Masked Autoencoders Are More Than Scalable Vision Learners

DSA5204 - Project Group 5

April 2024

Kho Tze Jit	A0215110E
Liaw Zheng Kai	A0222733M
Muhammad Haidi Bin Azaman	A0216941E
Liu Han	A0194490X
Foong Xin Yu	A0213920R
Chia Yi Min Matthew	A0217187Y

Introduction

Model Architecture: Masked Autoencoders (MAE) with Vision Transformer (ViT)
Paper: “Masked Autoencoders are Scalable Vision Learners”
Task: Image reconstruction of a noisy image

Background

Masked Autoencoders as Scalable Vision Learners

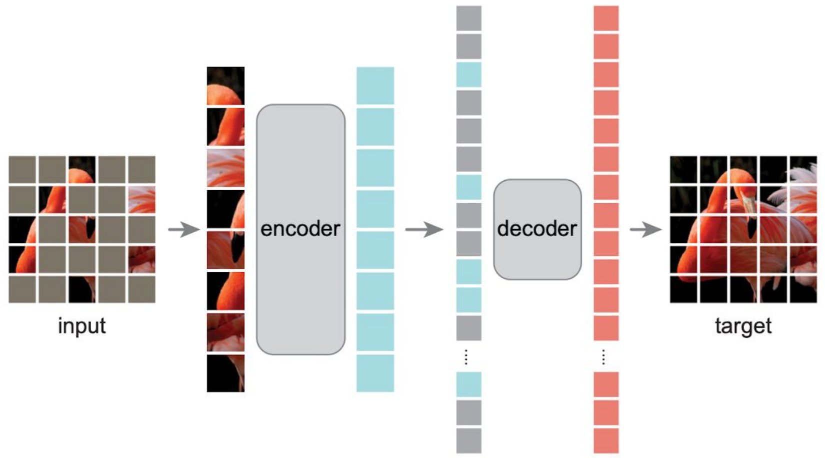
- MAEs introduced in 2010 by Pascal Vincent et al. (1)
- What benefits does MAE bring?
    - Self-supervised learning
    - Generalized models
- "Masked Autoencoders are Scalable Vision Learners" hypothesizes
    - MAE demonstrates a rich hidden representation of learned data when masked

Note:
(1) Extracting and Composing Robust Features with Denoising Autoencoders

Background

Masked Autoencoders as Scalable Vision Learners

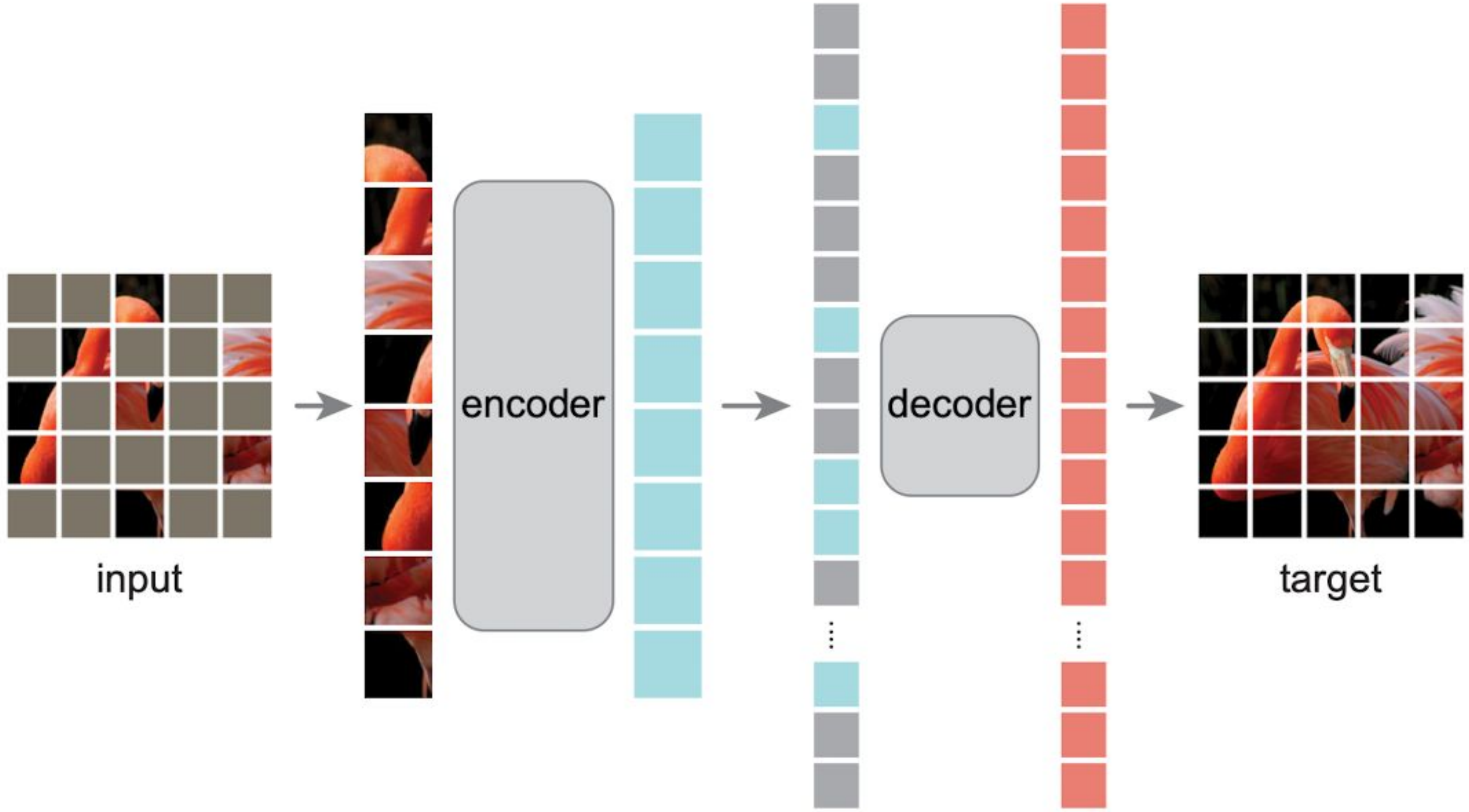
- MAEs introduced in 2010 by Pascal Vincent et al. (1)
- What benefits does MAE bring?
    - Self-supervised learning
    - Generalized models
- "Masked Autoencoders are Scalable Vision Learners" hypothesizes
    - MAE demonstrates a rich hidden representation of learned data when masked
    - Architecture:



Note:
(1) Extracting and Composing Robust Features with Denoising Autoencoders

Background: Proposed Architecture

Masked Autoencoders as Scalable Vision Learners



Project Objectives

In light of the benefits of MAE, we seek to explore:

