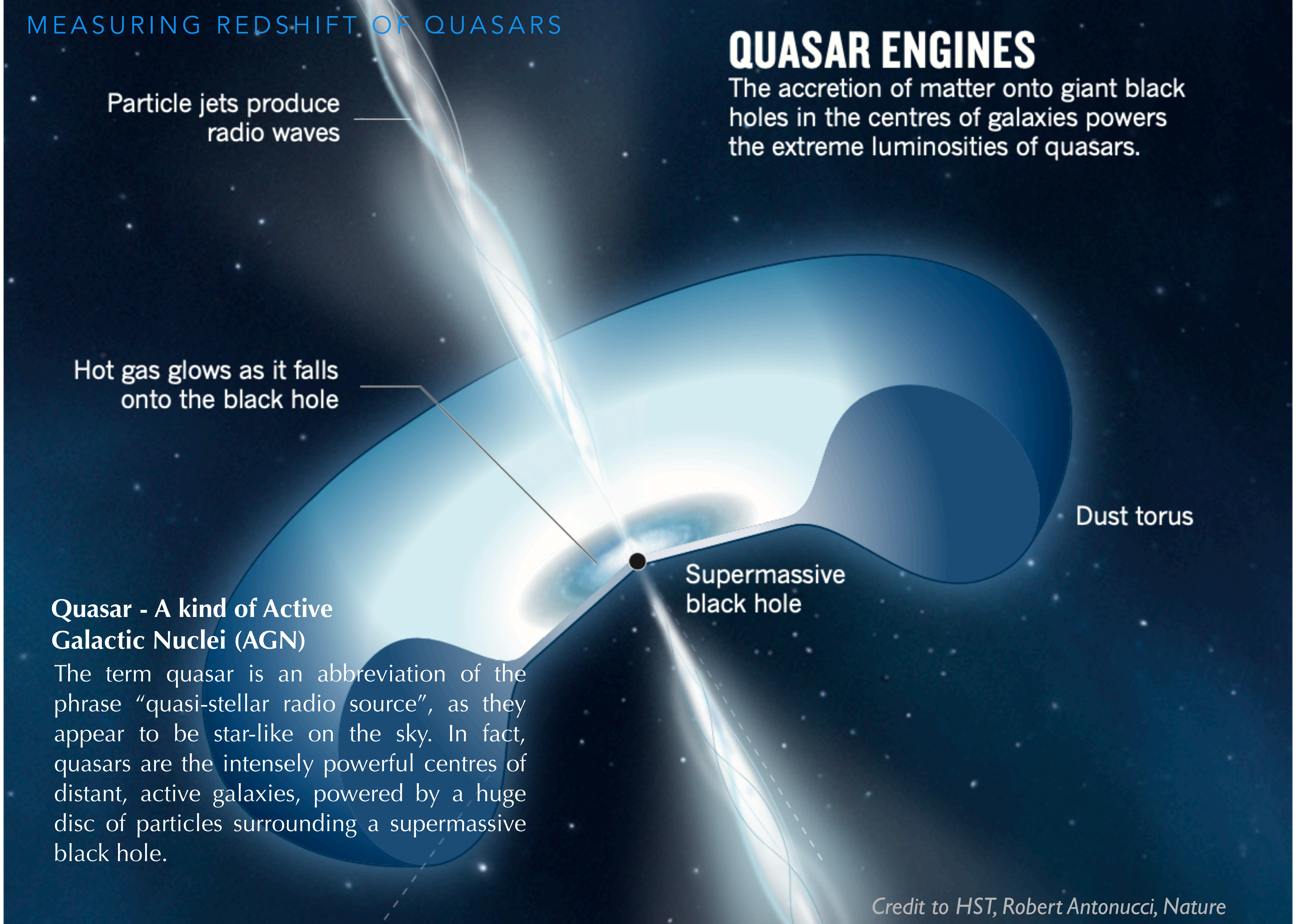
T. Mirtorabi

Yu. Wang

A. Vafaeisadr

Outline

* Quasars
* Deep Learning and Convolutional Neural Networks (CNN)
* Data; SDSS data of astrophysical objects
* How to download useful data
* Data preprocessing
* CNN architecture
* Results



Quasar - A kind of Active Galactic Nuclei (AGN)

The term quasar is an abbreviation of the phrase “quasi-stellar radio source”, as they appear to be star-like on the sky. In fact, quasars are the intensely powerful centres of distant, active galaxies, powered by a huge disc of particles surrounding a supermassive black hole.

Credit to HST, Robert Antonucci, Nature

Why we do this work?

1. ***Quasar is important:*** the most luminous non-transient source locating till redshift z>7, itself is important and it can be applied as a too to, for e.g.,

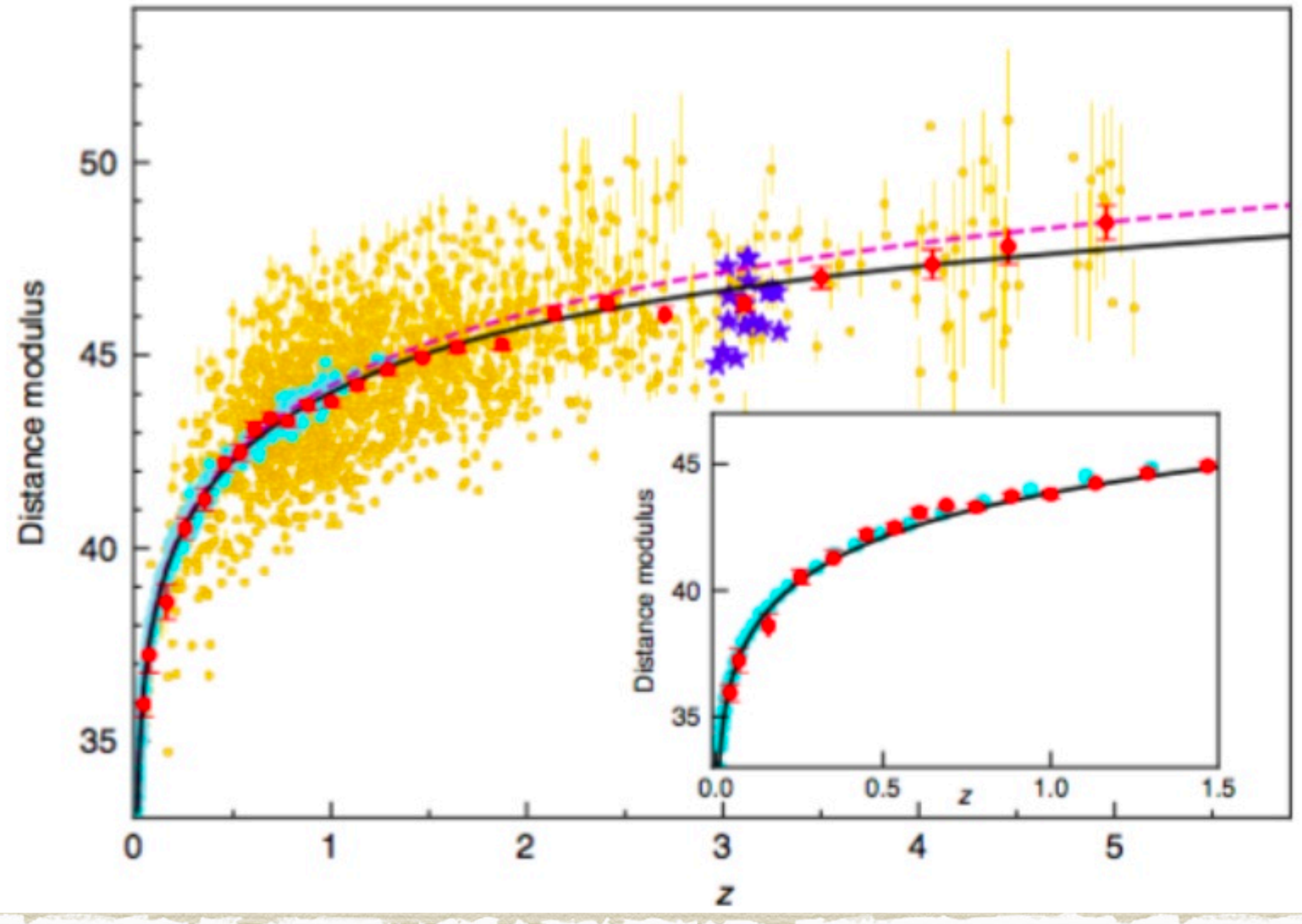
    - Constrain the cosmological parameters
    - Understand the formation of supermassive black hole
    - Explore the formation of galaxies
    - Trace the history of the reionization of early universe
    - Infer the intergalactic medium properties
    - …

2. ***Quasar has enormous data:*** one of the best surveyed cosmological objects, the total number closes to one million. For comparison, supernova has ~70,000 samples and gamma-ray burst has ~2000 samples.

Measurements of the expansion rate of the Universe based on a Hubble diagram of Quasars

- Quasars are the most luminous persistent sources in the Universe, observed up to redshifts of $z \approx 7.5$
- The ΛCDM model is poorly tested in the redshift interval between the farthest observed type Ia supernovae and the cosmic microwave background.
- The distance modulus/redshift relation of quasars at $z < 1.4$ is in agreement with that of supernovae
- A deviation from the ΛCDM model emerges at higher redshifts

G. Risaliti and E. Lusso: Nature Astronomy volume 3, pages272–277(2019)



High Redshift Quasar

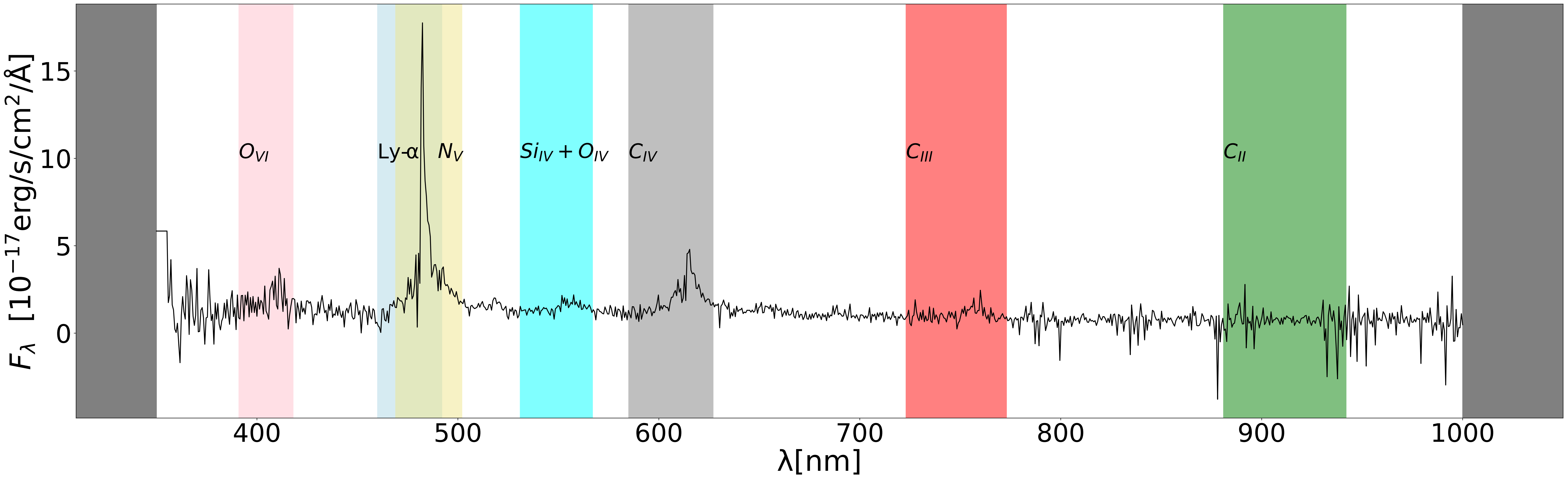
Quoted from Astro2020 Science White Paper:

"establishing large statistical sample of luminous quasars at z = 7 – 9, should be a high priority of high-redshift quasar community"

"Surveys of … quasars and AGNs at high redshift will allow studies of the earlier phases of the BH accretion process. These early phases … should be another high priority of high-redshift quasar research in the next decade."

