# Constraining Cosmology using Self-Attention

Se Yeon Hwang, Cristiano Sabiu
University of Seoul

2022.11.03
Workshop for Korea-UK AI/ML Research in Fundamental Sciences

Motivation

Prior research : 1711.02033v1, 1908.10590v5

Cosmological parameters

Motivation

Prior research : 1711.02033v1, 1908.10590v5


Cosmological parameters

Some universe!
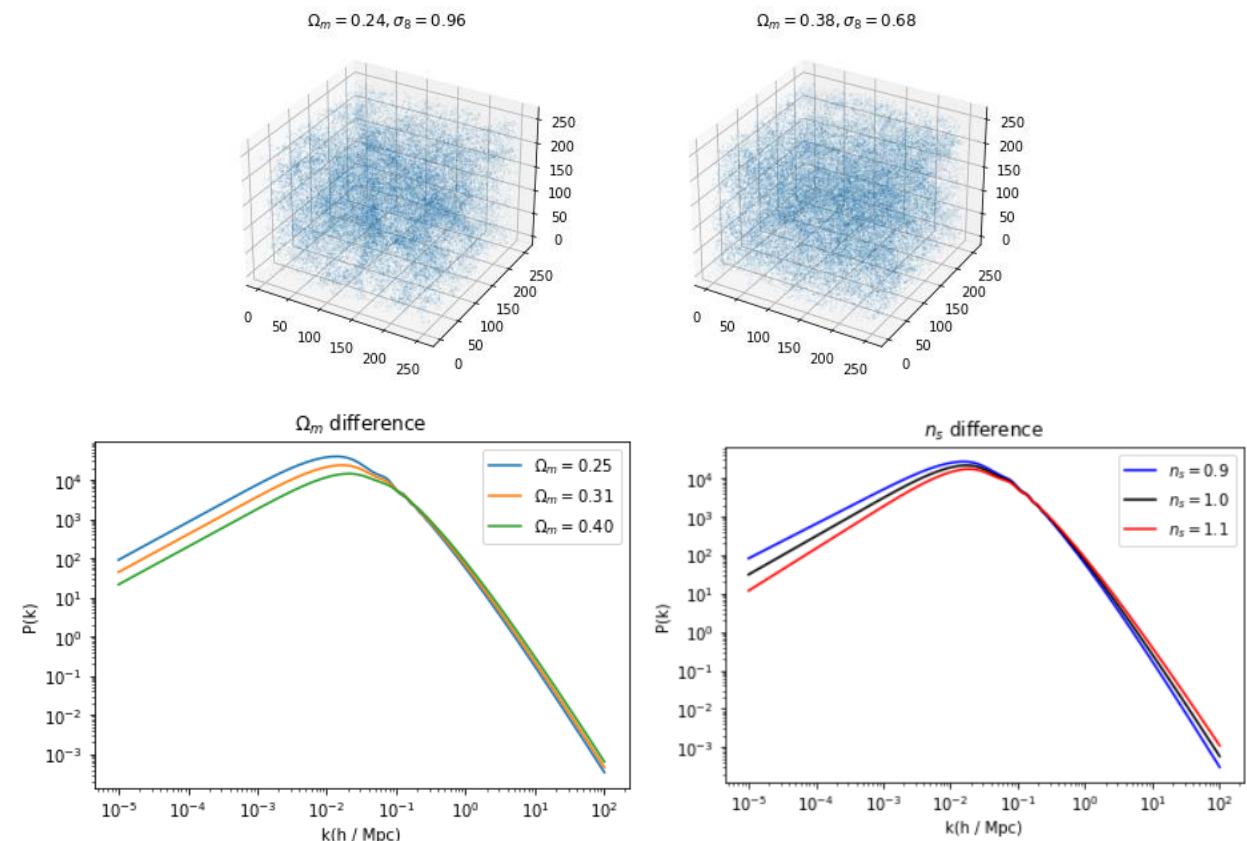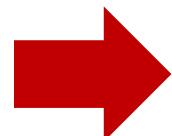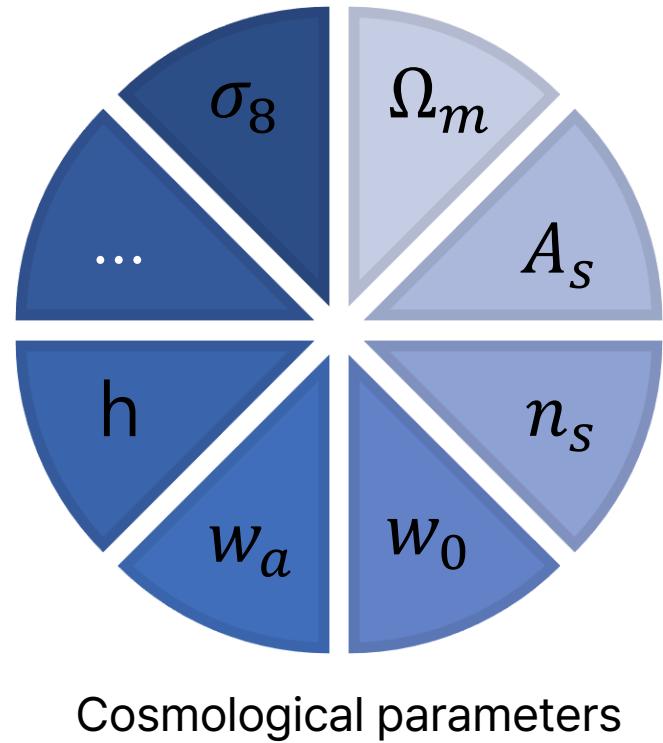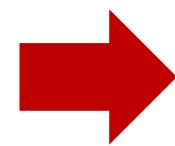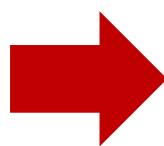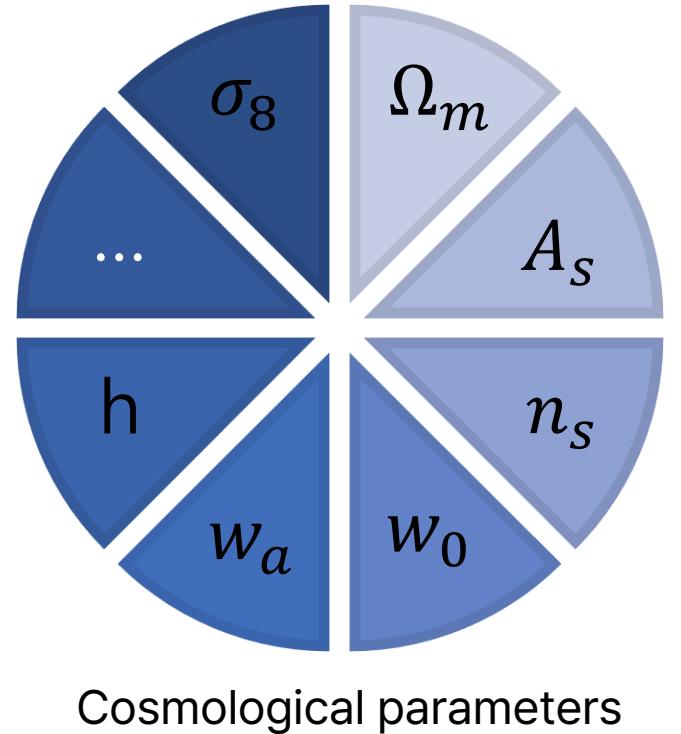
Machine Learning (vision task)