<u>Reproduction:</u>

Re-implementation of the paper's main experiment

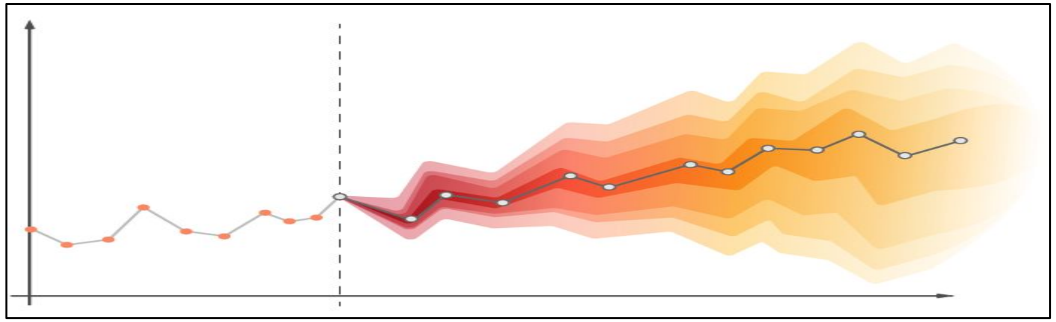
<u>Extensions:</u>

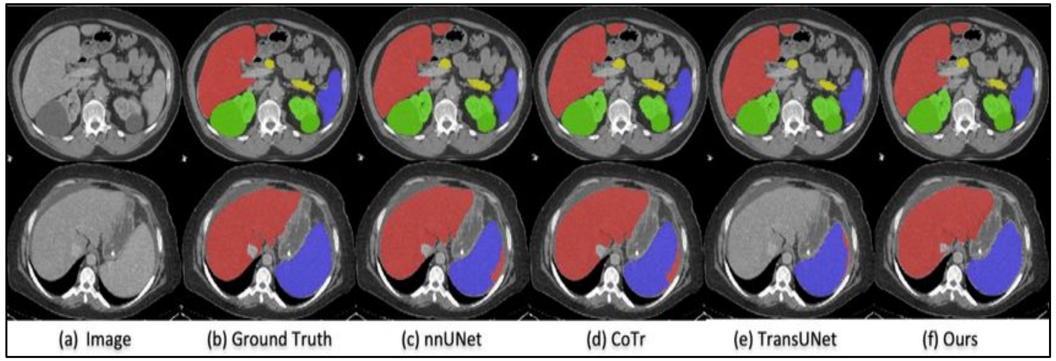
1. Time series reconstruction and prediction

2. Semantic Segmentation

3. 3D segmentation of medical CT scans

4. Generating appropriate samples for data imputation





Reproduction

Reproduction

Main Experiment: Comparison of image classification performance between pre-training with MAE and no pre-training

- Methodology:
  - Pre-training MAE with ViT-B backbone for 600 epochs and fine-tuning for 100 epochs
  - Training ViT-B from scratch for 200 epochs

Reproduction

Main Experiment: Comparison of image classification performance between pre-training with MAE and no pre-training

- Methodology:
  - Pre-training MAE with ViT-B backbone for 600 epochs and fine-tuning for 100 epochs
  - Training ViT-B from scratch for 200 epochs
- Dataset: Tiny ImageNet
  - 100,000 images of 200 classes (500 for each class) downsized to 64 x 64 coloured images
  - Smaller and fewer images as compared to ImageNet
- Metrics: Top-1 Accuracy

Reproduction

Main Experiment: Comparison of image classification performance between pre-training with MAE and no pre-training

- Methodology:
  - Pre-training MAE with ViT-B backbone for 600 epochs and fine-tuning for 100 epochs
  - Training ViT-B from scratch for 200 epochs
- Dataset: Tiny ImageNet
  - 100,000 images of 200 classes (500 for each class) downsized to 64 x 64 coloured images
  - Smaller and fewer images as compared to ImageNet

Reproduction

Main Experiment: Comparison of image classification performance between pre-training with MAE and no pre-training

- Training Details:
  - Image Patch Size: 4
  - Masking Ratio: 0.75

<table>
<tr><th></th><th></th><th>ViT from scratch
(our implementation)</th><th>Baseline MAE</th></tr>
<tr><td rowspan="2">Pre-Training</td><td>Number of Epochs</td><td rowspan="2">NA</td><td>600</td></tr>
<tr><td>Loss</td><td>MSE Loss</td></tr>
<tr><td rowspan="2">Main Training/
Fine-Tuning</td><td>Number of Epochs</td><td>200</td><td>100</td></tr>
<tr><td>Loss</td><td colspan="2">Categorical Cross Entropy</td></tr>
</table>

Reproduction

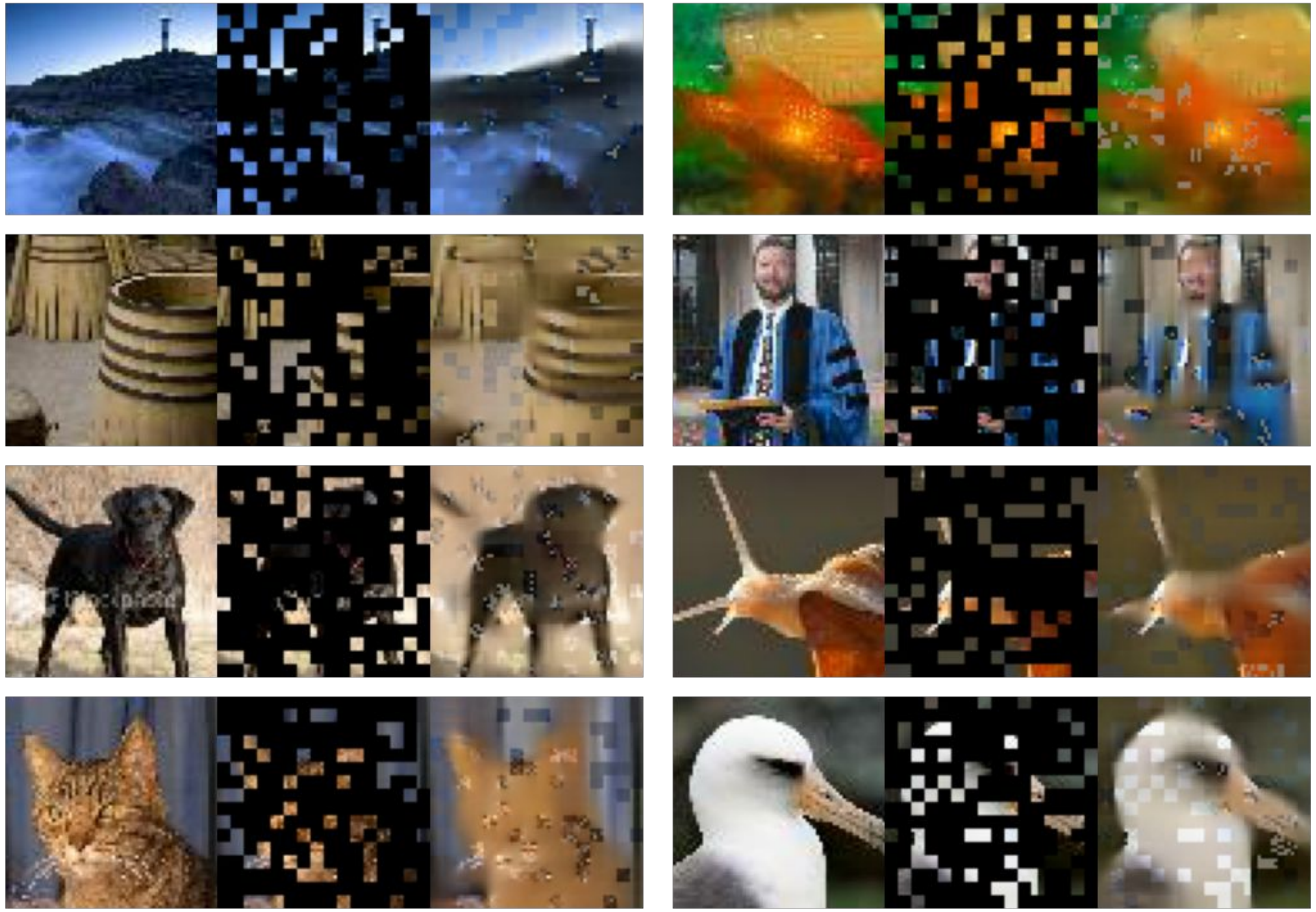


Figure 1: Example results of image reconstruction using MAE architecture

Reproduction

Main Experiment: Comparison of image classification performance between pre-training with MAE and no pre-training