— *Xiaohui Fan*

Xiaohui Fan, 2019


O_{VI}
Ly-α
N_V
$Si_{IV}+O_{IV}$
C_{IV}
C_{III}
C_{II}
F_λ [10^{-17}erg/s/cm^2/Å]
λ[nm]
0
5
10
15
400
500
600
700
800
900
1000

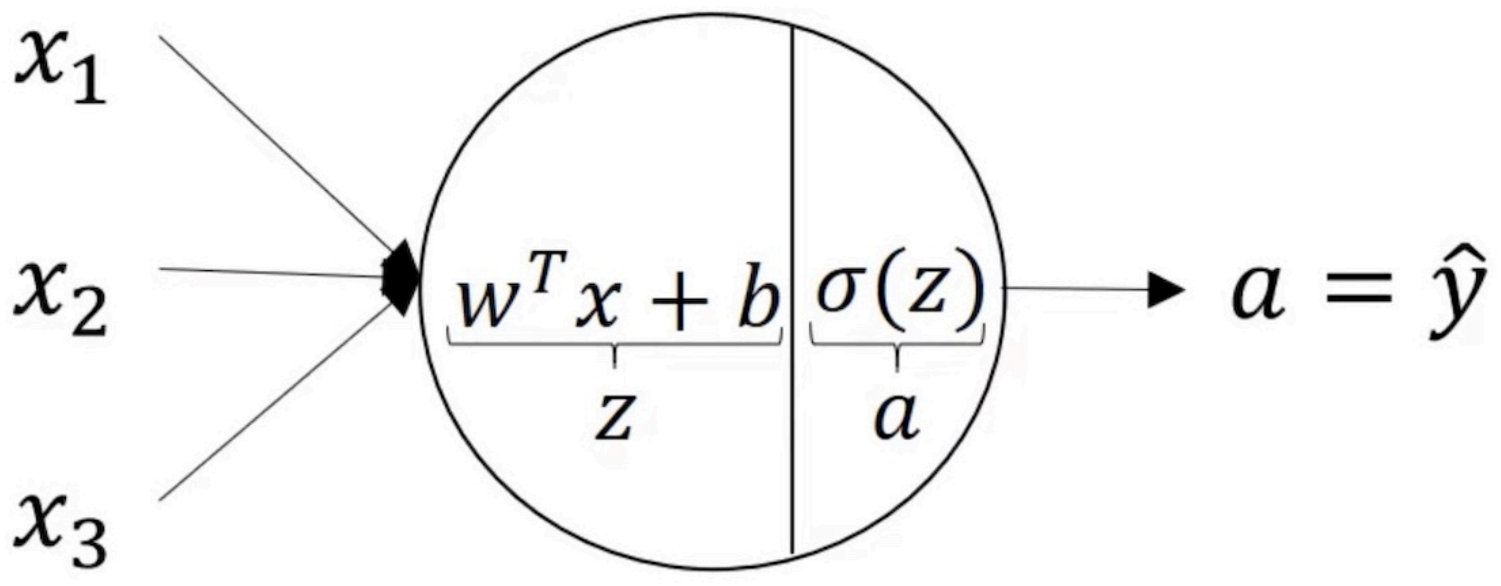
Purpose

Training a net to measure the redshift of quasars, especially it shall be able to measure the high redshift.

***Difficulty of traditional method**: Lyman-α lines, which is widely used to determine the high redshift quasars, are shifted to near infrared if z>6.4 and become undetectable . Hence, only 30% quasars at z~6.5 and 10% quasars at z>7 can be identified, and rare high redshift quasars have confirmed redshift, only 3 quasars with z>7 till 2020.*

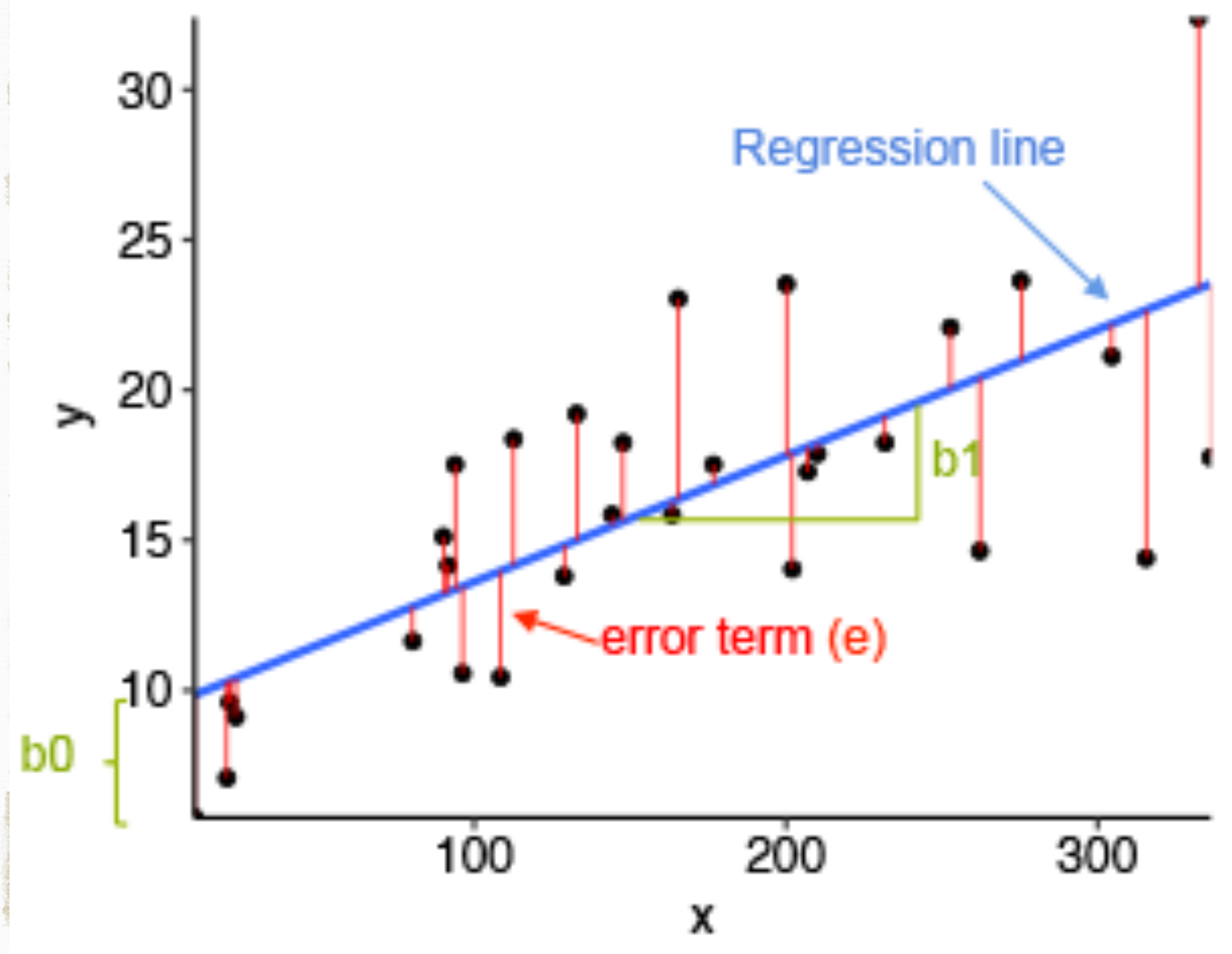
***Deep learning method**: The aim of the net is to seek and recognize, by the optimization procedure, all available common characteristics and patterns in data, which helps in turn to solve unseen problems and to measure the redshift.*

Basic Idea of Machine Learning

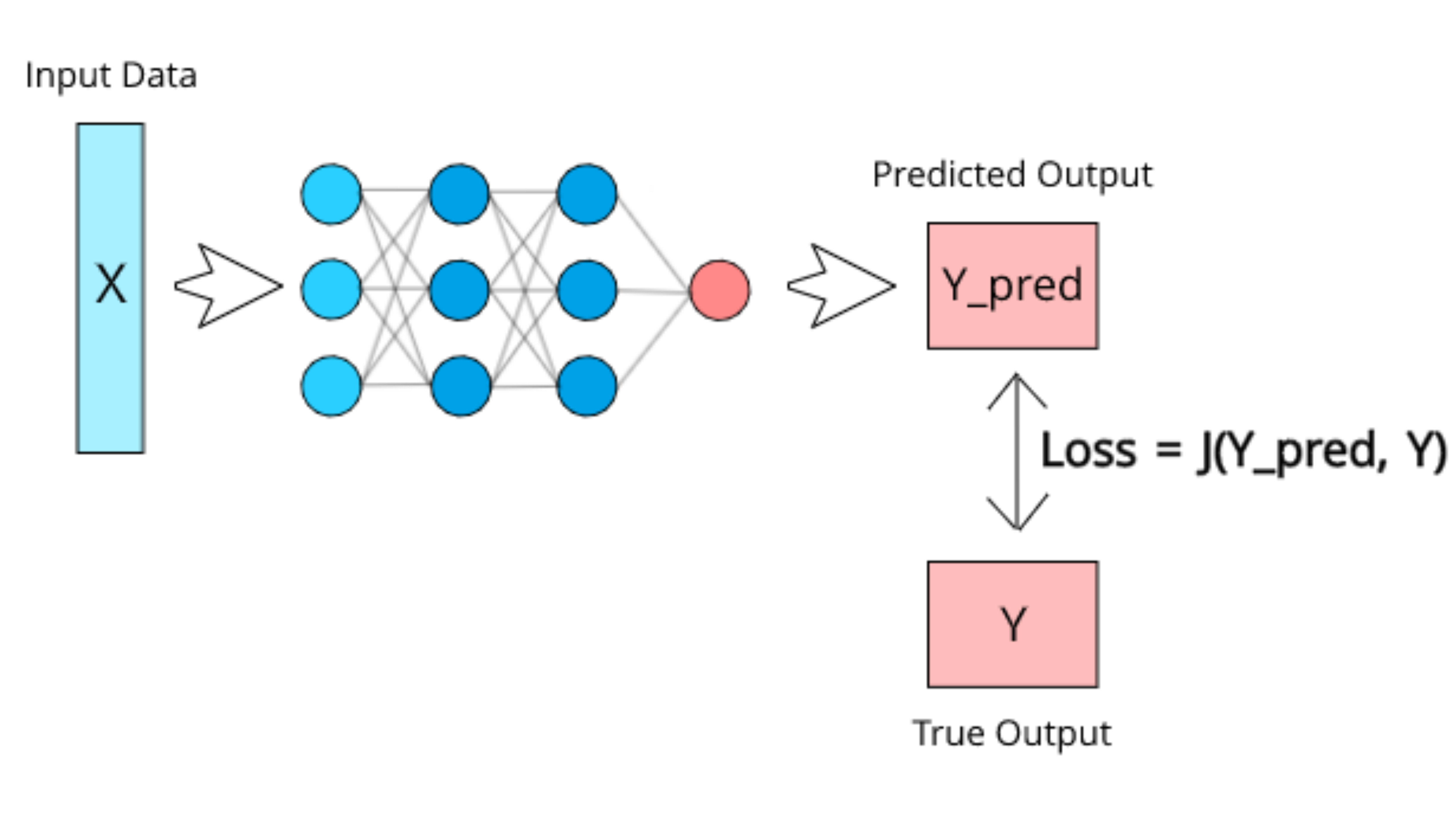


$$z = w^T x + b$$

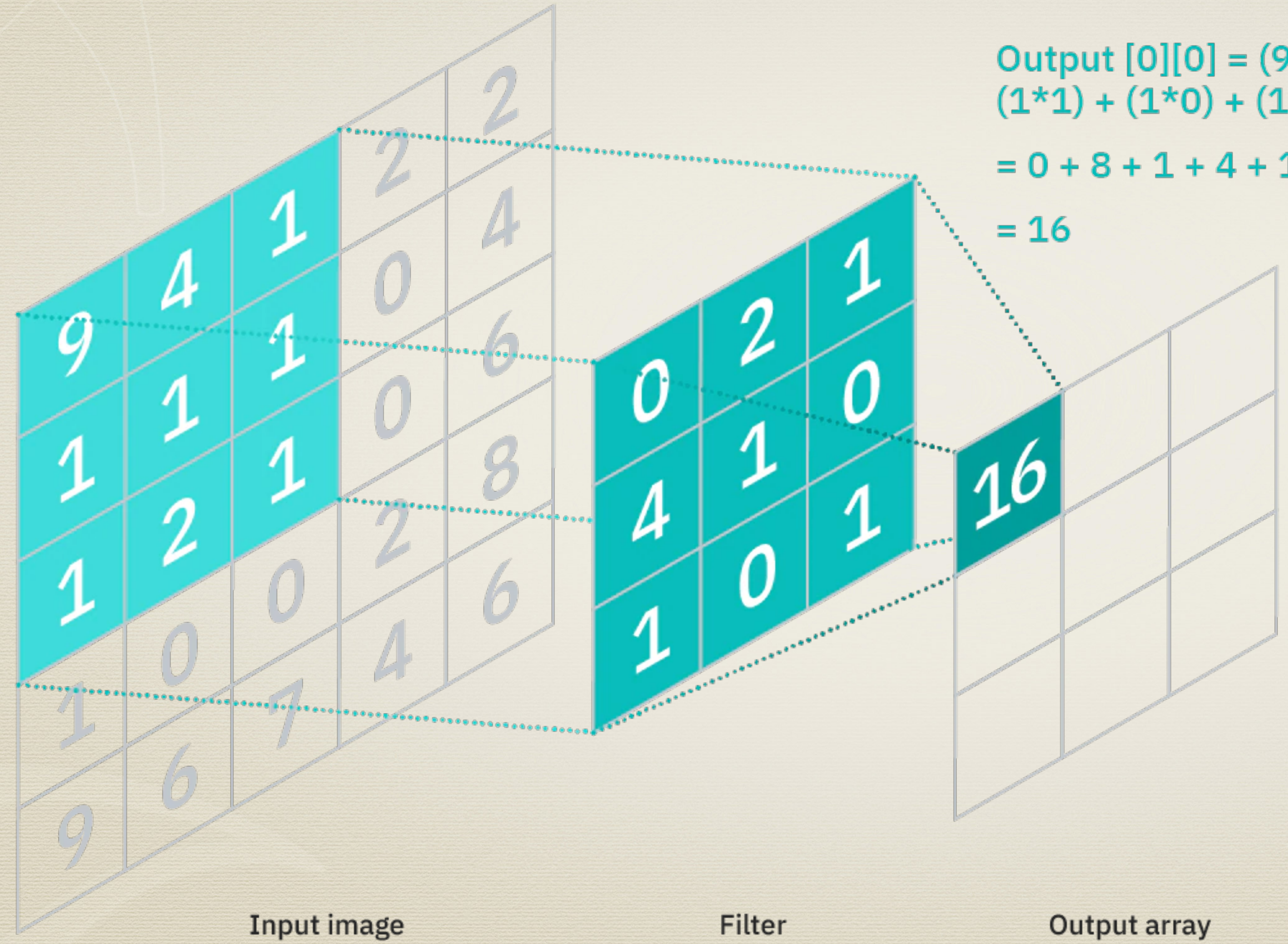
$$a = \sigma(z)$$



$$J = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y} - y)^2$$



Convolution?