Previous work

COSMOLOGICAL PARAMETER ESTIMATION FROM LARGE-SCALE STRUCTURE DEEP LEARNING
(Shuyang Pan, 2020, submitted to ApJ)[arXiv : 1908.10590v5]
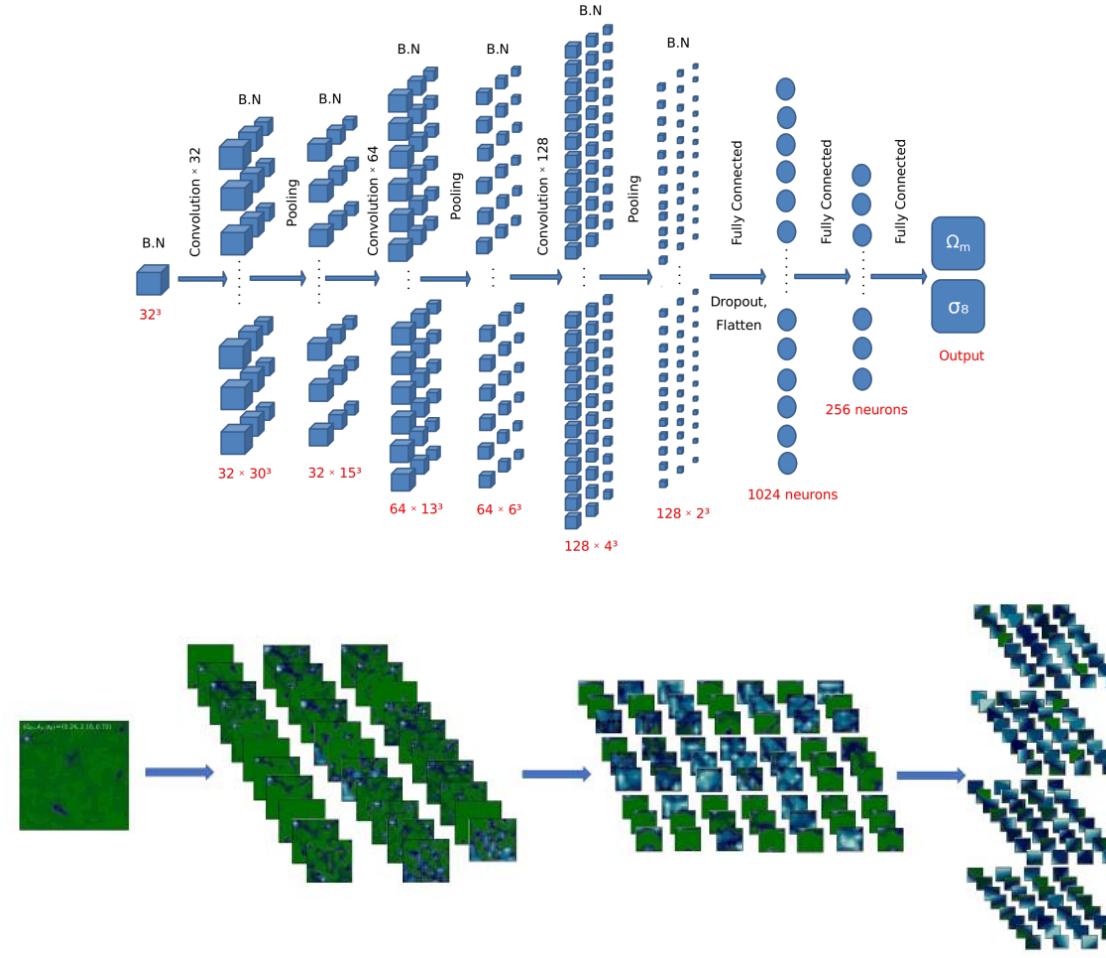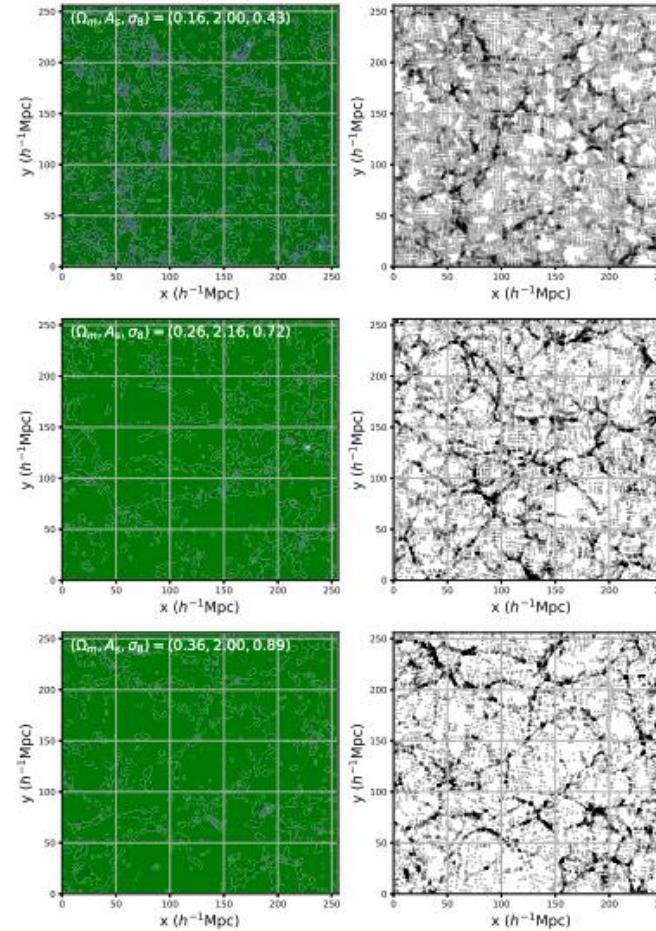
Previous work