- Results:

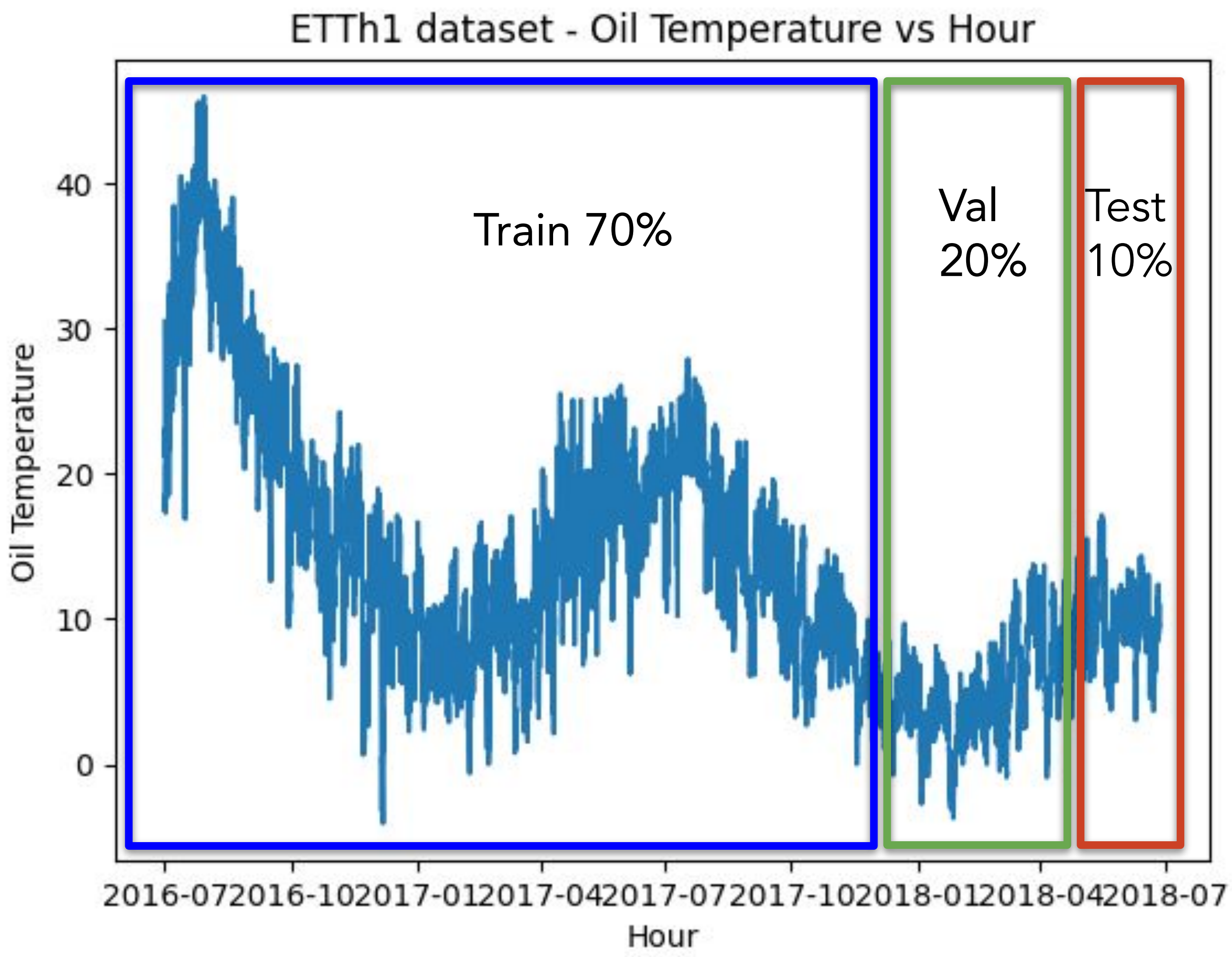
ViT from scratch (our implementation)	Baseline MAE
37.7%	45.6%

Extensions

Time Series Forecasting

Extensions of MAE: Time Series Forecasting

Can MAE improve forecast accuracy?



ETTh1 dataset

- Oil Temperature and Power Load data of 2 Electricity Transformers in China
- Target: Oil Temperature (OT) for univariate forecasting
- Features: Previous timesteps OT
- 17320 hourly timesteps

Extensions of MAE: Time Series Forecasting

Can MAE improve forecast accuracy?

Dataset Preprocessing

- shifted_df using <u>lookback = 100</u>
- train-val-test split: 70-20-10
- no random shuffling to maintain order

<u>Original Dataframe</u>

	date	OT
0	2016-07-01 00:00:00	30.531000
1	2016-07-01 01:00:00	27.787001
2	2016-07-01 02:00:00	27.787001
3	2016-07-01 03:00:00	25.044001
4	2016-07-01 04:00:00	21.948000

<u>Shifted Dataframe; lookback=3</u>

N timesteps

date	OT	OT(t-1)	OT(t-2)	OT(t-3)
2016-07-01 03:00:00	25.044001	27.787001	27.787001	30.531000
2016-07-01 04:00:00	21.948000	25.044001	27.787001	27.787001
2016-07-01 05:00:00	21.174000	21.948000	25.044001	27.787001
2016-07-01 06:00:00	22.792000	21.174000	21.948000	25.044001
2016-07-01 07:00:00	23.143999	22.792000	21.174000	21.948000
...	...	...	...	...

y_train
shape: (N,1)

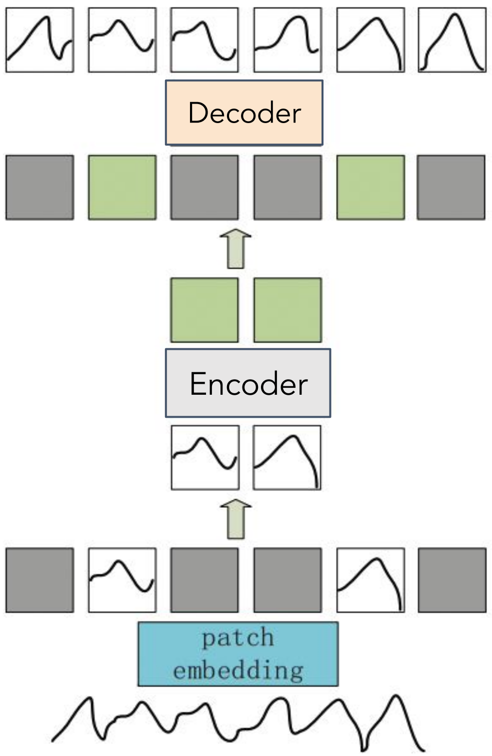
X_train
shape: (N,3,1)