Output [0][0] = (9*0) + (4*2) + (1*4) + (1*1) + (1*0) + (1*1) + (2*0) + (1*1)

= 0 + 8 + 1 + 4 + 1 + 0 + 1 + 0 + 1

= 16

Edge Finder!

10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0

6 x 6

*

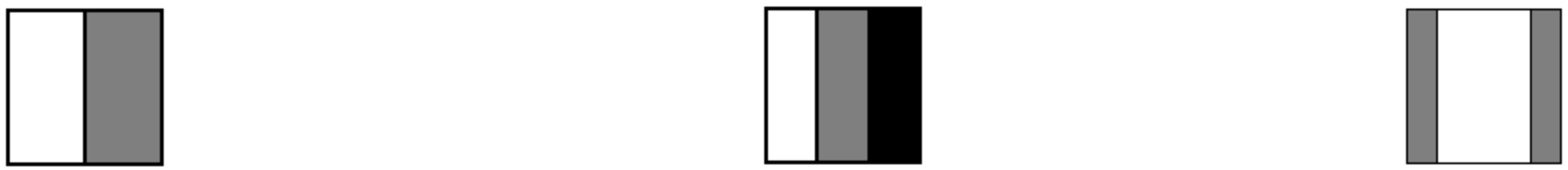
1	0	-1
1	0	-1
1	0	-1

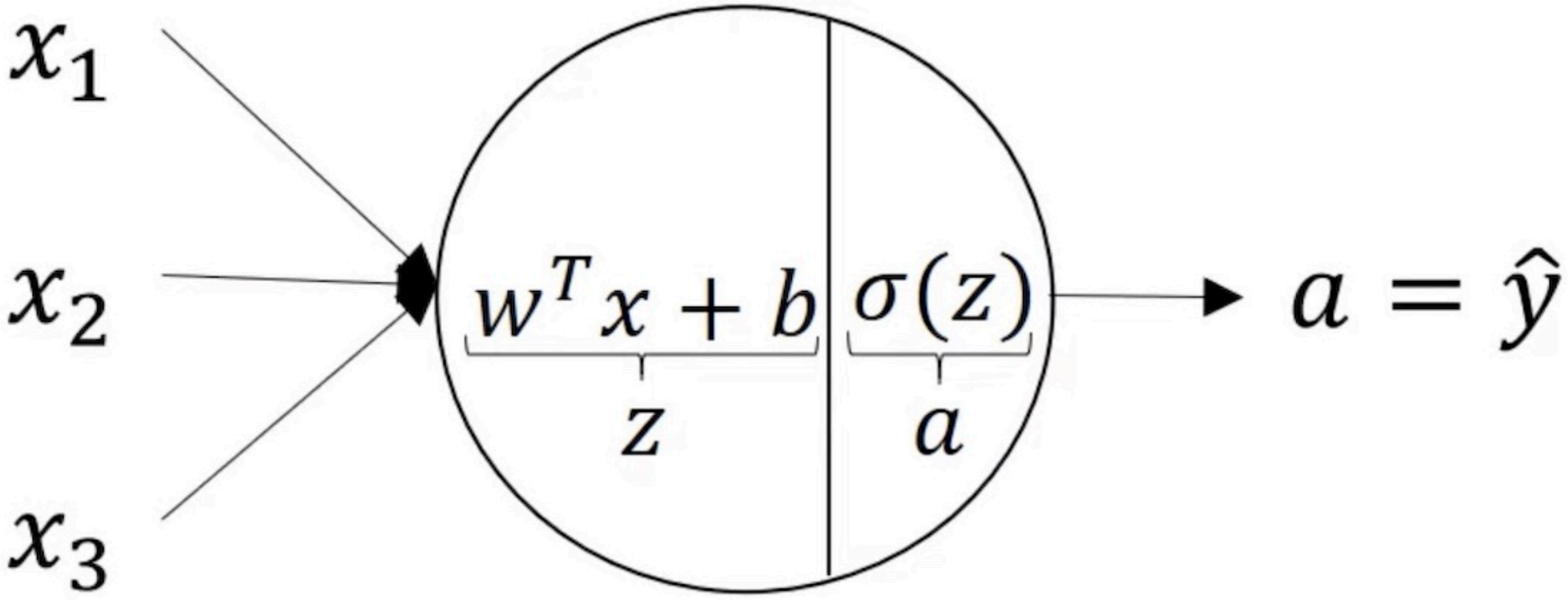
3 x 3

=

-0	30	30	0
0	30	30	0
0	30	30	0
0	30	30	0

4 x 4

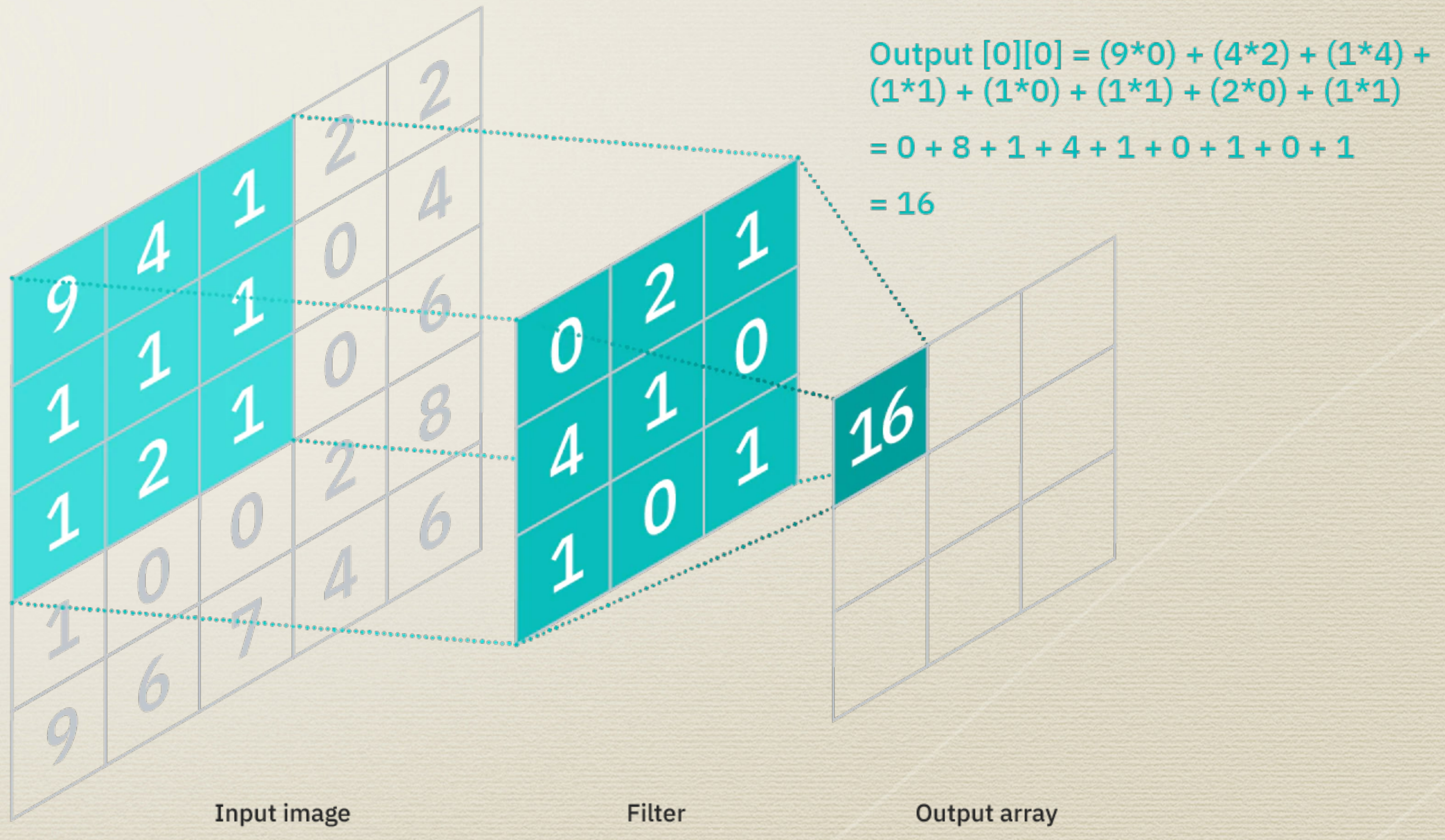


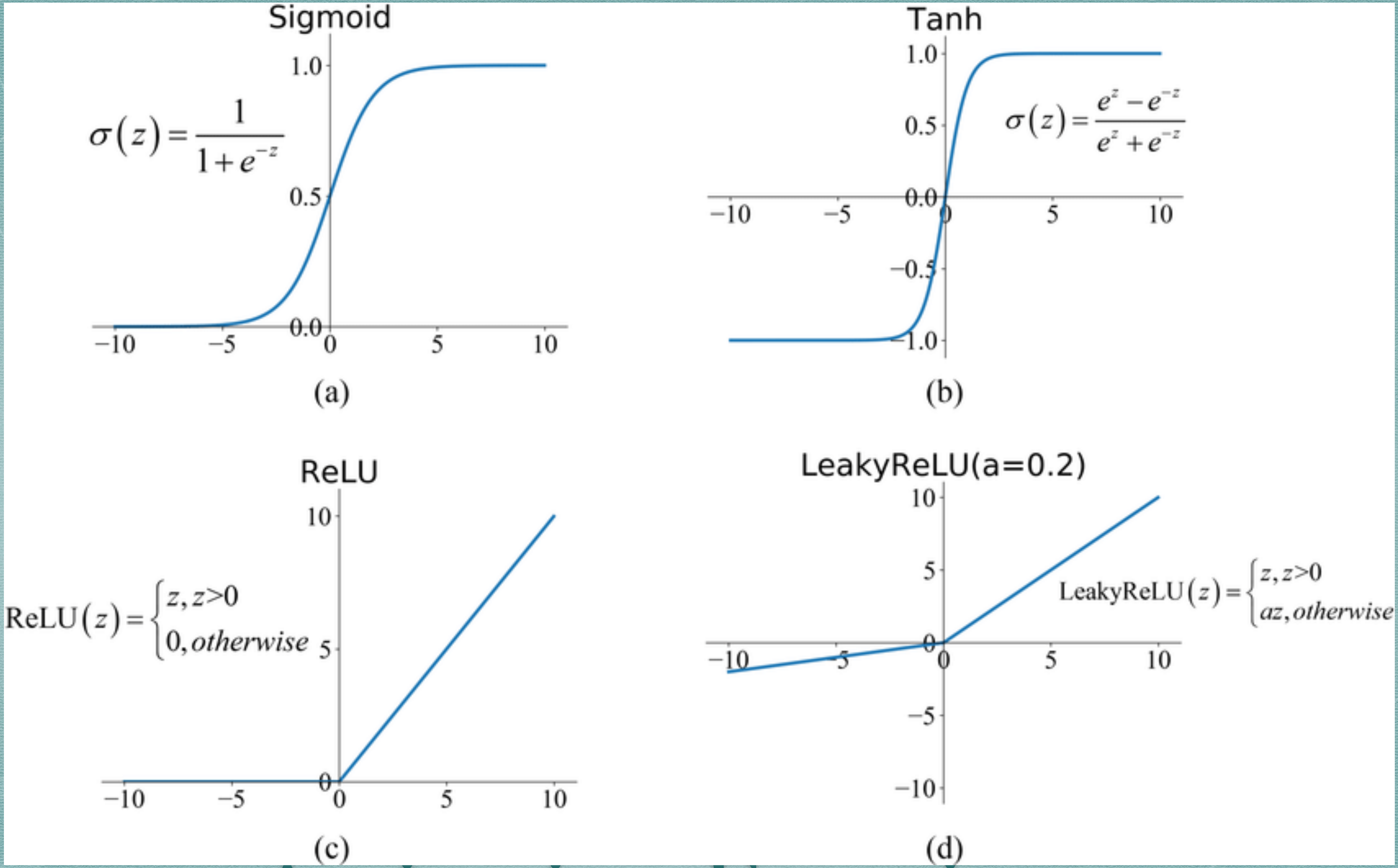


$$z = w^T x + b$$

$$a = \sigma(z)$$

$$\begin{bmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{bmatrix}$$

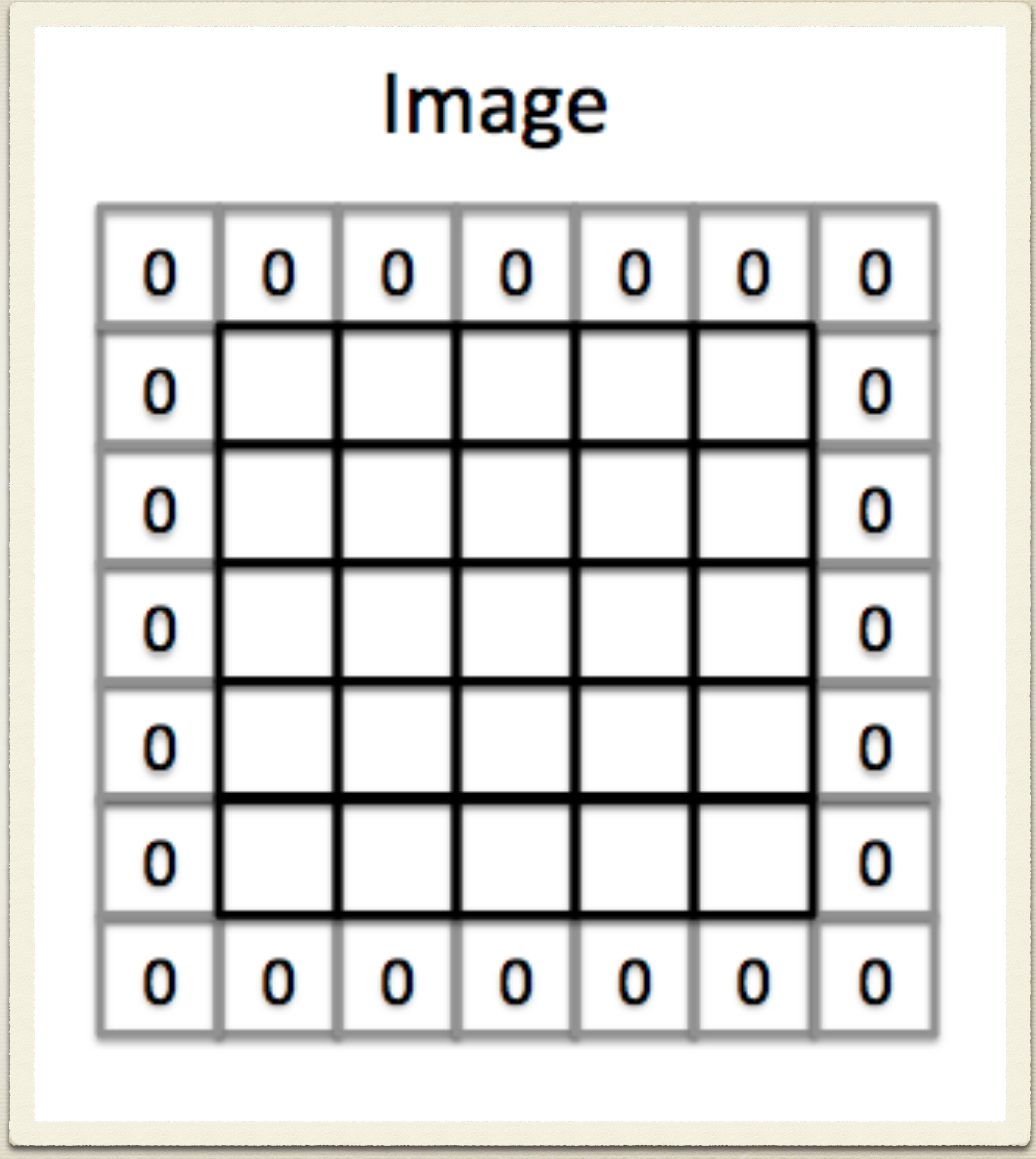




Activation Function

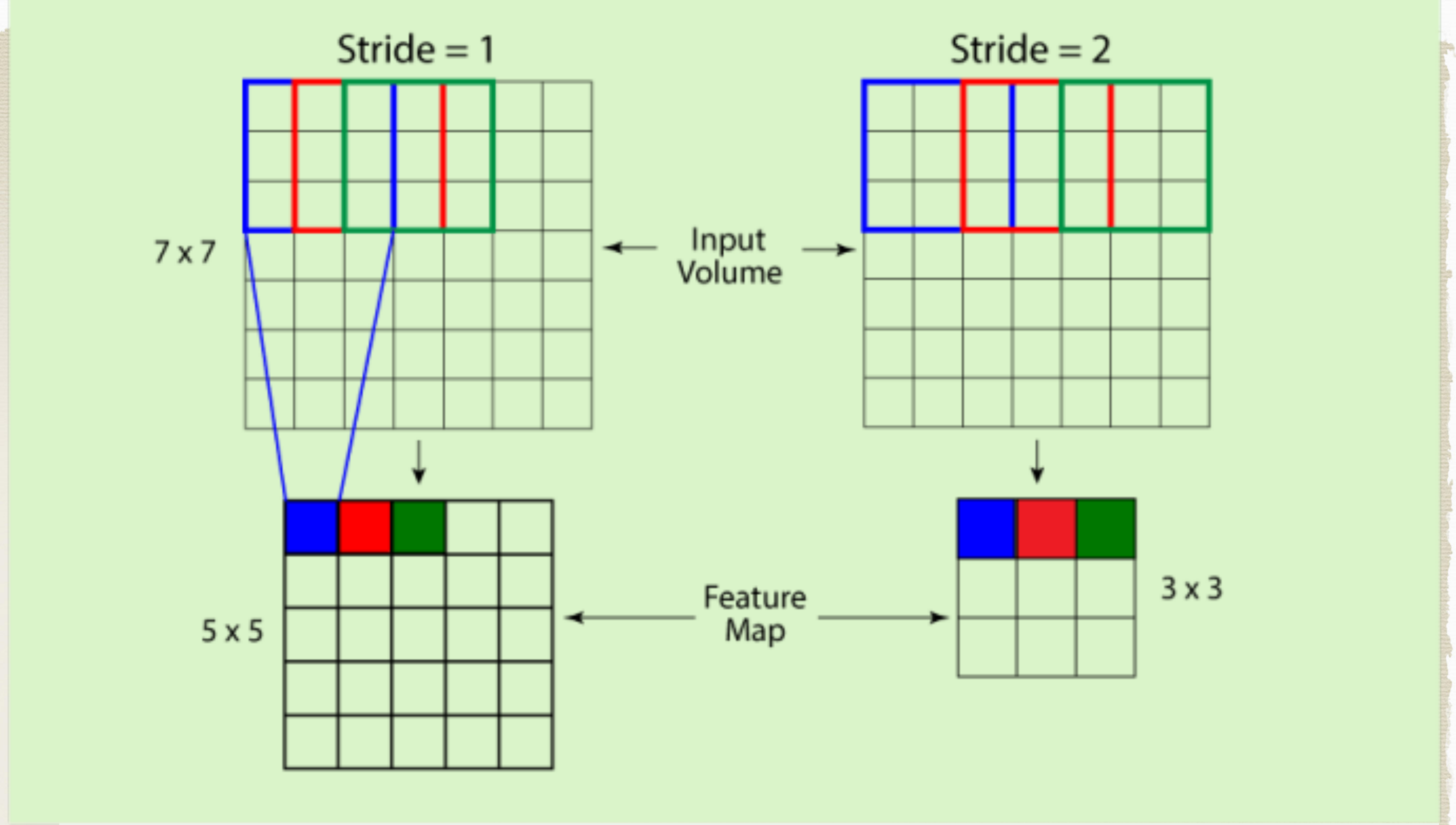
Padding

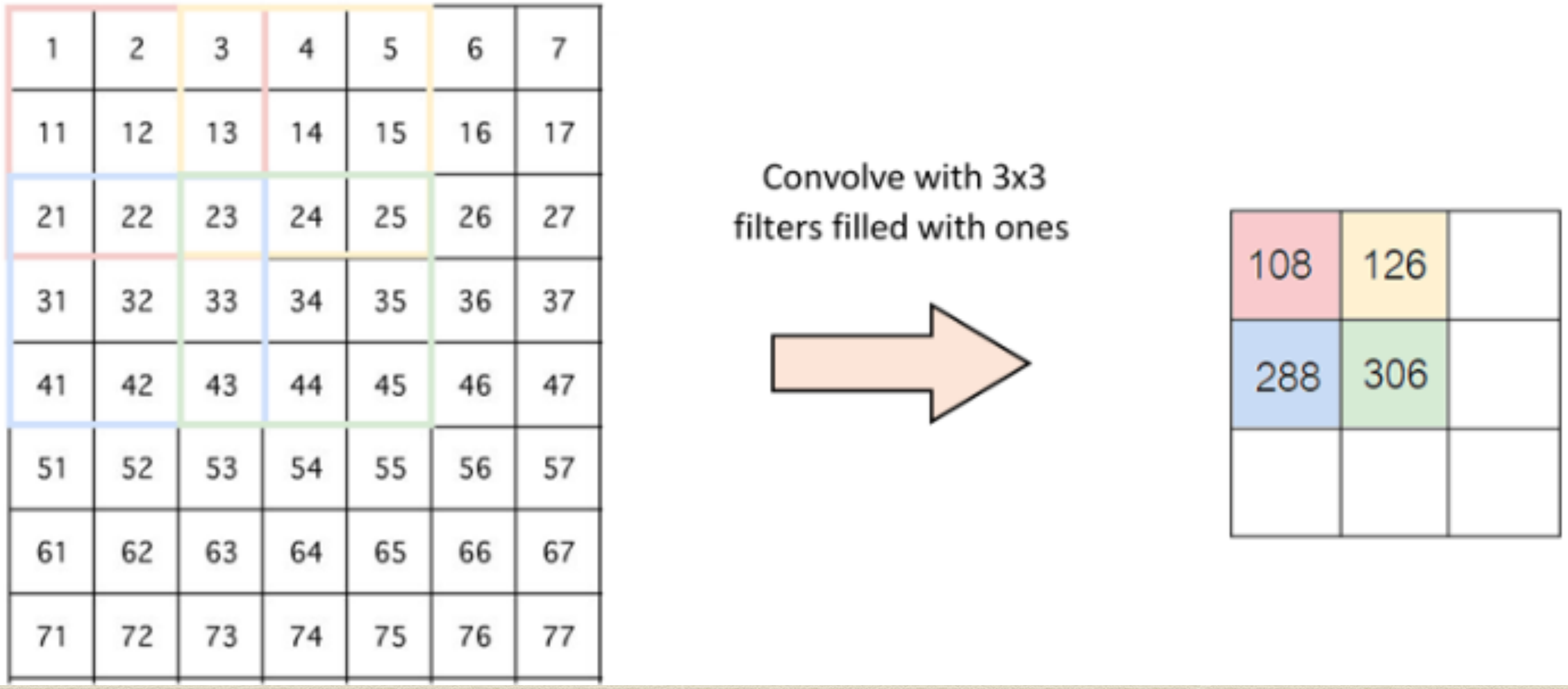
Information about borders



Strided Convolutions

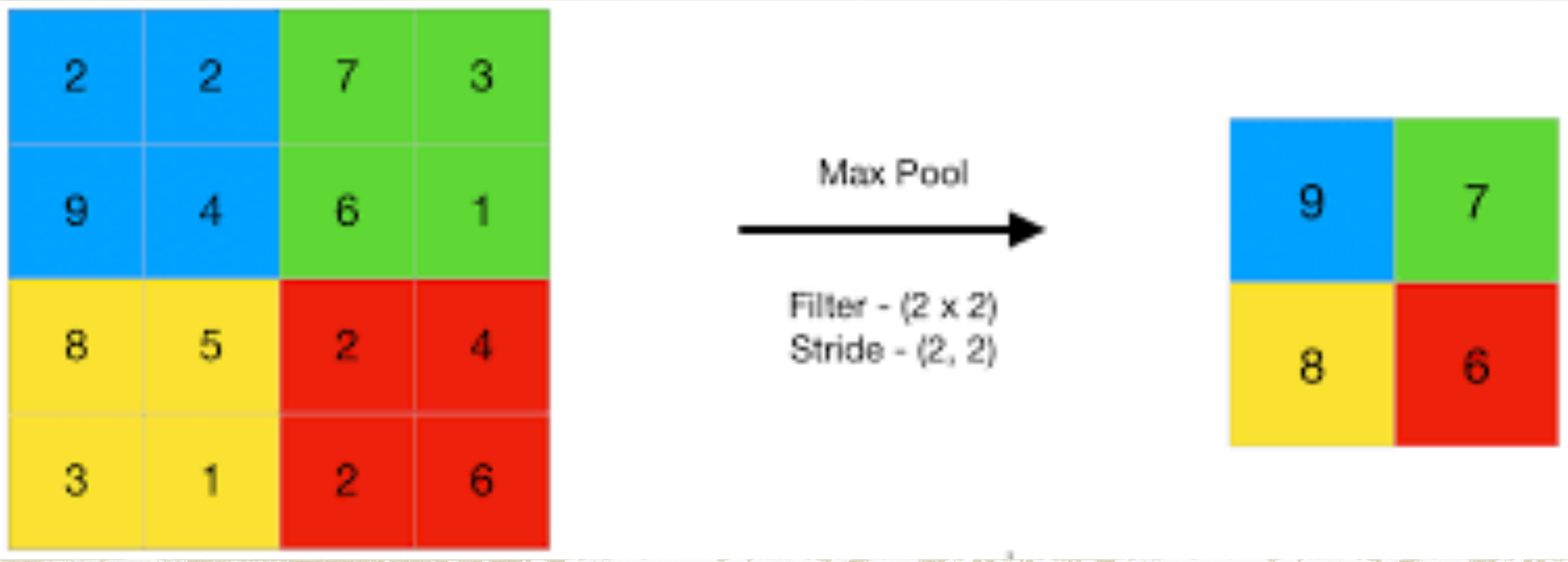
How deep one wants to find the patterns



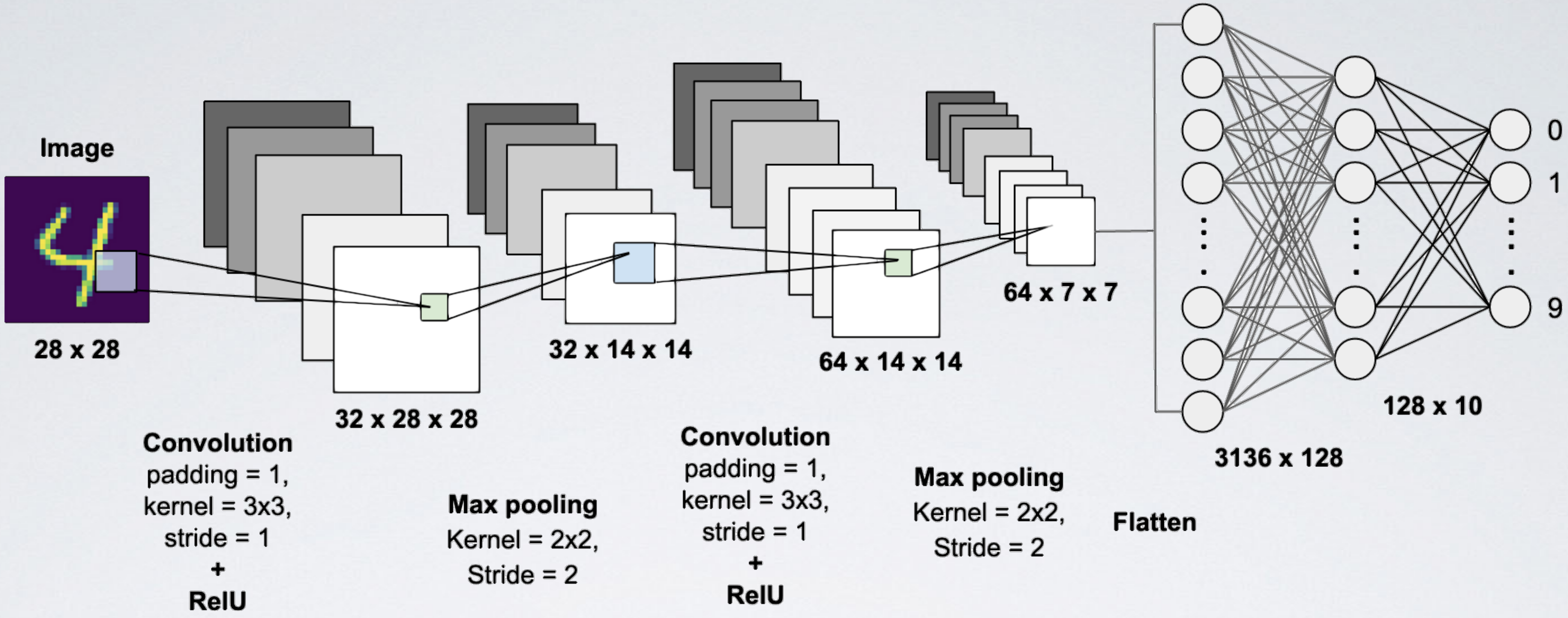