COSMOLOGICAL PARAMETER ESTIMATION FROM LARGE-SCALE STRUCTURE DEEP LEARNING
(Shuyang Pan, 2020, submitted to ApJ)[arXiv : 1908.10590v5]
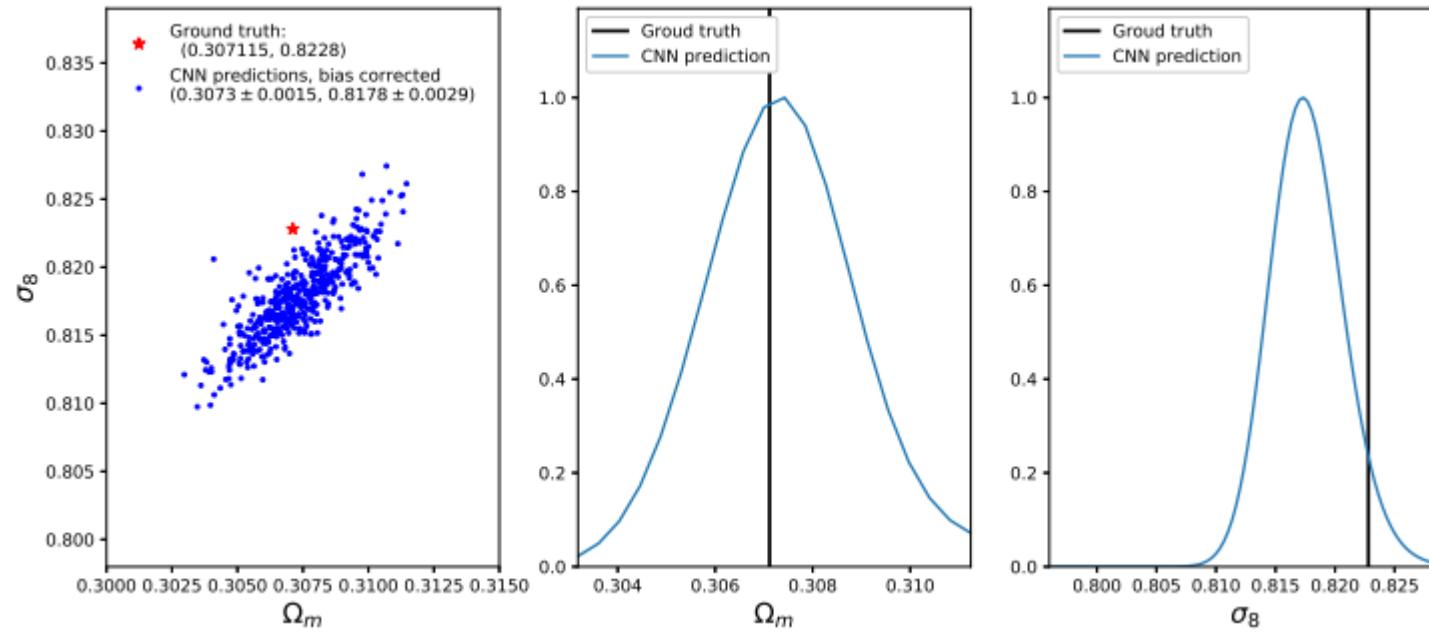
Previous work

DeepLSS: breaking parameter degeneracies in large scale structure with deep learning analysis of combined probes(Tomasz Kacprzak, 2022, submitted to PhysRevX) [arXiv : 2203.0961]
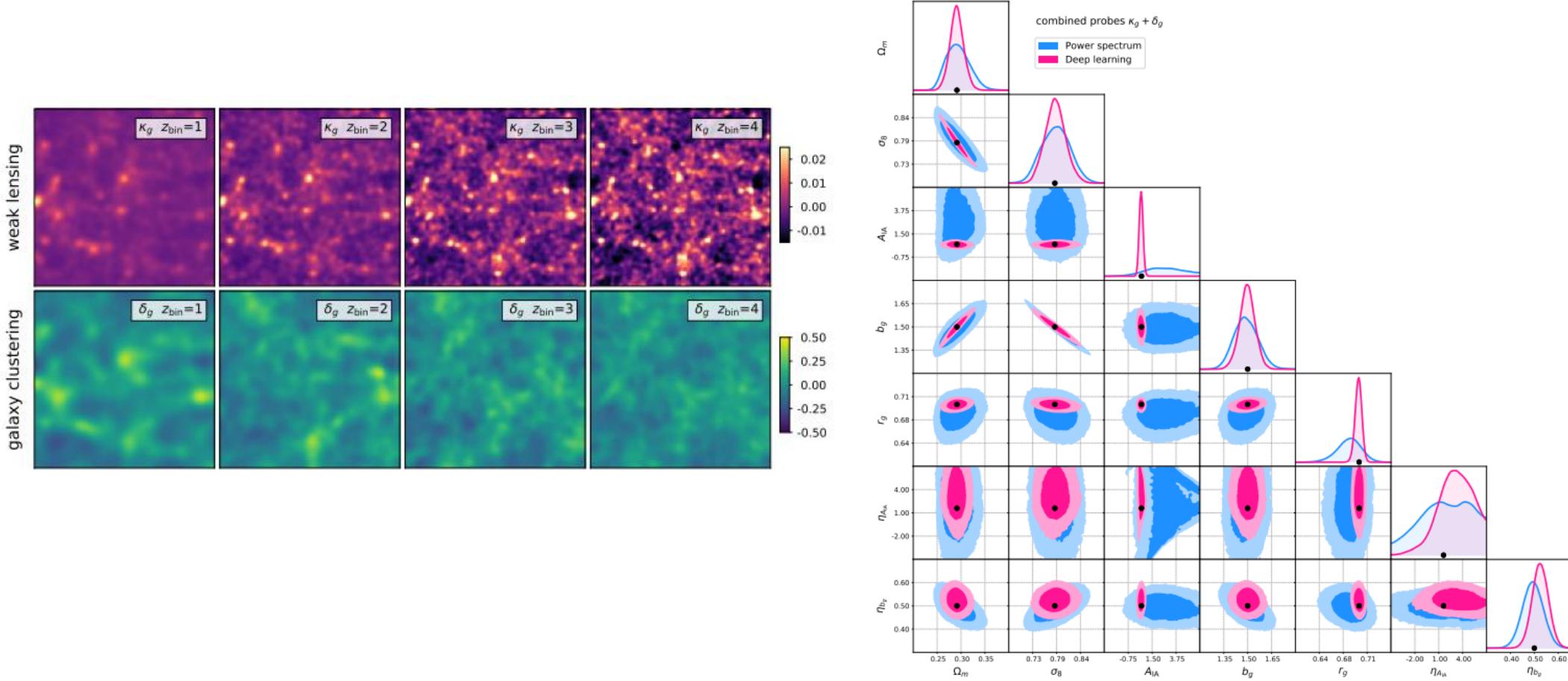
# Using deep learning on predicting cosmological parameters