1 since univariate
time series

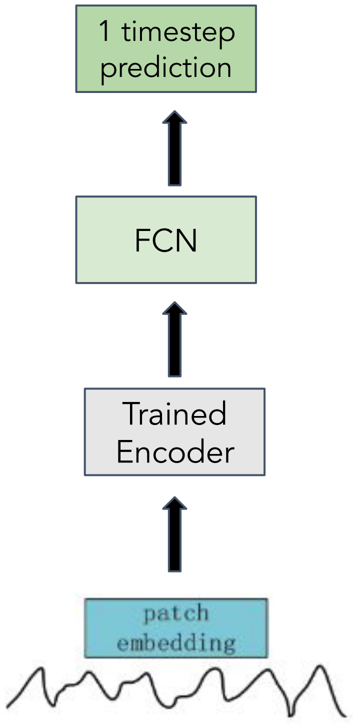
Extensions of MAE: Time Series Forecasting

Model Architecture and Evaluation Metrics

Pretraining



Finetuning



$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

References:

(1) TI-MAE: Self-supervised Masked Time Series Autoencoders

(2) MTSMAE: Masked Autoencoders for Multivariate Time-Series Forecasting

(3) Time Series Forecasting with Masked Autoencoder

Extensions of MAE: Time Series Forecasting

Results

Main Experiments

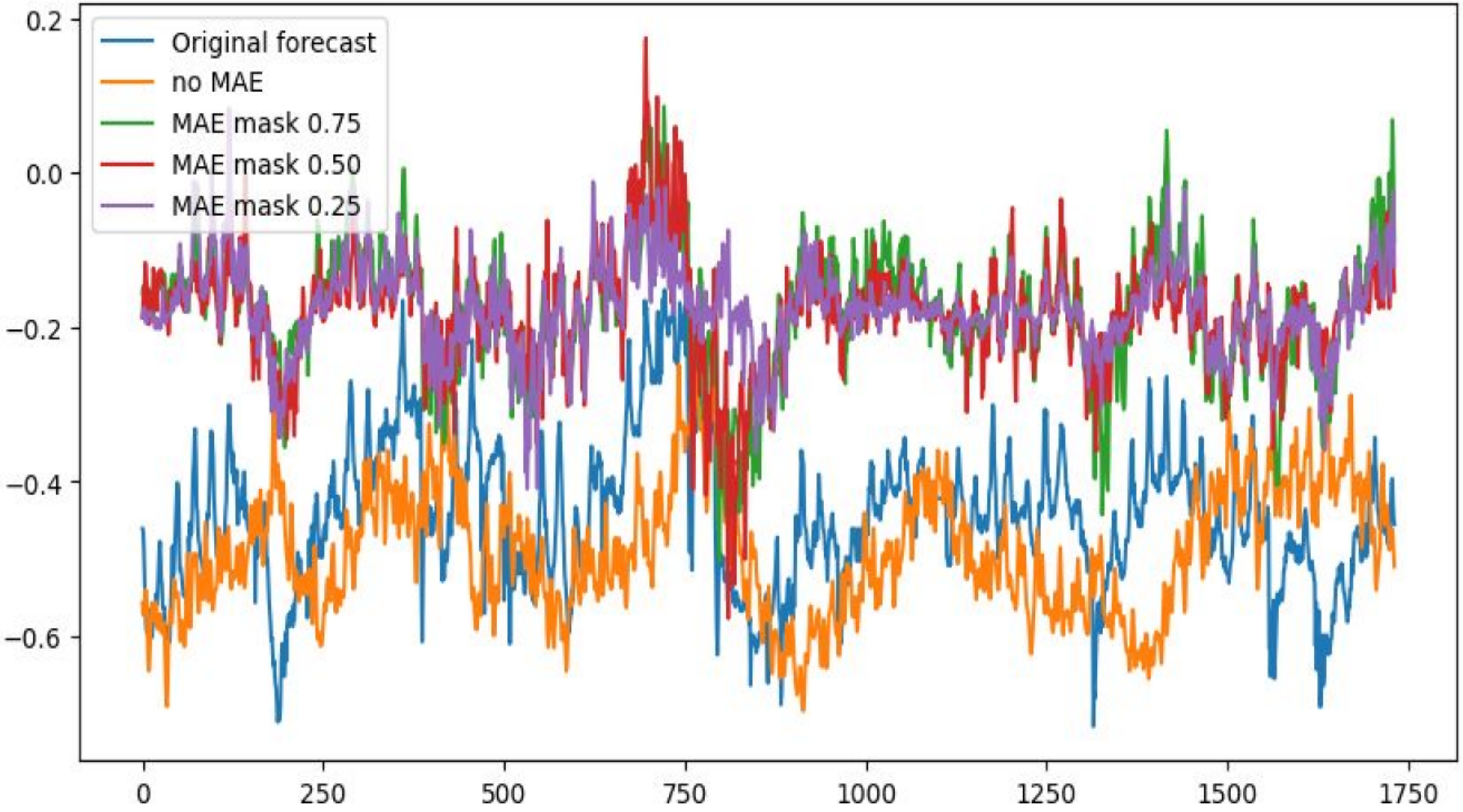
Models	Mean Sqr. Err.	Mean Abs. Err.
no MAE	0.01538	0.10043
MAE, msk=0.75	0.07757	0.26845
MAE, msk=0.50	0.07546	0.26391
MAE, msk=0.25	0.07778	0.26851

- MAE does not improve forecasting accuracy
- Models trained with different masking ratios exhibit similar performance
- Possible explanations
  - 1D time series signals are sparse
  - Neighbouring signal points are crucial for forecasting
  - Dataset is too small → transformers are data-hungry models