Pooling layers

Max
Average



Convolutional Neural Network




```
nn.Sequential( nn.Conv1d(1,32, 200), nn.ReLU(), nn.MaxPool1d(2, stride=2))
```

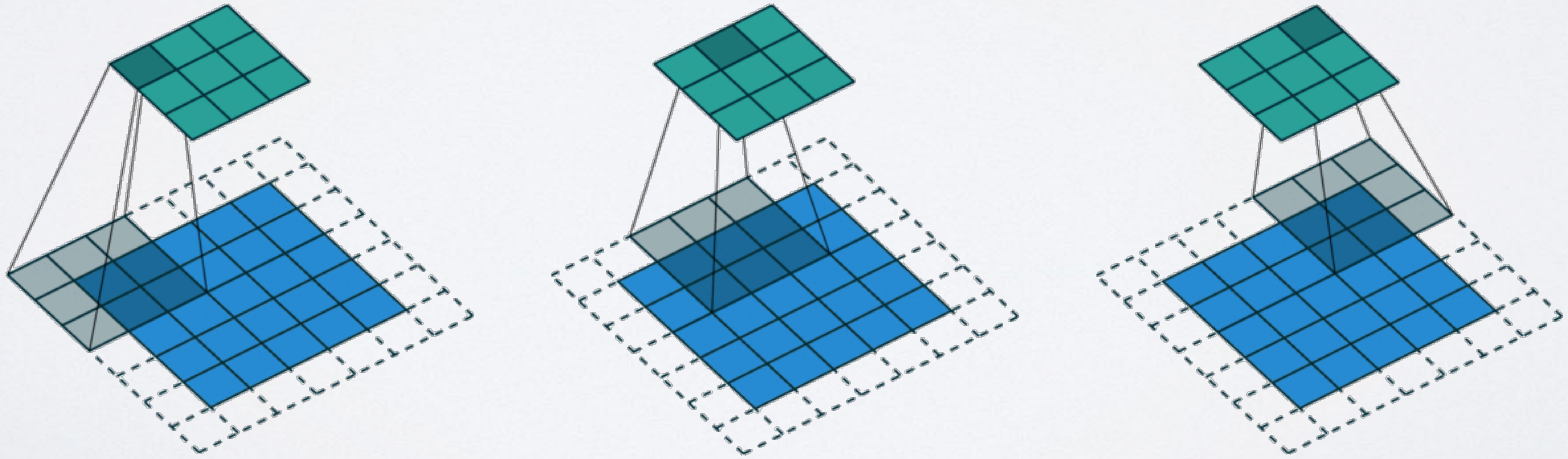
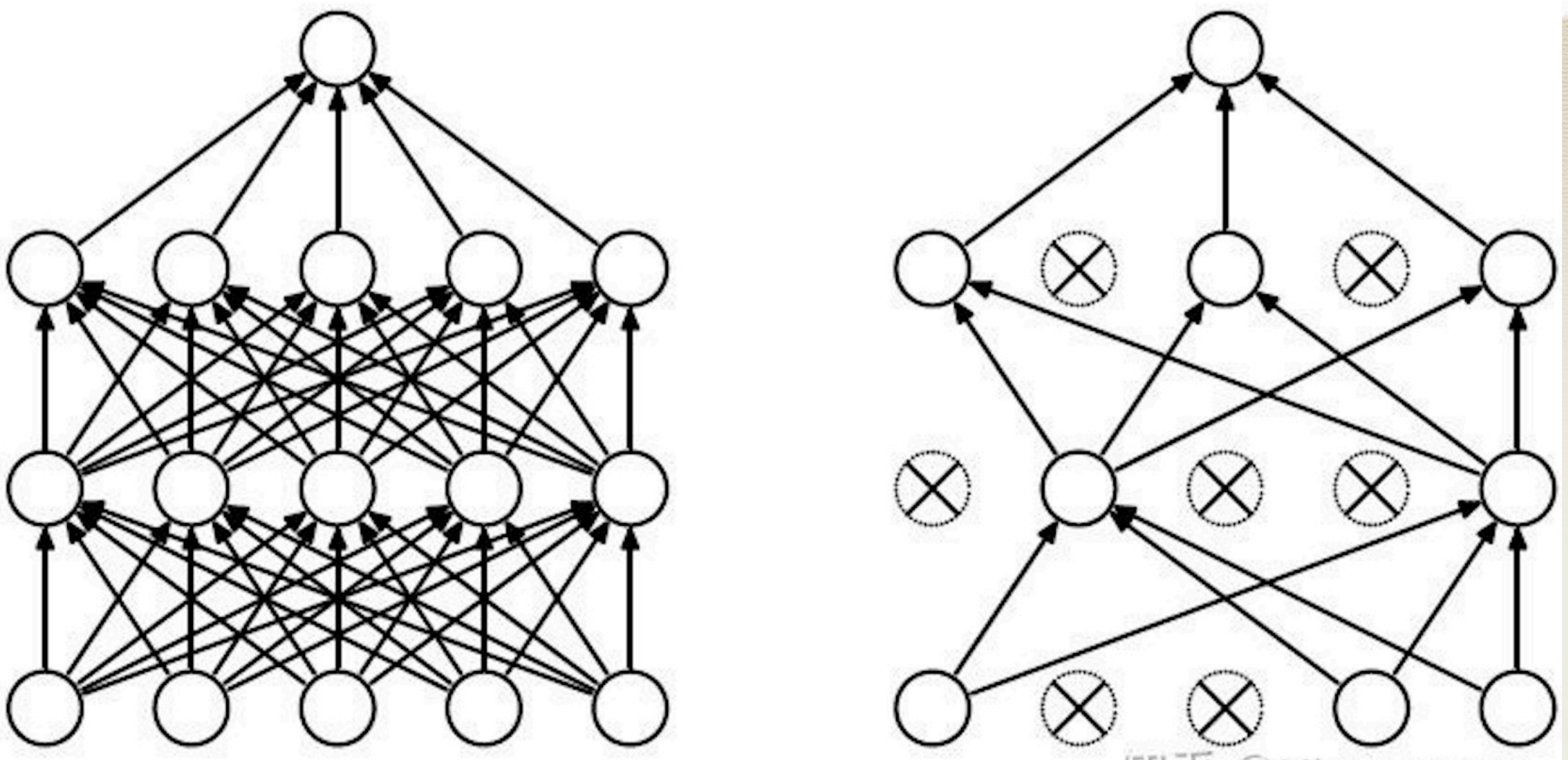
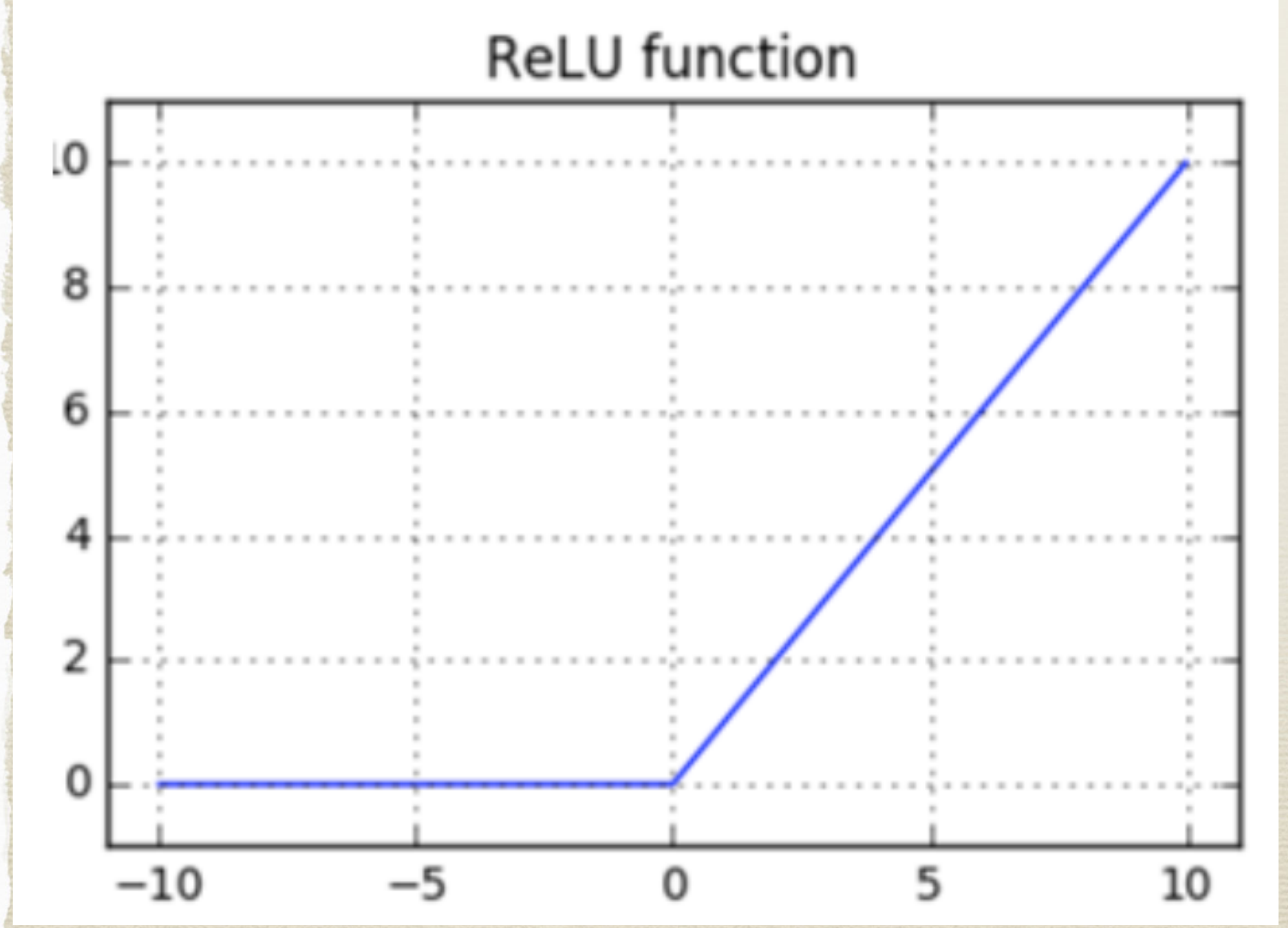


Image from towardsdatascience.com


知乎 @Microstrong

(a) Standard Neural Net (b) After applying dropout.




```python
class RNet(nn.Module):
    def __init__(self):
        super(RNet, self).__init__()
        self.layer1 = nn.Sequential(
            nn.Conv1d(1,32, 200),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.layer2 = nn.Sequential(
            nn.Conv1d(32,16,200),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.layer3 = nn.Sequential(
            nn.Conv1d(16, 128, 20),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.drop_out = nn.Dropout()
        self.fc1 = nn.Sequential(
            nn.Linear(8576, 900),
            nn.ReLU())
        self.fc2 = nn.Sequential(
            nn.Linear(900, 100),
            nn.ReLU())
        self.fc3 = nn.Linear(100, 1)

    def forward(self, x):
        x = x.unsqueeze(1)
        out = self.layer1(x)
        out = self.layer2(out)
        out = self.layer3(out)
        out = out.reshape(out.size(0), -1)
        out = self.drop_out(out)
        out = self.fc1(out)
        out = self.fc2(out)
        out = self.fc3(out)
        return out
```

SDSS

The Sloan Digital Sky Survey or SDSS is a major multi-spectral imaging and spectroscopic redshift survey
2.5-m wide-angle optical telescope
at Apache Point Observatory in New Mexico

eBOSS
Was designed to measure the expansion rate of the Universe.
Detect the characteristic scale imprinted by baryon acoustic oscillations in the early universe.