How CNN works

input

output

Ω_m

Fully Connected
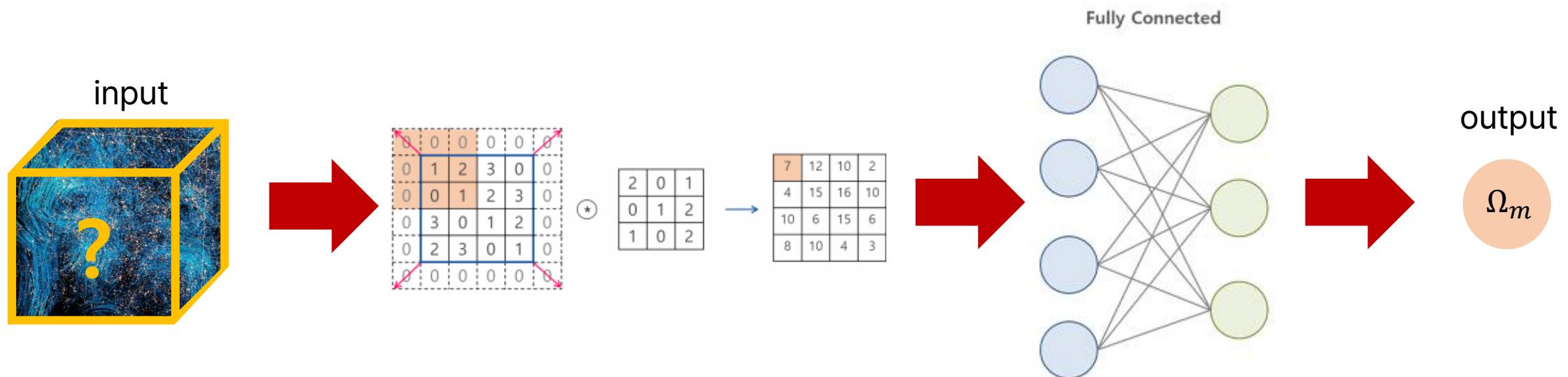
How CNN works
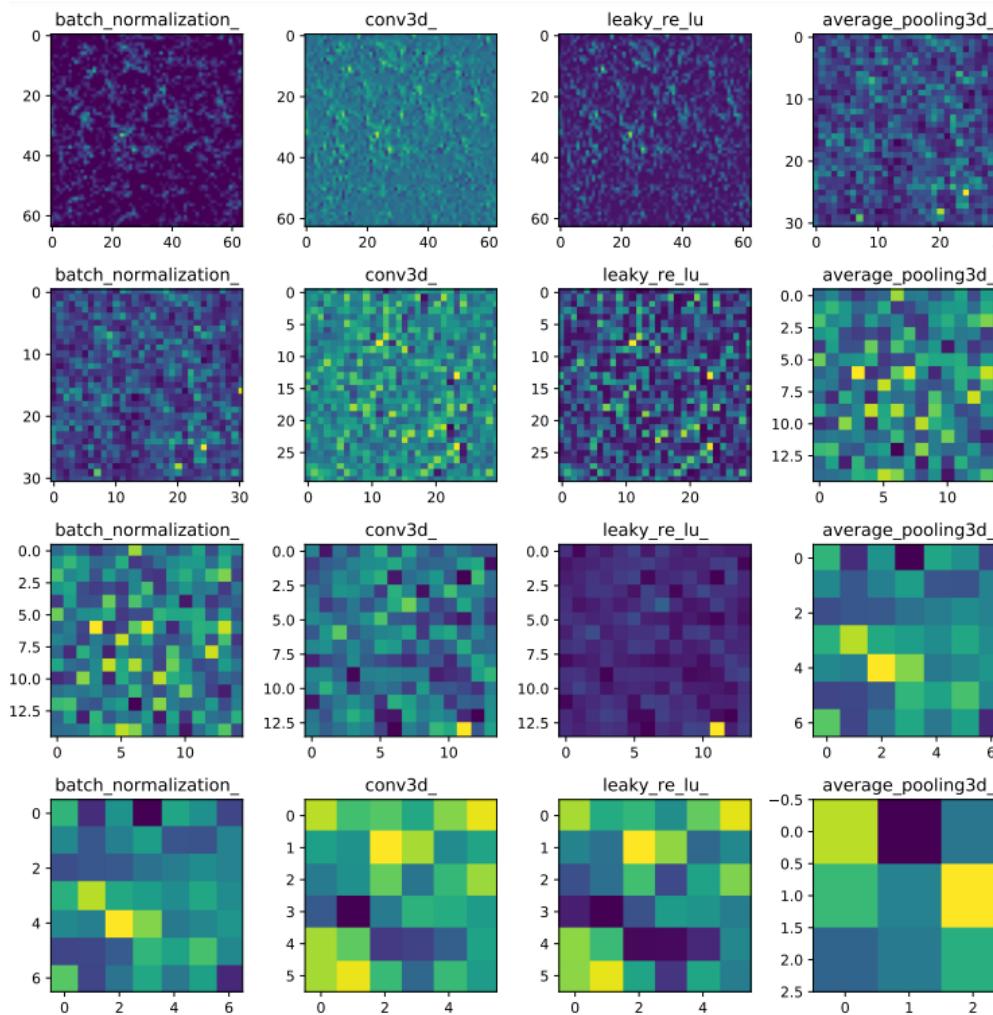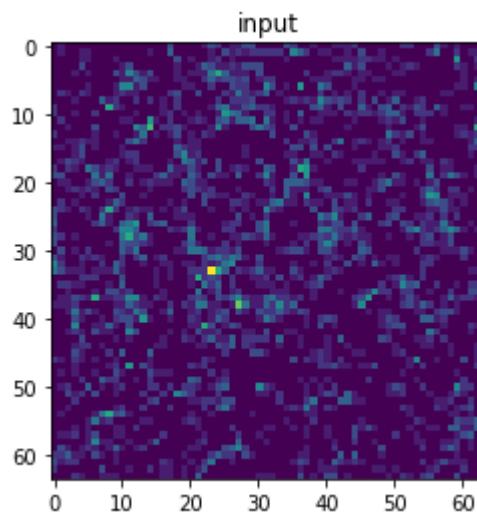
What is ViT and how it works?
And what is self attention?