Extensions of MAE: Time Series Forecasting

Results

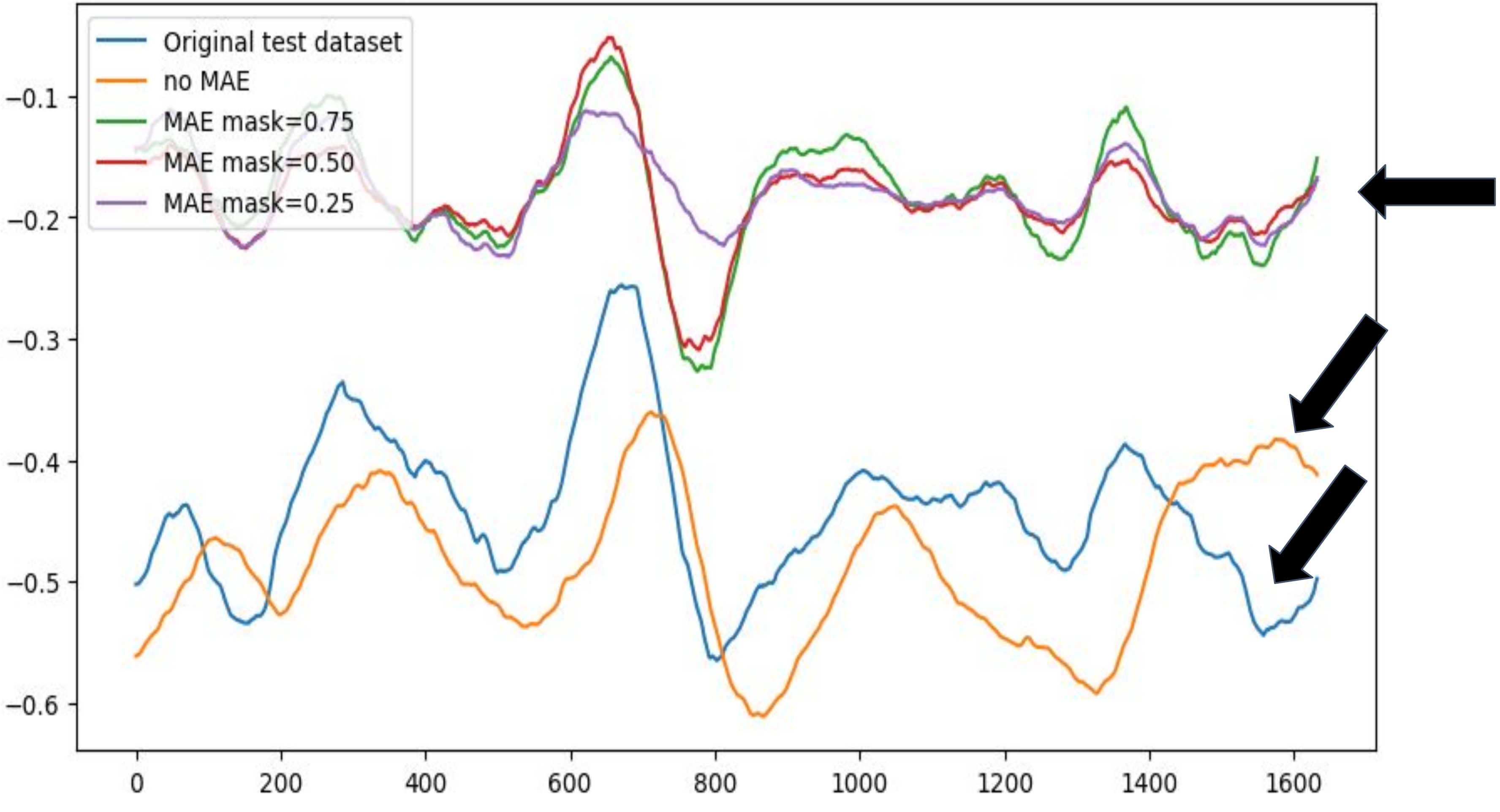
Original Test Dataset Plot



Extensions of MAE: Time Series Forecasting

Results

Smoothed Test Dataset Plot



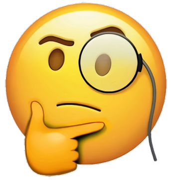
Extensions of MAE: Time Series Forecasting

Results

<u>Additional Experiments</u>

Models	Mean Sqr. Err.	Mean Abs. Err.
no MAE	0.01538	0.10043
RNN	0.00072	0.01822
LSTM	0.00071	0.01799
GRU	0.00071	0.01791

significant improvement



- Interesting insights:
  - Shallower architectures → better forecasting accuracy
  - Suggests that the dataset used was small
- Future work:
  - Use a larger dataset with more timesteps
  - Multivariate forecasting → increased information density

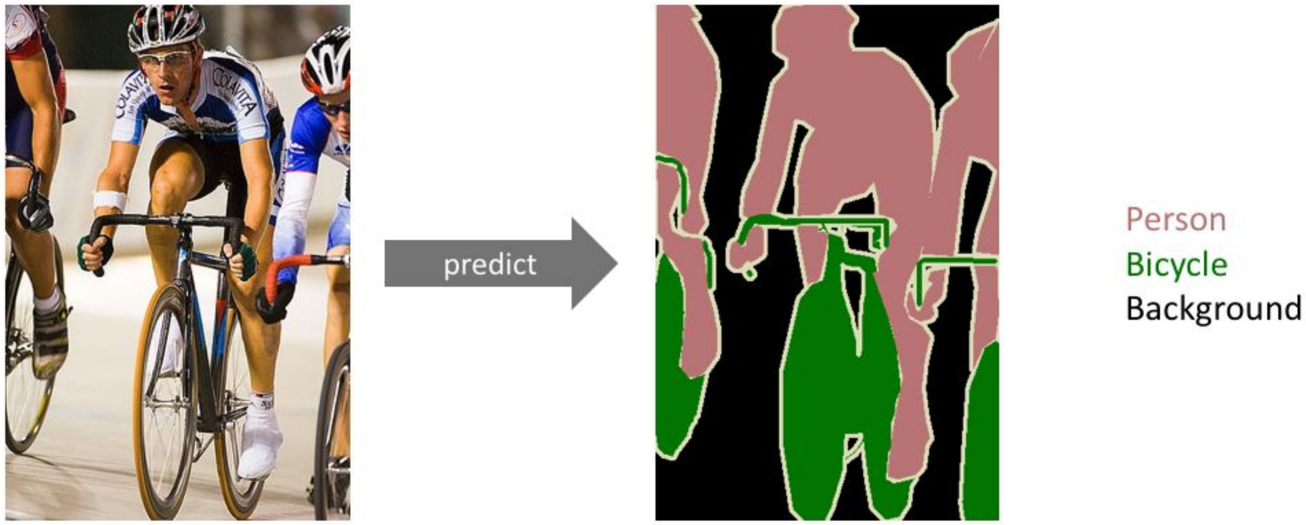
Extensions

Semantic Segmentation

Extensions of MAE: Semantic Segmentation

What is semantic segmentation

- Classifying every pixel of an input image to 1 of n semantic classes
- Supervised learning task with images with labelled pixels as target output