The most comprehensive observed quasar spectra to date is provided by the Sixteenth Data Release Quasar-only (DR16Q) of SDSS extended Baryon Oscillation Spectroscopic Survey (eBOSS).

This catalog contains 750,414 quasars in the automated redshift range $0 \leq z \leq 7.1$.

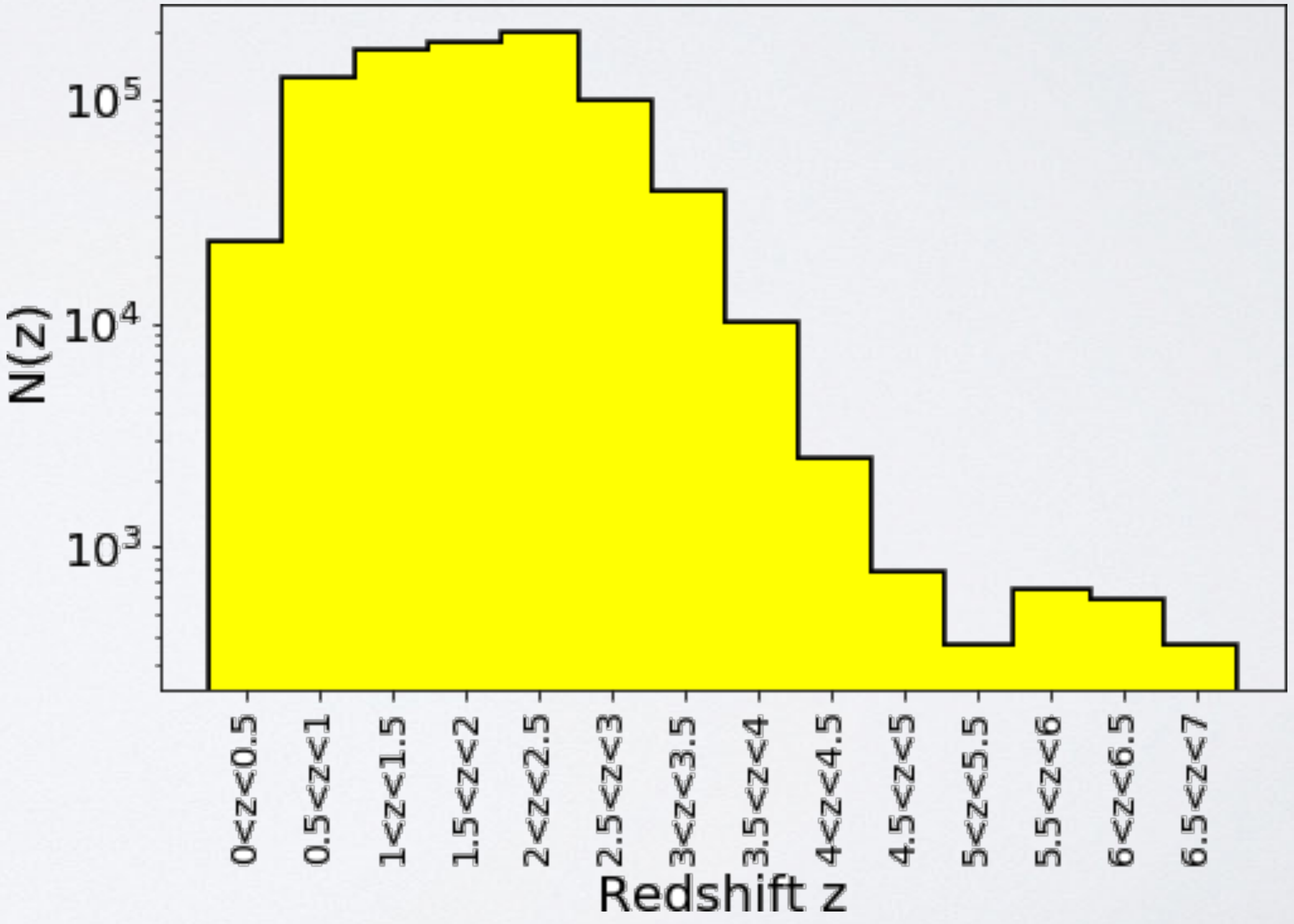
Sample - Feature & Label

Statistics: *Dataset from Sixteenth Data Release (DR16) of SDSS extended Baryon Oscillation Spectroscopic Survey (eBOSS). More than 700,000 quasar spectra have been detected in the redshift range 0 < z < 7.1.*

We consider the selection of quasars in range 2.8-3, which contains 11227 spectra.

Feature: *eBOSS spectrograph covers the wavelength range from 361nm to 1014nm with the resolution R = 2000. We take the spectrum as one dimensional image of 2000 pixels (flux).*

Label: *The identified redshift by traditional method.*



Network Design

Convolutional Neural Network (CNN) has been widely adopted to recognize patterns from an image. Therefore, here, 1-dimensional CNN is selected to learn higher-order features from the input flux via convolutional layers then to make a prediction on redshift as a output layer.

Thoughts: To design a purposive network, we follow the thought that, based on the spectrum, redshift can be estimated by the combination of three patterns:

1. The shift of the emission and absorption lines at different redshifts.
2. Some specific signals may appear at given redshfits.
3. Some features which we cannot understand!

Design: In the convolutional part, we specially construct a large size filter of 200 pixels covering ~ 20% of the data to capture the global shift of the spectrum, and in series a small size filter of 20 pixels to capture the minor shift and those specific signals.

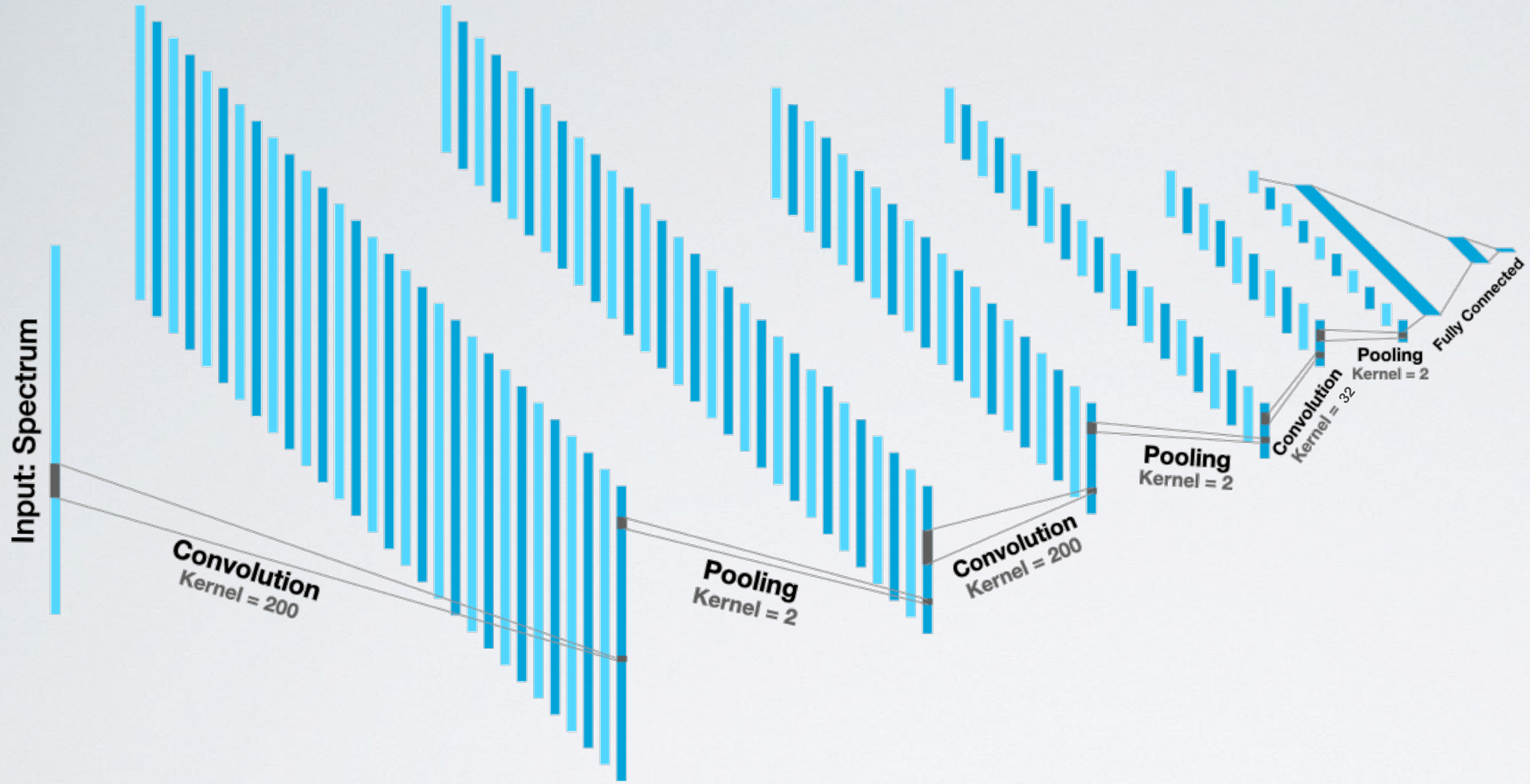
Redshift Caused by Universe Expansion

Observation on Earth

YU WANG

MEASURING REDSHIFT OF QUASARS

One Dimensional Convolutional Neural Network



Structure of one dimensional CNN. The spectrum of quasar is input as a one-dimensional array, which goes through the convolutional layer of kernel size = 200, 200, 20 respectively in order to search for the global and local pattern. The fully connected layers output the redshift.

```
class RNet(nn.Module):
    def __init__(self):
        super(RNet, self).__init__()
        self.layer1 = nn.Sequential(
            nn.Conv1d(1,32, 200),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.layer2 = nn.Sequential(
            nn.Conv1d(32,16,200),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.layer3 = nn.Sequential(
            nn.Conv1d(16, 128, 20),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.drop_out = nn.Dropout()
        self.fc1 = nn.Sequential(
            nn.Linear(8576, 900),
            nn.ReLU())
        self.fc2 = nn.Sequential(
            nn.Linear(900, 100),
            nn.ReLU())
        self.fc3 = nn.Linear(100, 1)

    def forward(self, x):
        x = x.unsqueeze(1)
        out = self.layer1(x)
        out = self.layer2(out)
        out = self.layer3(out)
        out = out.reshape(out.size(0), -1)
        out = self.drop_out(out)
        out = self.fc1(out)
        out = self.fc2(out)
        out = self.fc3(out)
        return out
```

YU WANG

Automated redshift estimates

- Z_VI, that is the redshift of visually inspected estimation the most accurate one, but it is exteremely time consuming process.
- Z_PIPE,The SDSS pipeline redshift estimate, is the result of a principal component analysis performed on a sample of visually-inspected quasars.
- Z_PCA, is also the result of a principal component analysis but, unlike Z_PIPE, the reference sample has been carefully chosen to have an automated redshift corresponding to the location of the maximum of the Mg ii emission line.