Transformer and self attention

Self attention⊂Transformer

(Ashish Vaswani et al, 2017) arXiv[1706.03762v5]

Natural language process(NLP)

Google BERT

Self attention⊂Vision Transformer

(Alexey Dosovitskiy et al, 2020) arXiv[2010.11929v2]

Computer Vision task

How ViT works

Dividing into patches

Flatten & make 1d array

Transformer

Output

Image patches are treated the same way
as tokens(words) in a sentence
e.g. I // am // a // student

Calculate the "attention score"
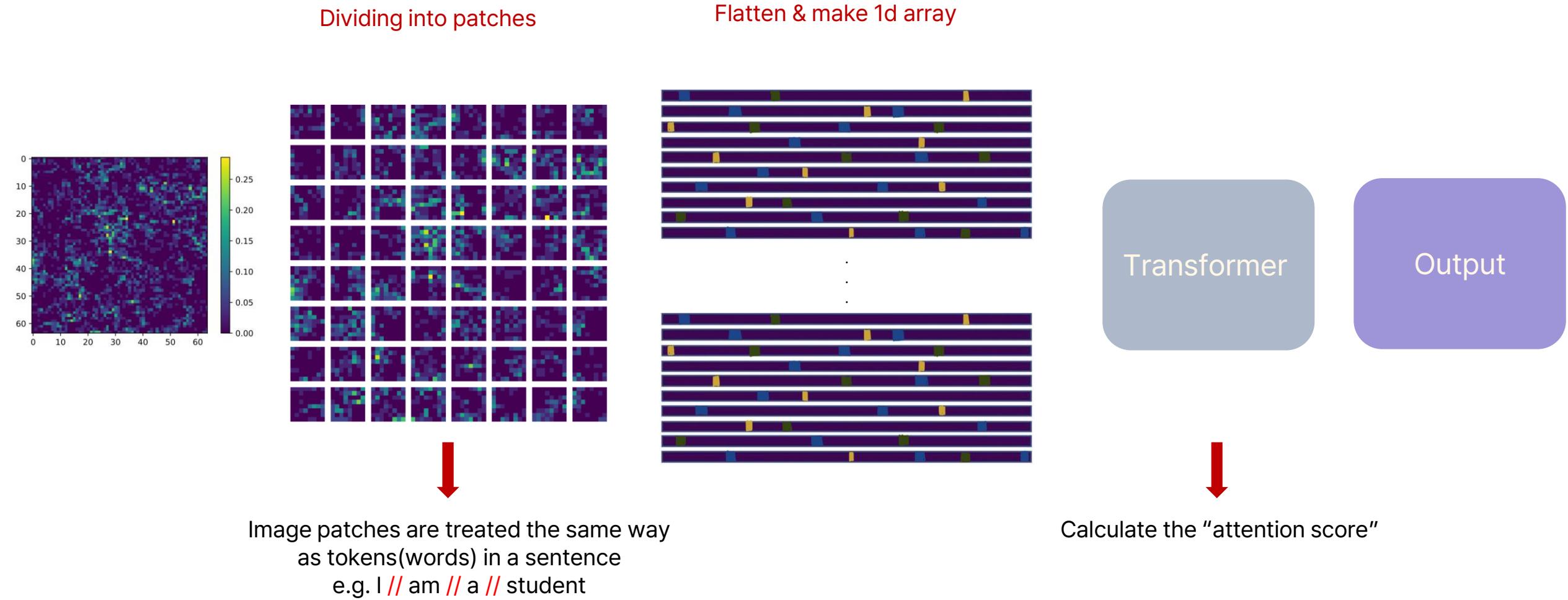
What is attention

Input Image

Reference : https://keras.io/examples/vision/probing_vits/
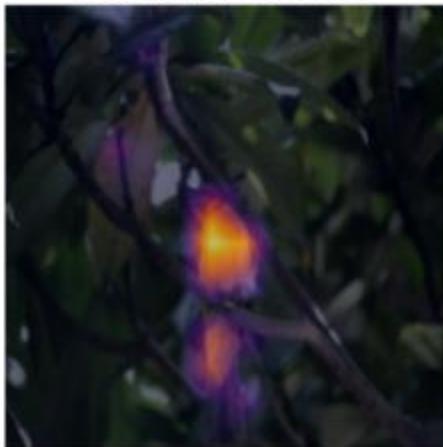
Attention head: 0
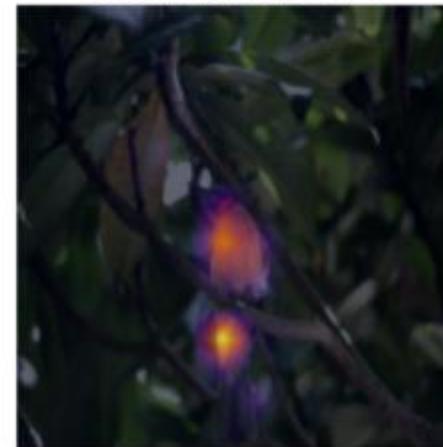
Attention head: 1
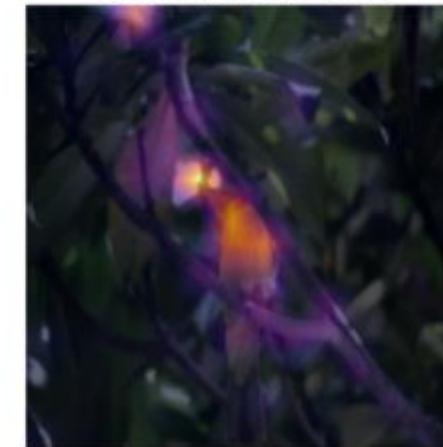
Attention head: 2