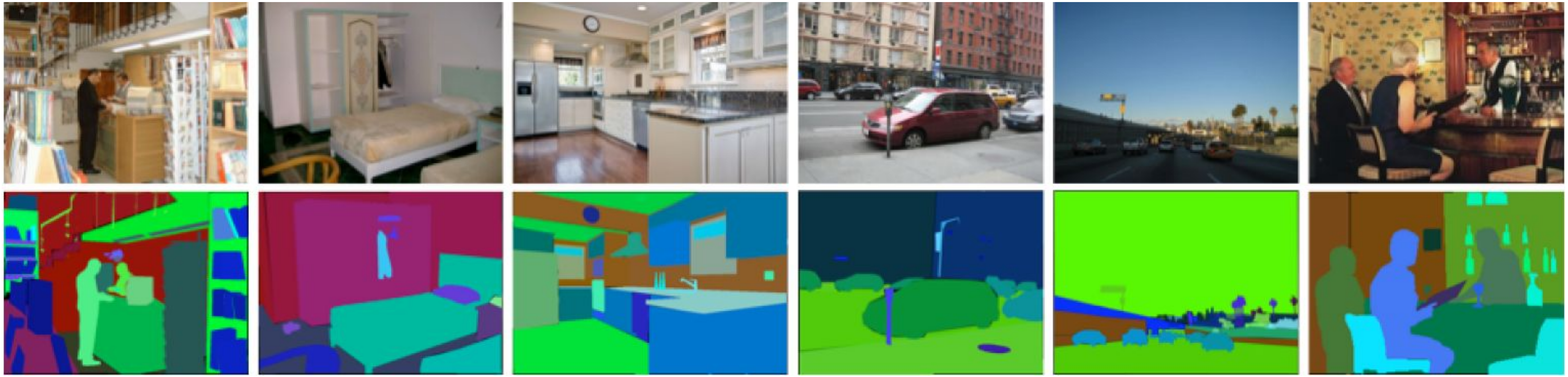


https://www.jeremyjordan.me/semantic-segmentation/

Extensions of MAE: Semantic Segmentation

Dataset - ADE20K

- Object segmentation exhaustively labelled manually (20,210 training data images, 3,169 semantic labels)

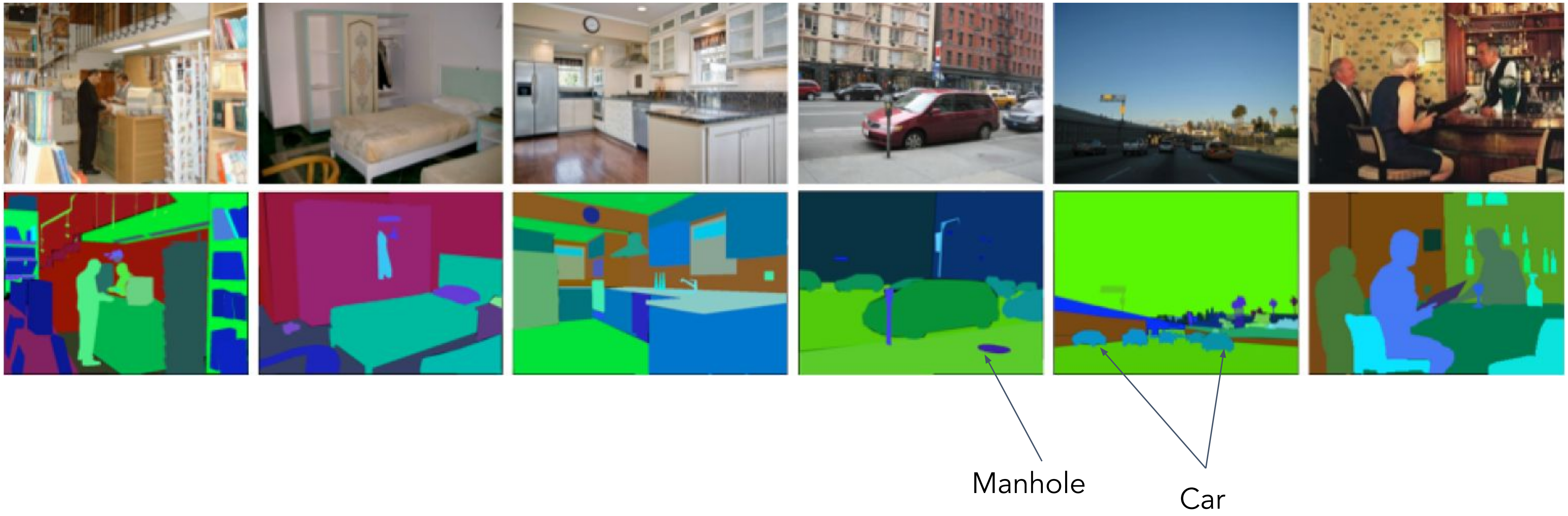


https://groups.csail.mit.edu/vision/datasets/ADE20K/

Extensions of MAE: Semantic Segmentation

Dataset - ADE20K

- Object segmentation exhaustively labelled manually (20,210 training data images, 3,169 semantic labels)