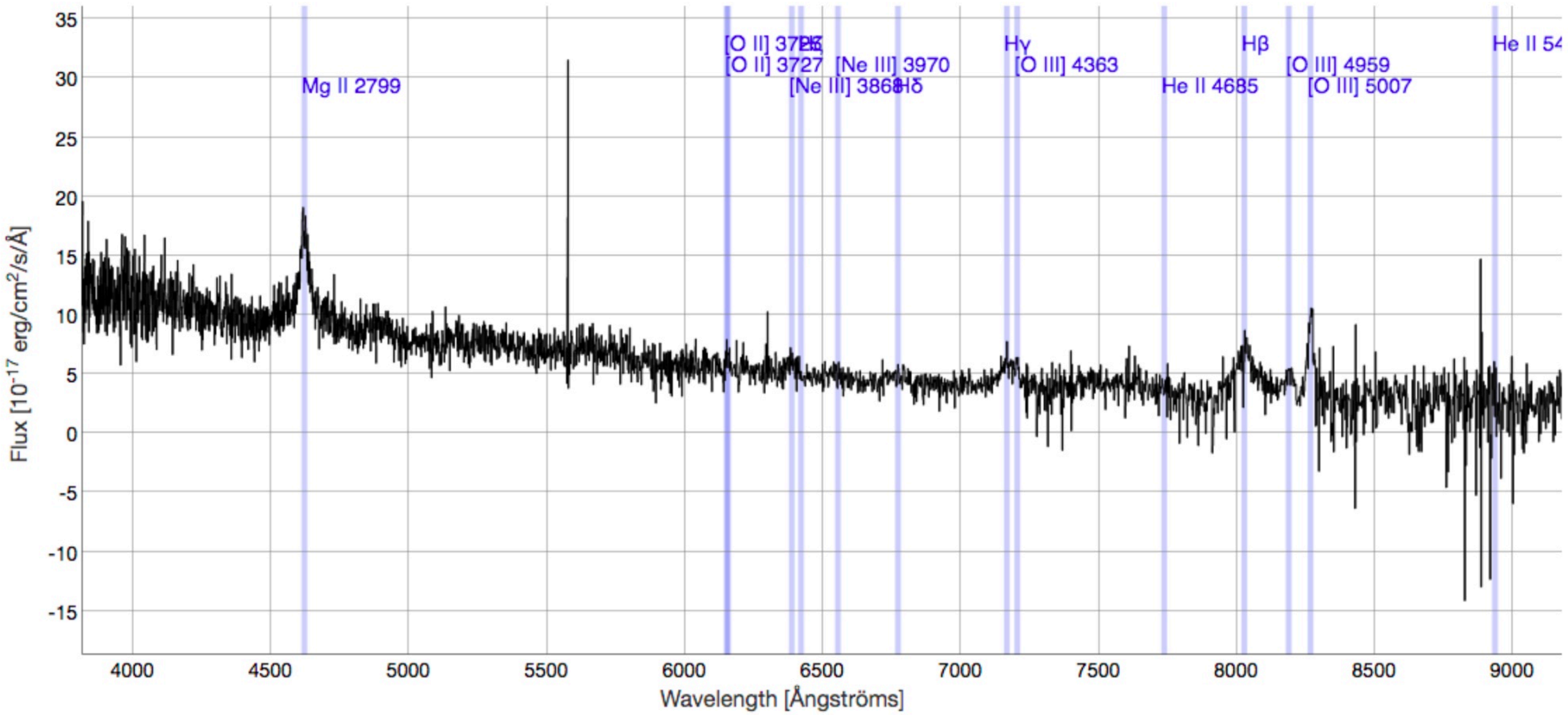


Figure: Quasar spectrum at z=0.65

QuasarNET

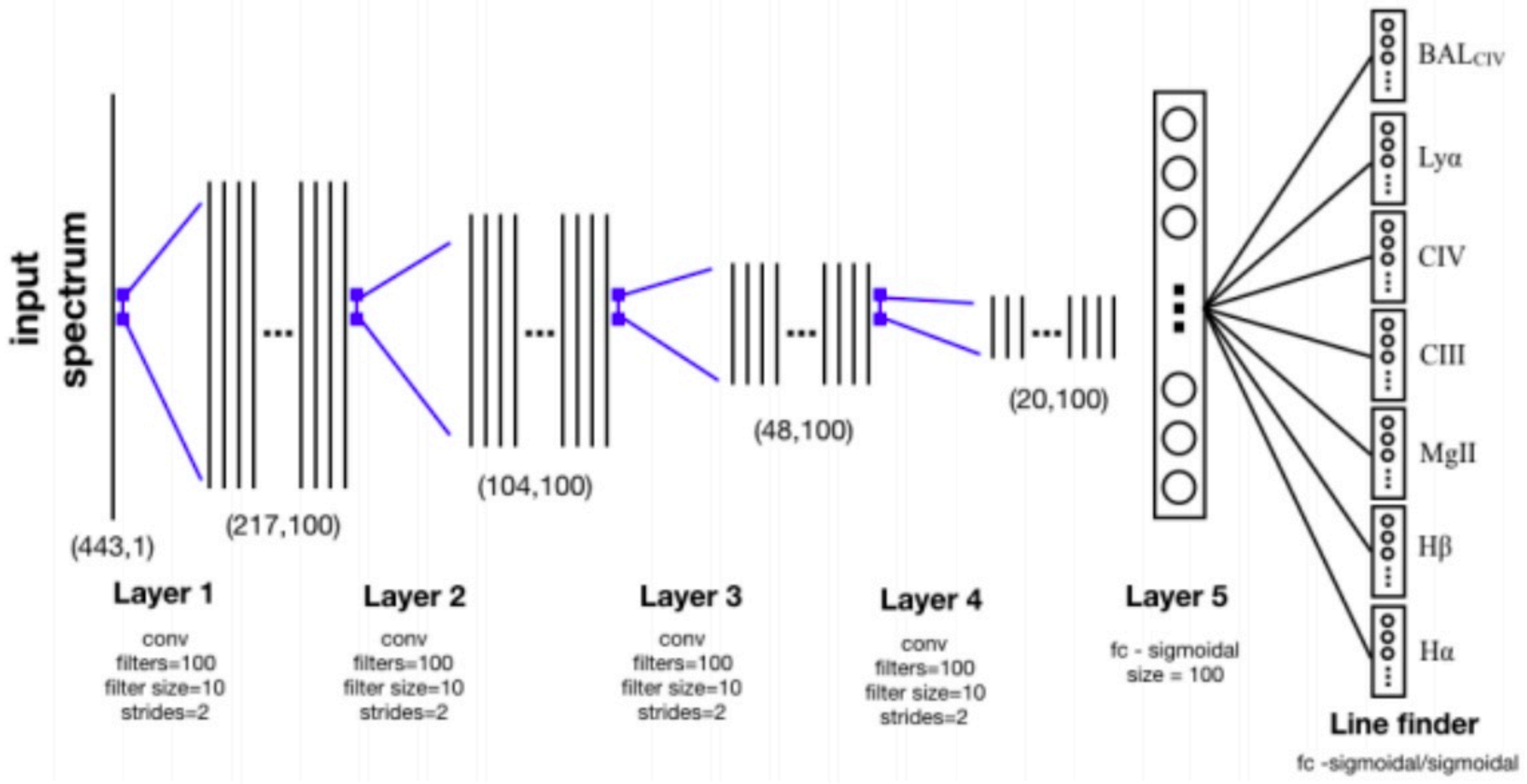


Figure: QuasarNET architecture

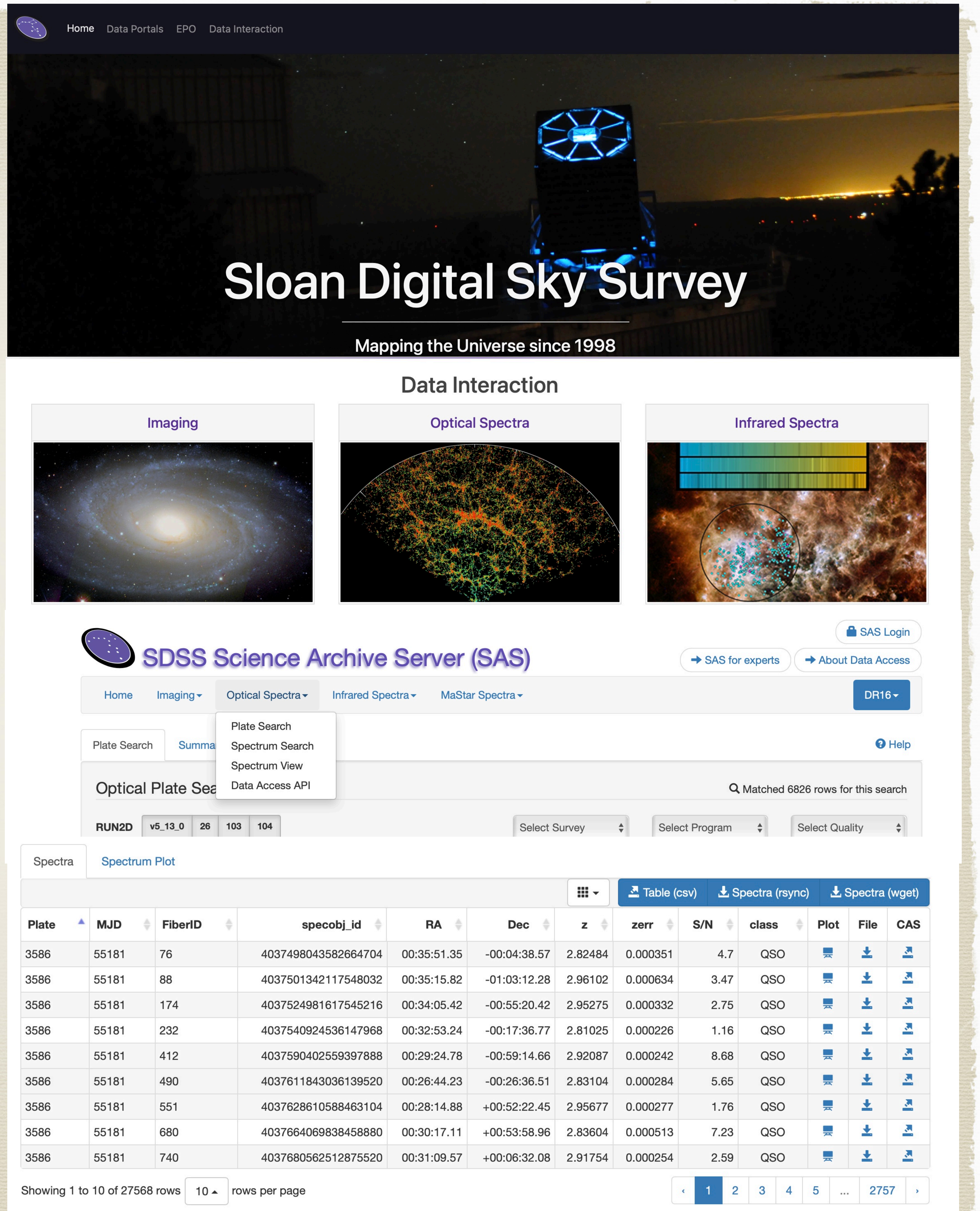
- The QuasarNET exploits a CNN to find local characteristics of spectra,
- it trains a network by DR12Q which contains confirmed quasars via visual inspection of the spectra,

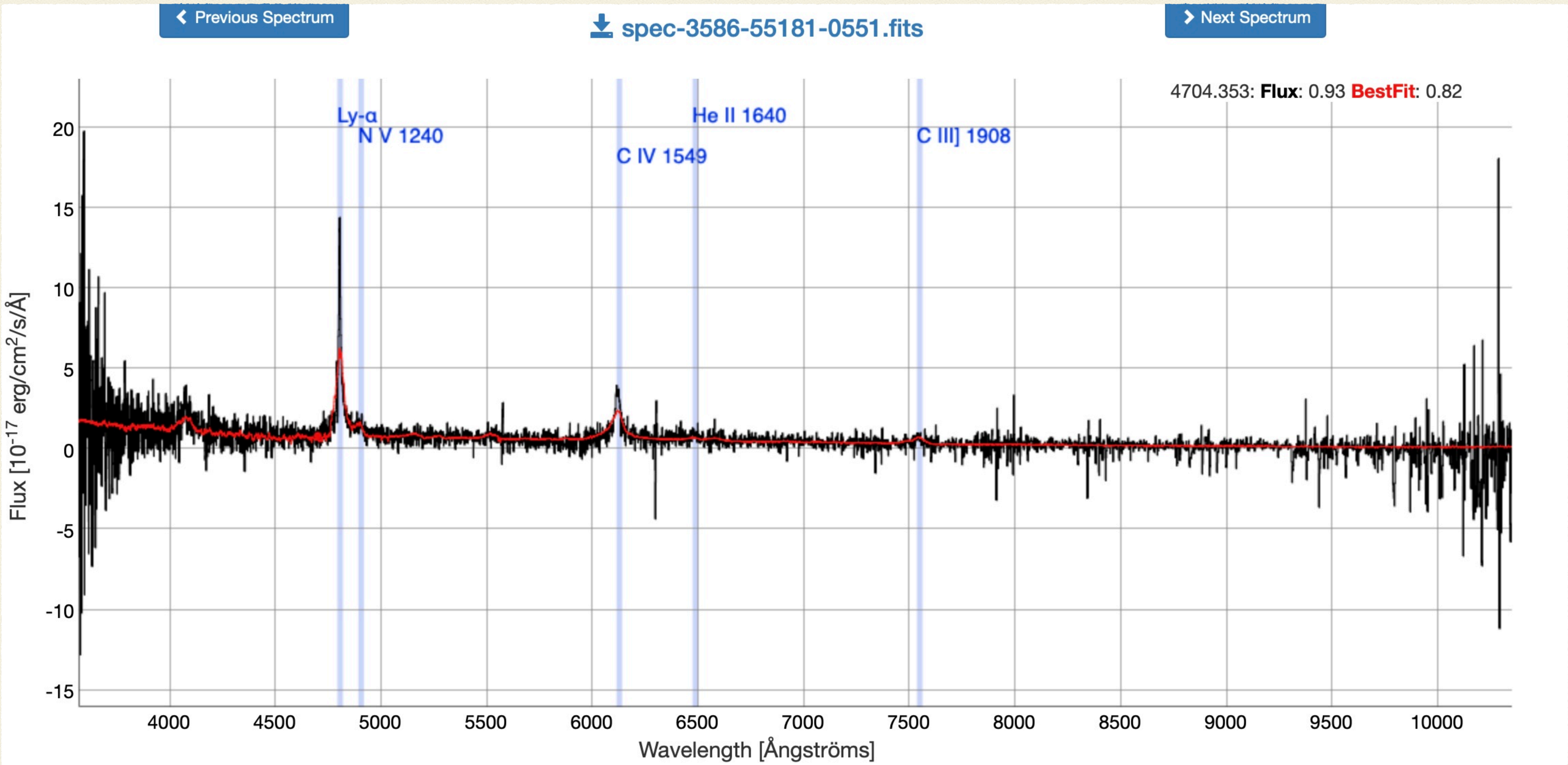
QuasarNET

- more than 250000 samples to train
- it detects at least two emission line from seven emission lines in the quasar's spectra: Lyα (121.6 nm), CIV (154.9 nm), CIII (190.9 nm), MgII (279.6 nm), $\mathrm{H}\beta$ (486.2 nm) and $\mathrm{H}\alpha$ (656.3 nm) as well as a CIV line with a broad absorption feature.
- the accuracy of this network in predicting redshift in the range of 0$<$ z $<$5.45 is 99.8$\%$ for $|\Delta\nu| <$ 6000 km/s.