Attention head: 3
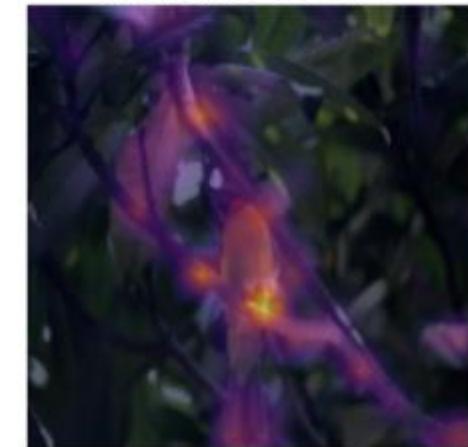
What is attention

Reference : https://keras.io/examples/vision/probing_vits/

Attention head: 0 Attention head: 1 Attention head: 2 Attention head: 3
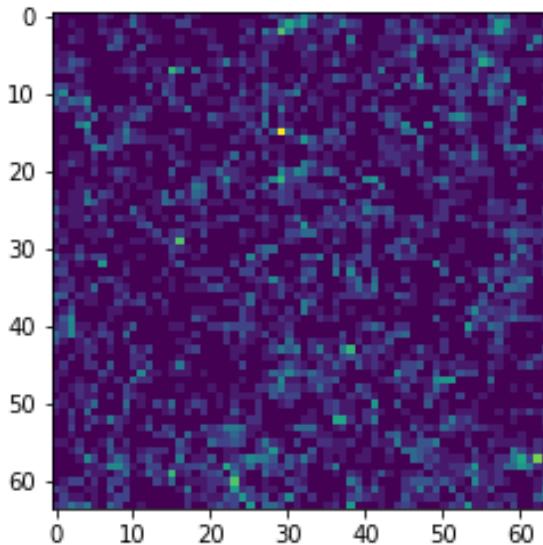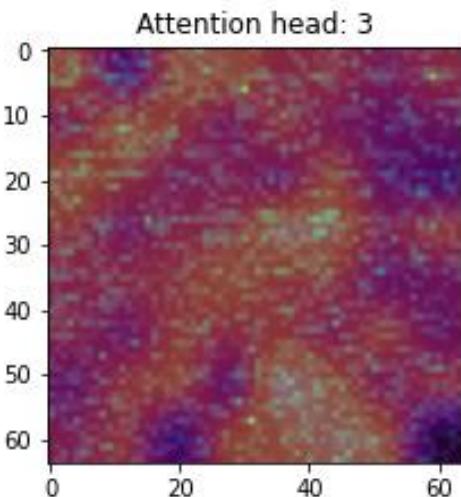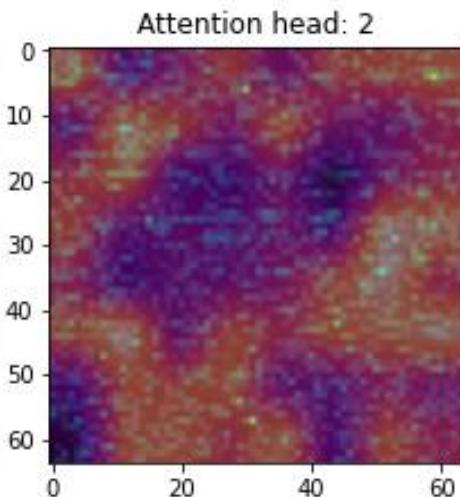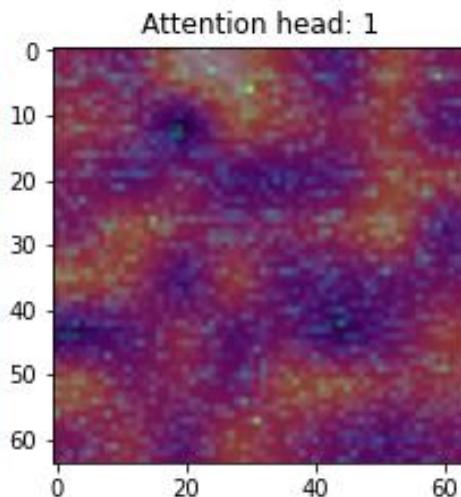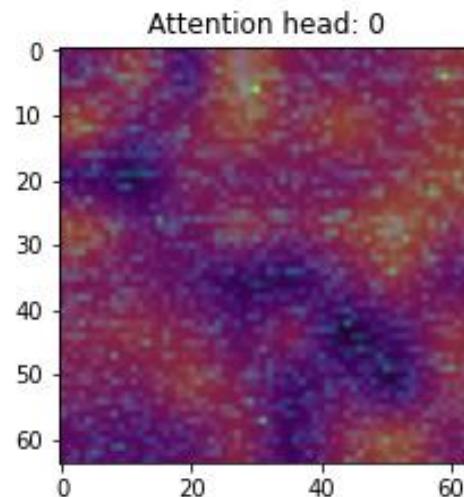
Difference between CNN and ViT
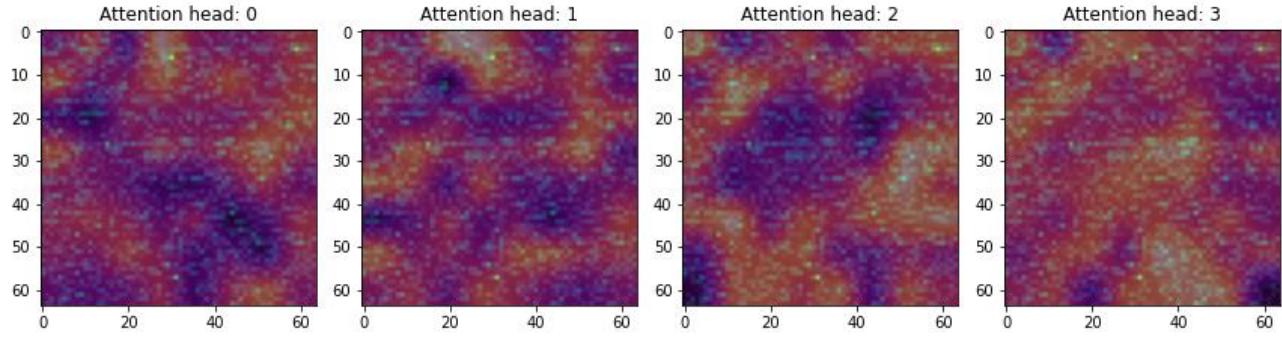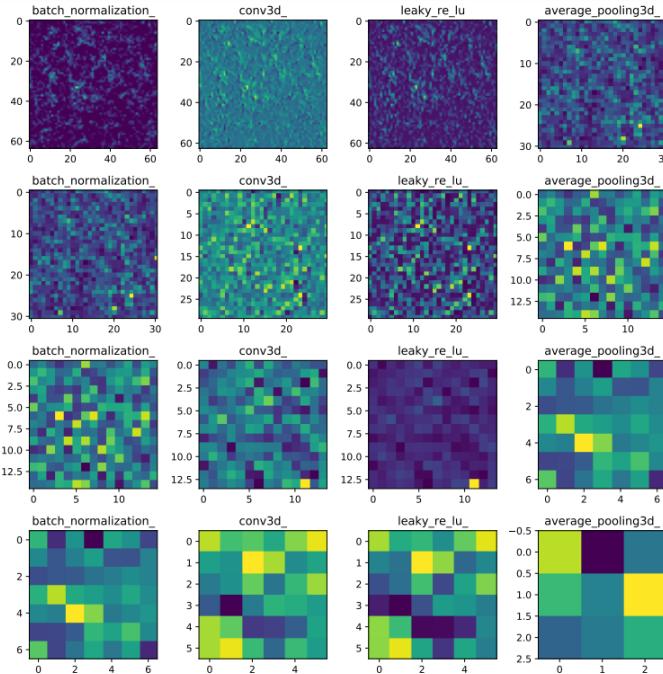
- ❖ ViT looks global relationship between patches while CNN is doing convolution between neighborhood patches
- ❖ This can cause ViT to have less inductive bias which is ability to inference the data which is not seen in training process because the ViT view whole data while CNN compressed down the data
- ❖ It is important to pre-train with large-dataset and also how you divide the entire image into patches(patch size) is important because ViT's locality mainly comes from this.