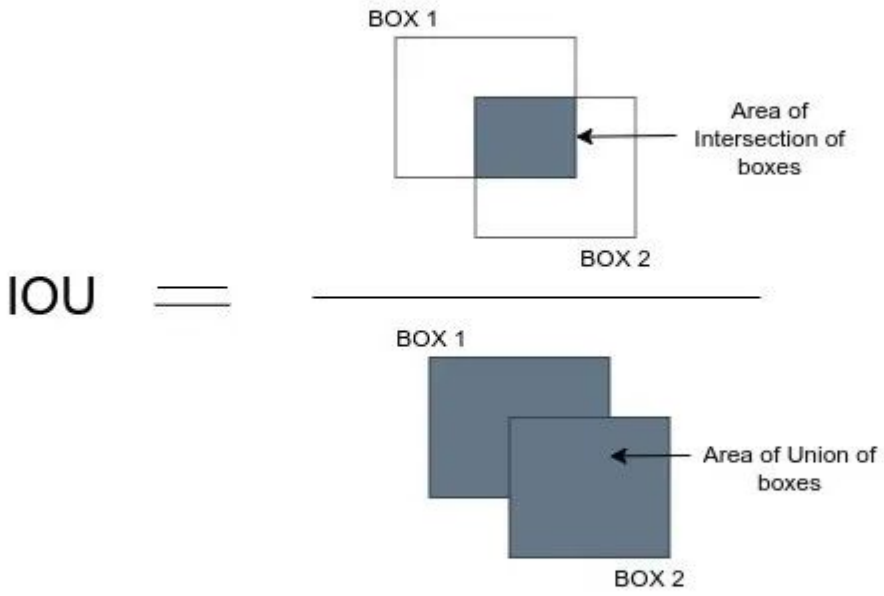


https://groups.csail.mit.edu/vision/datasets/ADE20K/

Extensions of MAE: Semantic Segmentation

Metrics

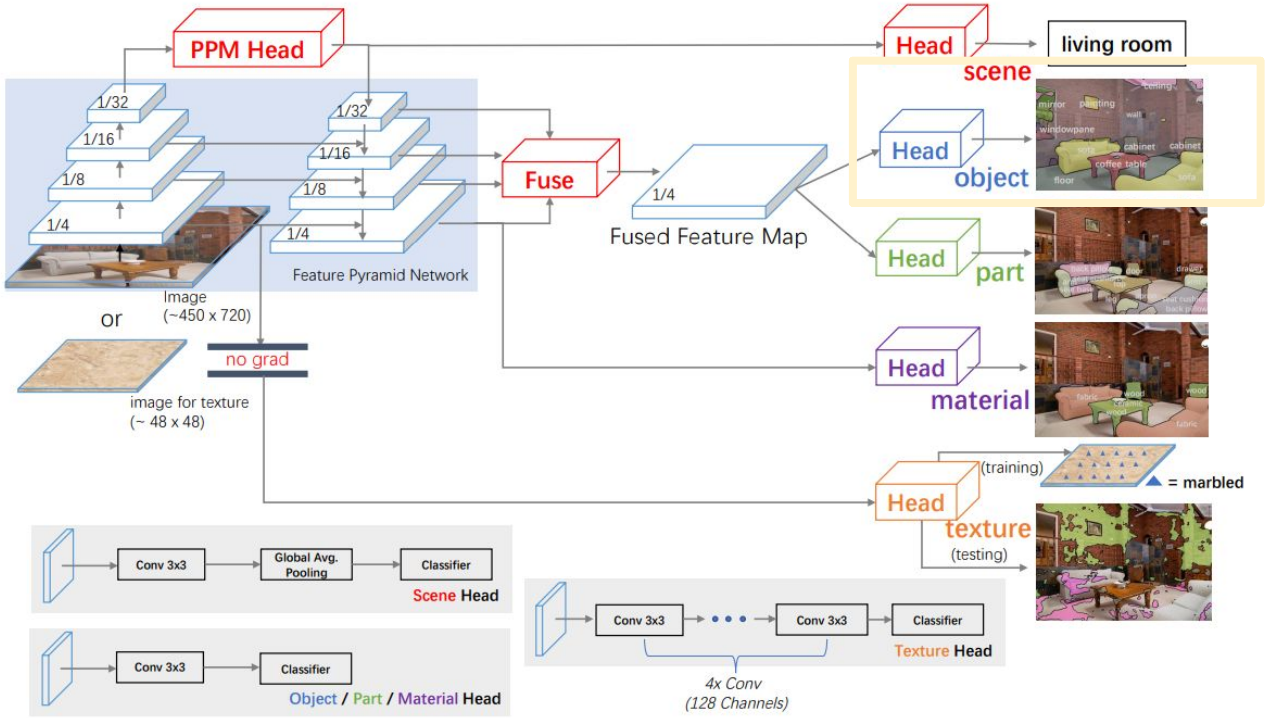
- Pixel accuracy
- MIoU (Mean Intersection over Union)
    - Area of overlap / area of union



https://medium.com/analytics-vidhya/iou-intersection-over-union-705a39e7acef

Extensions of MAE: Semantic Segmentation

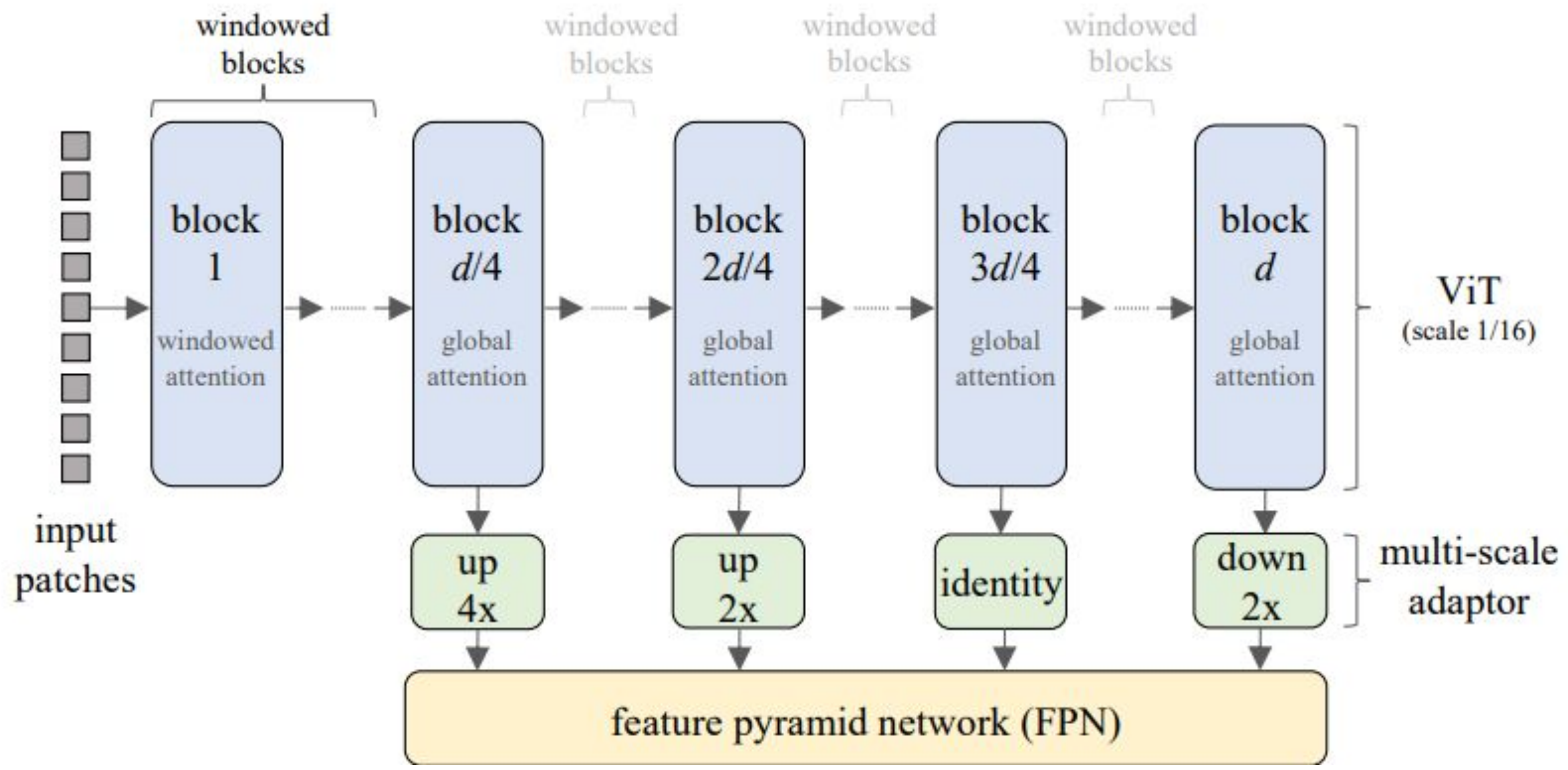
Model architecture - UperNet (Unified Perceptual Parsing Network)