Optical data download

* https://dr16.sdss.org
* Click on: ‘Optical Spectra’
* Optical Spectra >> ‘Spectrum search’
* Click on ‘BOSS’+’QSO’
* Define a range for redshift: here I choose 2.8-3
* Click on ‘Search’!
* Click on Spectra (wget)





Z= 2.95677

Class QSO

```
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3609/spec-3609-55201-0890.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3639/spec-3639-55205-0614.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3650/spec-3650-55244-0354.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3657/spec-3657-55244-0052.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3662/spec-3662-55182-0814.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3663/spec-3663-55176-0290.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3664/spec-3664-55245-0120.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3666/spec-3666-55185-0564.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3666/spec-3666-55185-0986.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3667/spec-3667-55186-0476.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3667/spec-3667-55566-0470.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3679/spec-3679-55209-0156.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3681/spec-3681-55243-0486.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3683/spec-3683-55178-0586.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3683/spec-3683-55245-0600.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3684/spec-3684-55246-0920.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3686/spec-3686-55268-0896.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3689/spec-3689-55180-0262.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3691/spec-3691-55274-0946.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3692/spec-3692-55186-0162.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3693/spec-3693-55208-0966.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3693/spec-3693-55208-1000.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3701/spec-3701-55540-0156.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3745/spec-3745-55234-0122.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3754/spec-3754-55488-0668.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3755/spec-3755-55504-0578.fits
https://data.sdss.org/sas/dr16/eboss/spectro/redux/v5_13_0/spectra/lite/3755/spec-3755-55504-0982.fits
```

```
wget -i download_url.txt
```

wget -i filename.txt

DATA PREPROCESSING

Manipulation or remove some parts, or add some parts to data in order to ensure the quality (speed, accuracy, etc) of performance

Normalization: Transforming features to be on a similar scale

Vectorization: If it is an image we have to convert it into to the pixel values which can be interpreted by the neural networks.

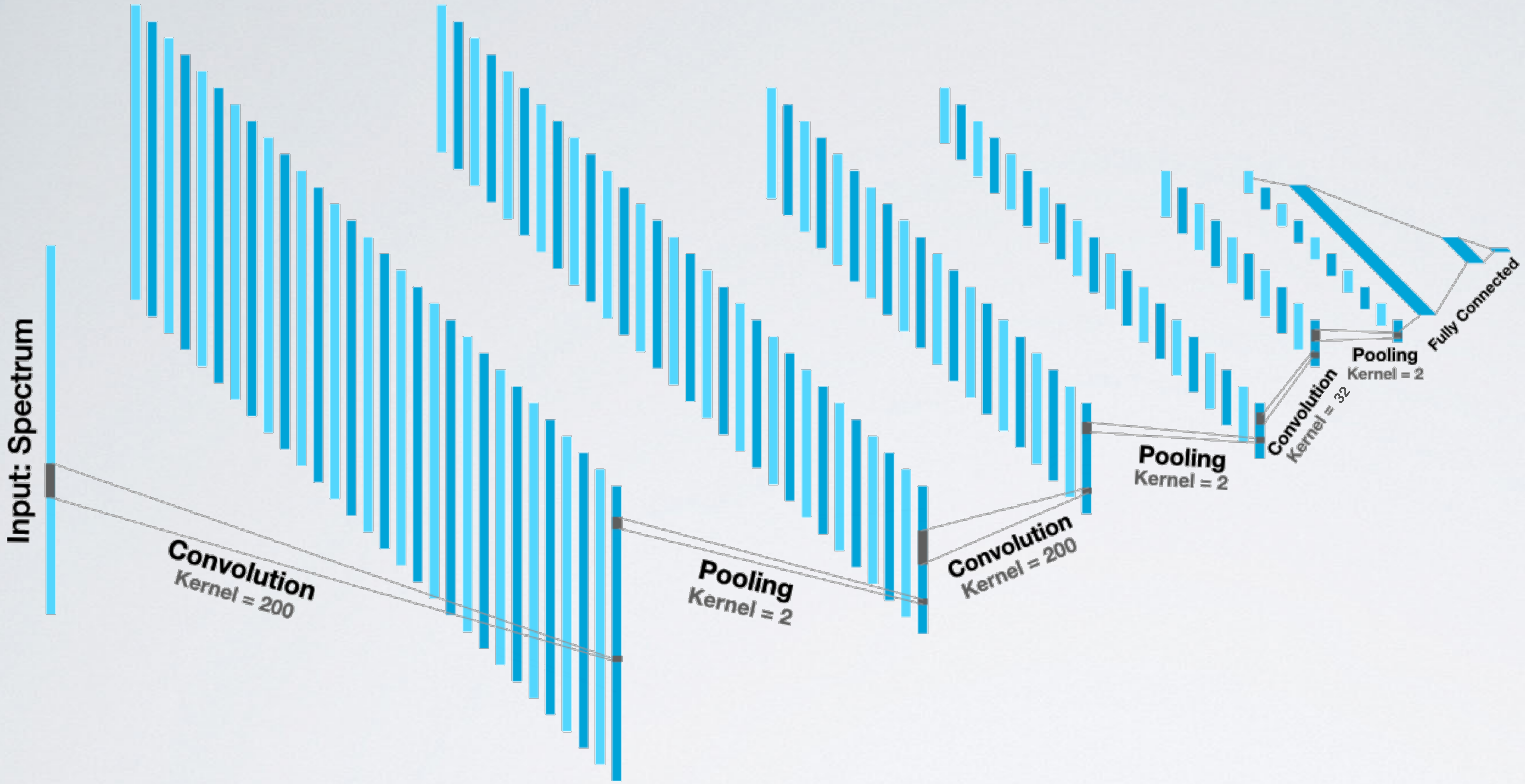
Handle missing values: Replace with mean, median value

```python
def _getData(filename,lower,upper,redshifted=False):

    obj=fits.open(filename)
    data = Table.read(obj,hdu=1).to_pandas()
    z = Table.read(obj,hdu=2)['Z'].item()
    data['lam'] = np.power(10,data['LOGLAM'])


    lam = np.linspace(lower,upper,2000)
    flux = np.interp(lam, data['lam'], data['FLUX'])

    # normalize data
    std = np.std(flux)
    avg = np.mean(flux)
    fluxNormalized = (flux-avg)/std
    dataSelected = pd.DataFrame({'lam': lam, 'flux': flux, 'fluxNormalized': fluxNormalized})
    return dataSelected,z
```

executed in 6ms, finished 17:01:42 2021-10-31

```python
folder='Dataset-2.8-3'
filenames = [x for x in os.listdir(folder) if x.endswith(".fits")]

pixels = []
zs = []

for filename in filenames:
    data, z = _getData(folder+'/'+filename ,lower=3500,upper=10000, redshifted=True )
    pixels.append(data['fluxNormalized'].tolist())
    zs.append(z)


pixels = np.asarray(pixels)
zs = np.asarray(zs)

pixels.tofile('flux-2.8-3.csv', sep=",")
zs.tofile('zpipe-2.8-3.csv', sep=",")
```

One Dimensional Convolutional Neural Network



Structure of one dimensional CNN. The spectrum of quasar is input as a one-dimensional array, which goes through the convolutional layer of kernel size = 200, 200, 20 respectively in order to search for the global and local pattern. The fully connected layers output the redshift.

```
class RNet(nn.Module):
    def __init__(self):
        super(RNet, self).__init__()
        self.layer1 = nn.Sequential(
            nn.Conv1d(1,32, 200),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.layer2 = nn.Sequential(
            nn.Conv1d(32,16,200),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.layer3 = nn.Sequential(
            nn.Conv1d(16, 128, 20),
            nn.ReLU(),
            nn.MaxPool1d(2, stride=2))
        self.drop_out = nn.Dropout()
        self.fc1 = nn.Sequential(
            nn.Linear(8576, 900),
            nn.ReLU())
        self.fc2 = nn.Sequential(
            nn.Linear(900, 100),
            nn.ReLU())
        self.fc3 = nn.Linear(100, 1)

    def forward(self, x):
        x = x.unsqueeze(1)
        out = self.layer1(x)
        out = self.layer2(out)
        out = self.layer3(out)
        out = out.reshape(out.size(0), -1)
        out = self.drop_out(out)
        out = self.fc1(out)
        out = self.fc2(out)
        out = self.fc3(out)
        return out
```

MNRAS **000**, i–?? (2015) Preprint 4 November 2021 Compiled using MNRAS LaTeX style file v3.0

Deep Learning in Searching the Spectroscopic Redshift of Quasars

F. Rastegar Nia,[1,2,3]★ M. T. Mirtorabi,[1]† R. Moradi,[2,3,4]‡ A. Vafaei. Sadr,[5]§ Y. Wang[2,3,4]¶
[1] *Department of Physics,Faculty of Physics and Chemistry, Alzahra University,Tehran, Iran*
[2] *ICRANet, Piazza della Repubblica 10, I-65122 Pescara, Italy*
[3] *ICRA, Dipartimento di Fisica, Università di Roma "La Sapienza", Piazzale Aldo Moro 5, I-00185 Roma, Italy*
[4] *INAF – Osservatorio Astronomico d'Abruzzo,Via M. Maggini snc, I-64100, Teramo, Italy*
[5] *Departement de Physique Theorique and Center for Astroparticle Physics, University of Geneva*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
Study the cosmological sources at their cosmological rest-frames are crucial in order to track the cosmic history and properties of the compact objects. In view of increasing data volume of existing and upcoming telescopes/detectors we here apply the 1–dimensional convolutional neural network (CNN) to estimate the redshift of quasars in Sloan Digital Sky Survey IV (SDSS-IV)

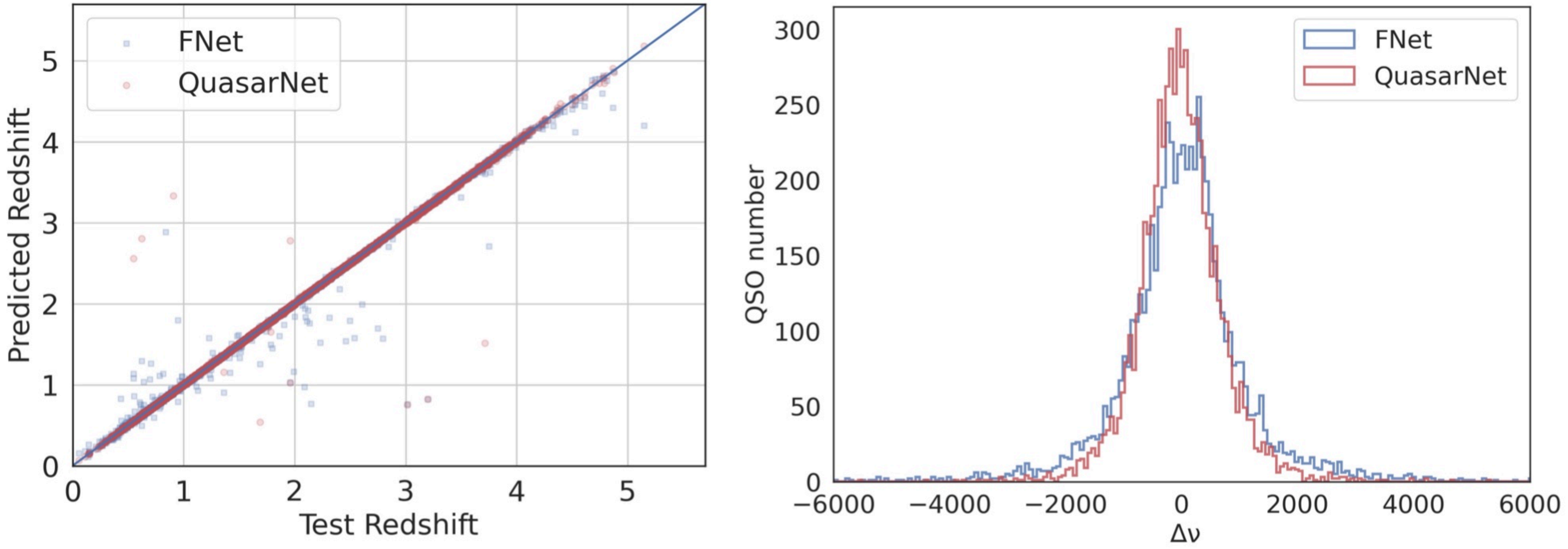


Figure: **left**: Redshift predicted by FNet and QuasarNet for $0 < z \leq 5.45$. The predicted redshift vs. the visually inspected redshift selected from the DR16Q catalog. **right:** The velocity difference for redshift predicted by FNet, QuasarNet and PCA when compared to visually inspected (VI) redshift. $|\Delta\nu| < 6000\ \mathrm{km}/s$ in $0 < z \leq 5.45$; The accuracy for FNet is 98.3%. For Quasar Net is 99.8%. The accuracy for $|\Delta\nu| < 12000\ \mathrm{km}/s$ is 99.0% for FNet and 99.8% for QuasarNet.