Data

Code : PINOCCHIO(PINpointing Orbit Crossing Collapsed HIerarchical Objects),(2002, Pierluigi Monaco)
- output: : dart matter halo catalog
- This used approximate Lagrangian Perturbation Theory(LPT)

Different cosmology

➔ To train and test model

1000 simulations
with different initial seed

single cosmology

➔ To check trained model error

100 simulations
with different initial seed

Data

Code : PINOCCHIO(PINpointing Orbit Crossing Collapsed HIerarchical Objects),(2002, Pierluigi Monaco)

Different cosmology

Parameter range
$\Omega_m \subset [0.25, 0.4]$, $\sigma_8 \subset [0.75, 0.85]$,
$\omega_0 \subset [-1.3, -0.7]$, $\omega_a \subset [-0.3, 0.3]$,
$n_s \subset [0.9, 1.1]$, $h \subset [0.6, 0.8]$
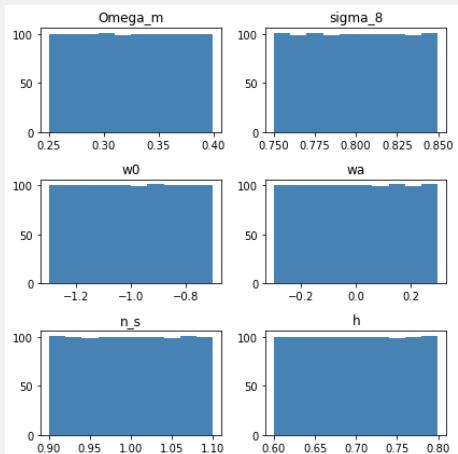
1000 simulations

single cosmology

|	
Ω_m	**0.3133**
σ_8	0.8079
ω_0	-1
ω_a	0
n_s	0.9649
h	0.6736

100 simulations

Data

- For the machine learning input data, I made 3d-histogram(3d-histogram axis: RA, DEC and observed redshift)
- Range cut : -20 < RA,DEC < 20 and 0.3 < z < 0.8
- Bin size = 64^3
- The color indicate the number of dark matter halos in the pixels


3-d histogram
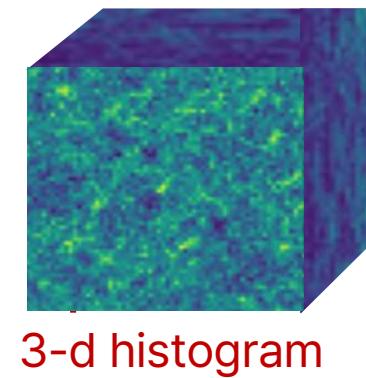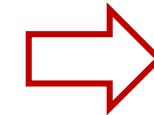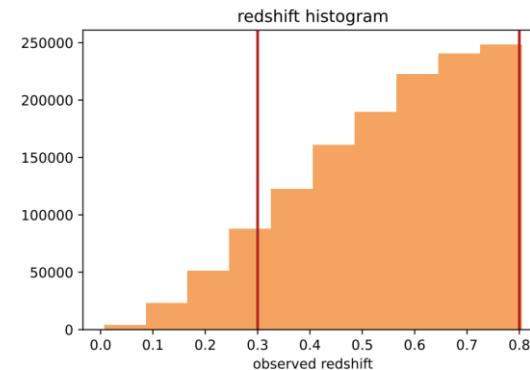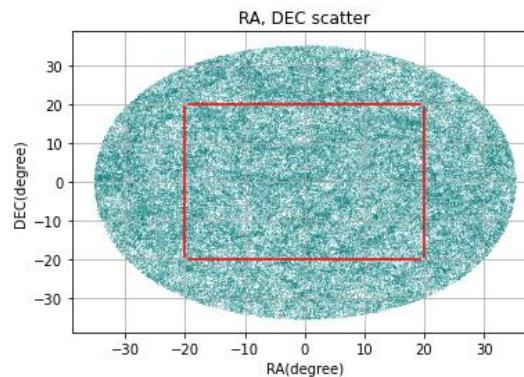
result
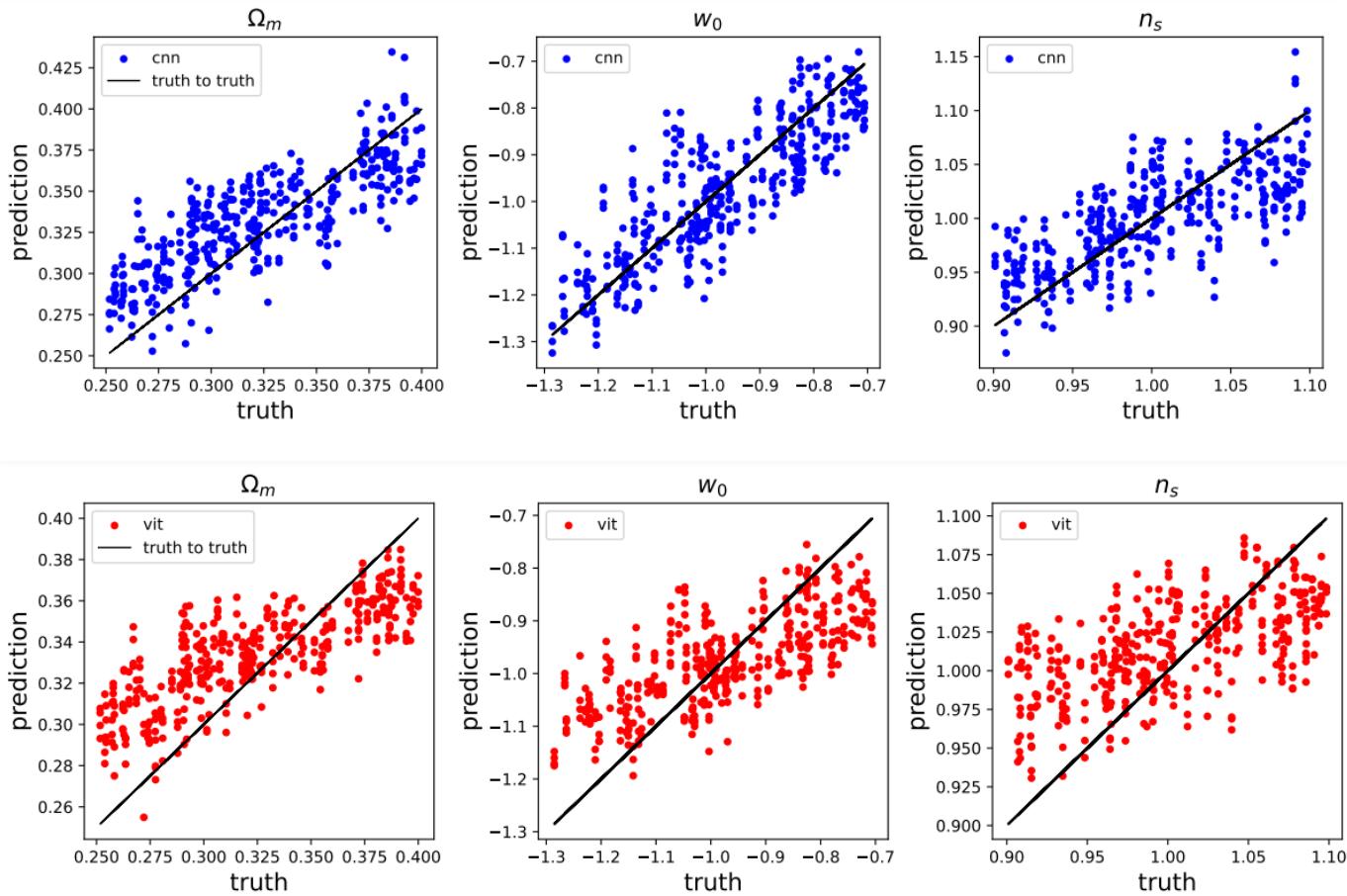
Blue : CNN
Red : ViT

	time	# of parameter
CNN	~40m	9,162,439
ViT	~90m	22,356,483

result

Blue : CNN
Red : ViT
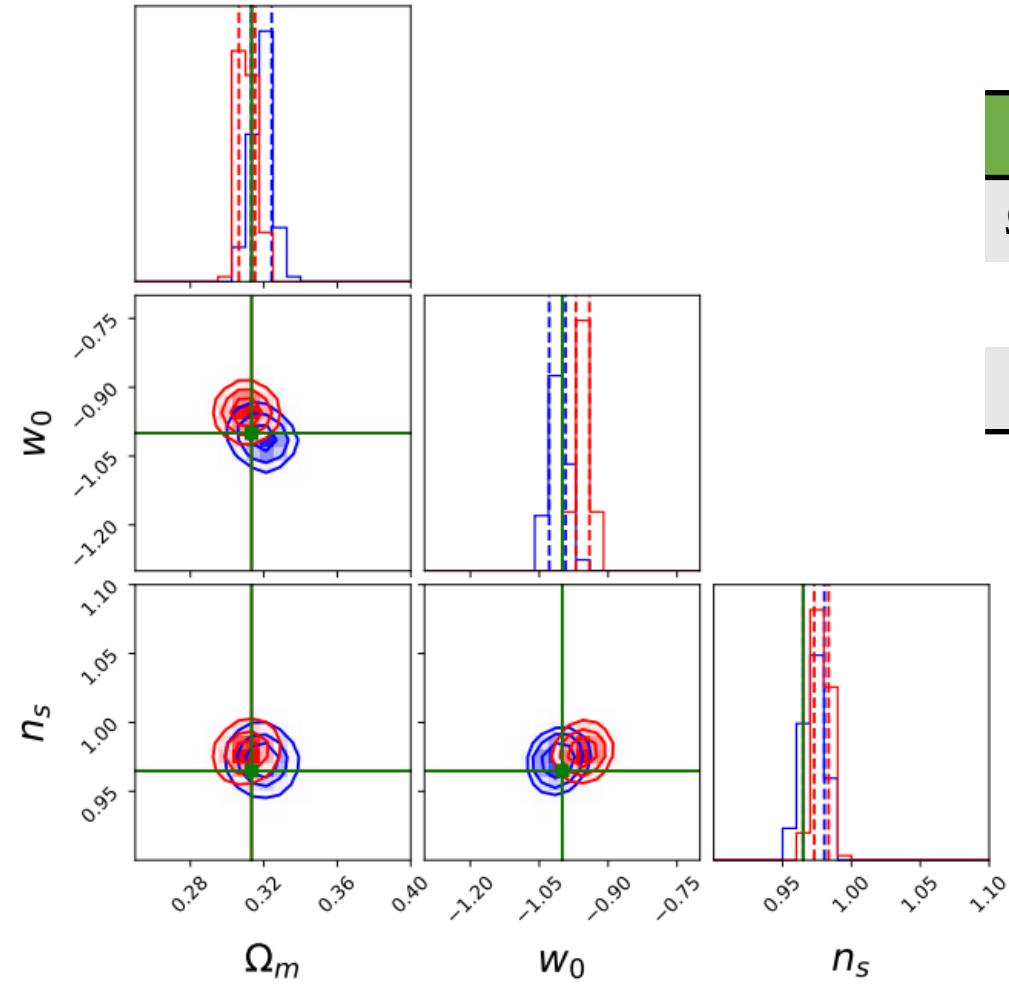
|	CNN	ViT
Ω_m	0.00578	0.00432
ω_0	0.01928	0.01213
n_s	0.00749	0.00626

summary
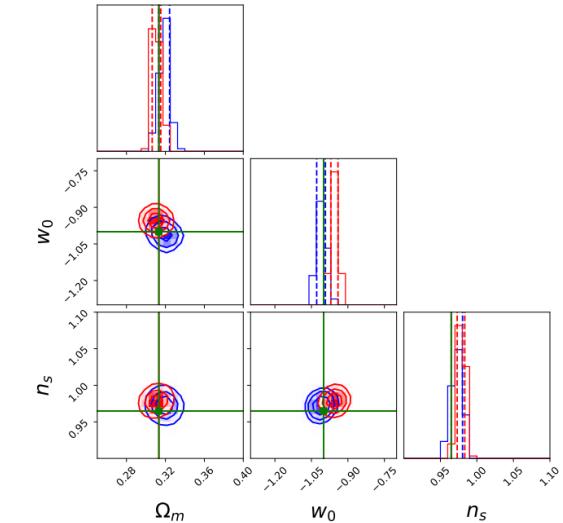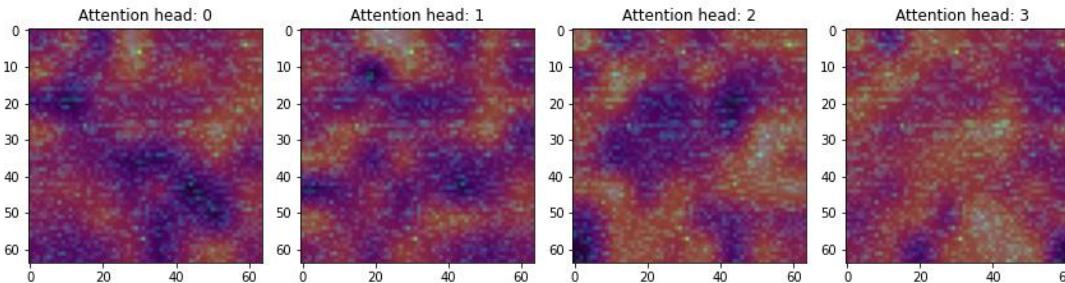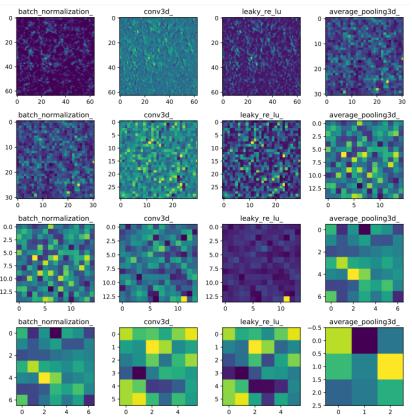
Take home message : ViT can work as CNN in predicting cosmological parameters
Future : Can ViT be more useful than CNN in aspect of time or performance?

Thank you for your attention!

result

Position embedding on the ViT