https://arxiv.org/pdf/1807.10221.pdf

Extensions of MAE: Semantic Segmentation

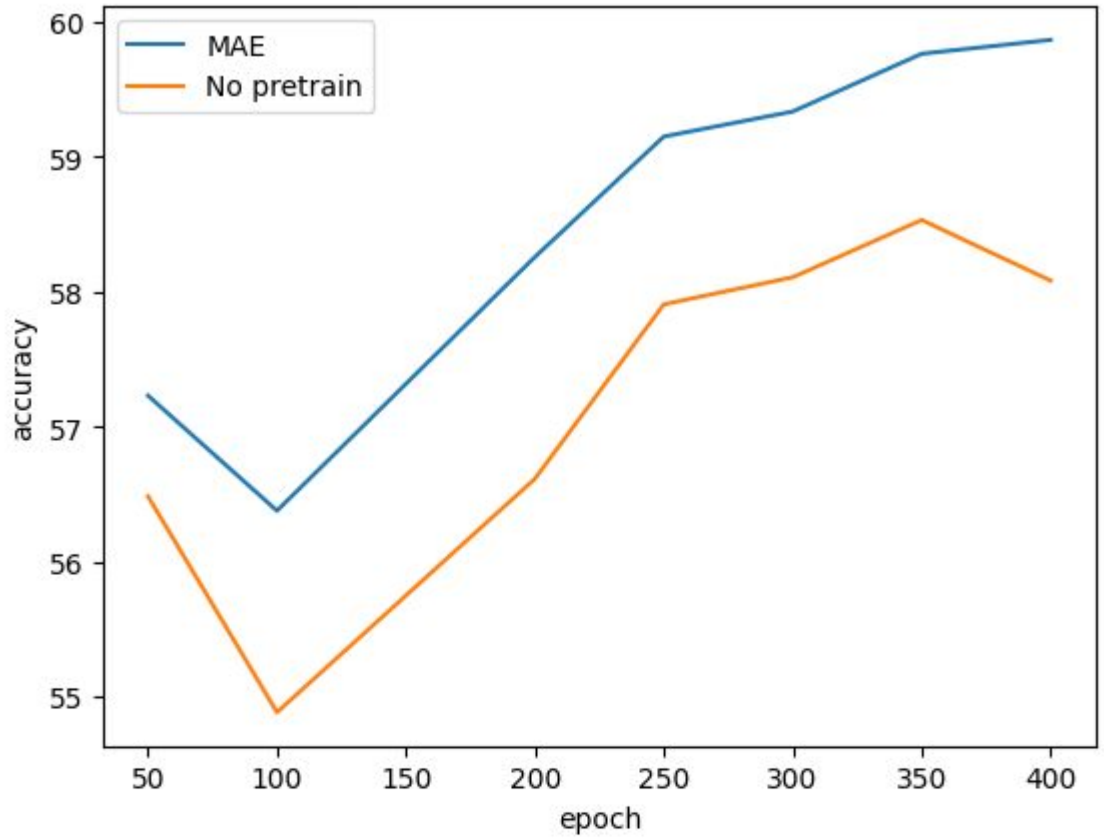
Model modification



https://arxiv.org/pdf/2111.11429.pdf

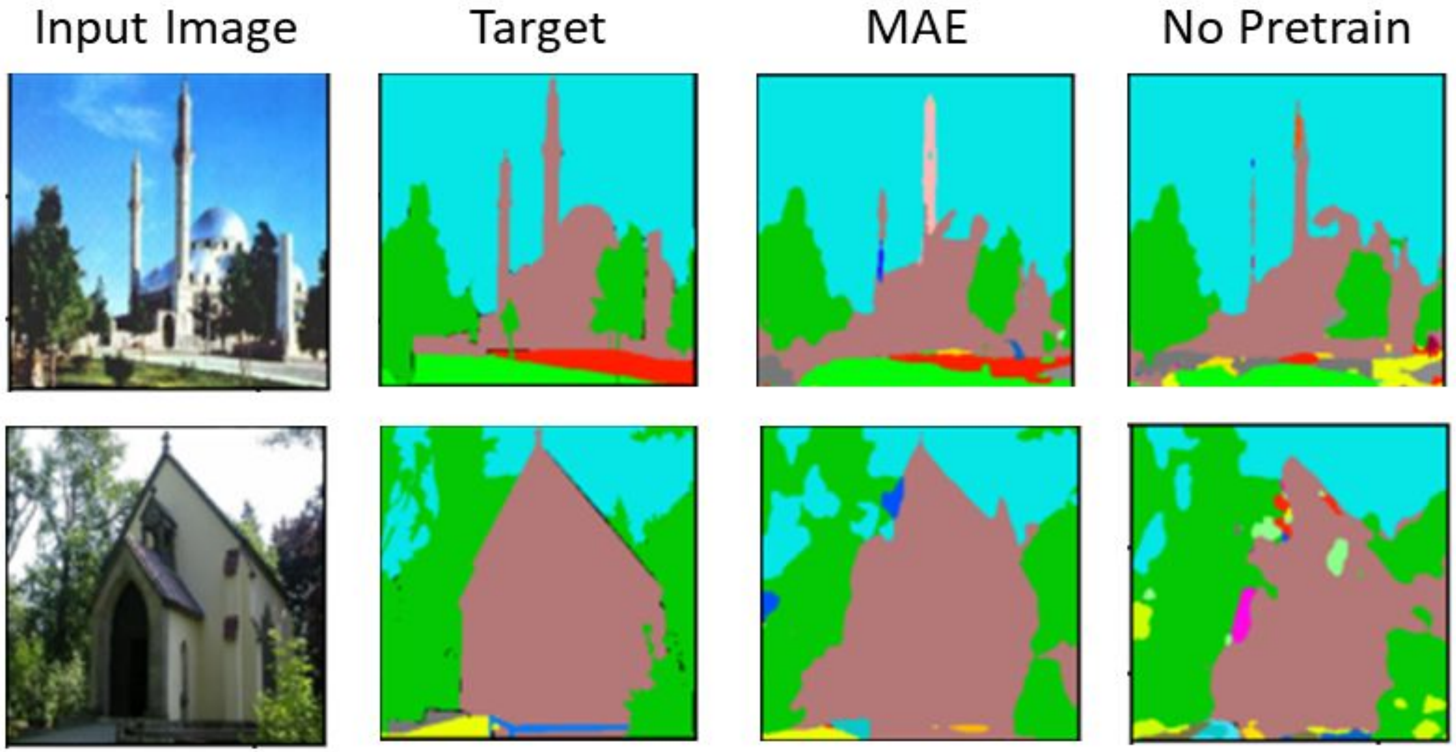
Extensions of MAE: Semantic Segmentation

Results



Extensions of MAE: Semantic Segmentation

Results



Extensions of MAE: Semantic Segmentation

Results

- Learns features of images using available unmasked patches
- Such features might be useful in recognizing objects and classifying them in semantic segmentation

Metric	MAE	No pretrain
Accuracy	59.86	58.08
MIoU	0.288	0.280

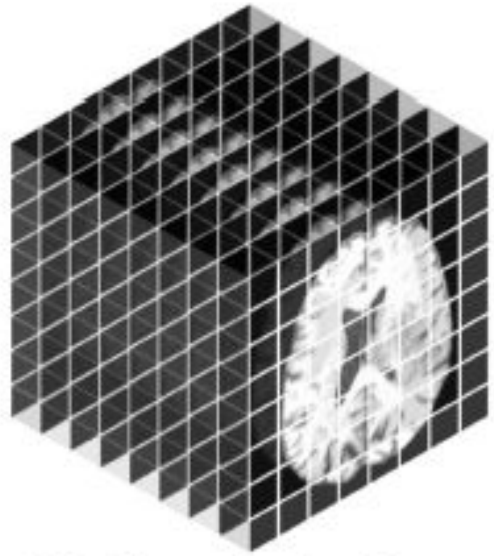
Extensions

3D Volumetric Medical Segmentation

Extensions of MAE: 3D Volumetric Medical Segmentation

What is 3D Volumetric Semantic Segmentation?

- Same as semantic segmentation, but in 3D.
- Since MAE pretraining gave better results for semantic segmentation, we can try extending this to the 3D images (volumes).
- We will look specifically at the task of medical image segmentation.

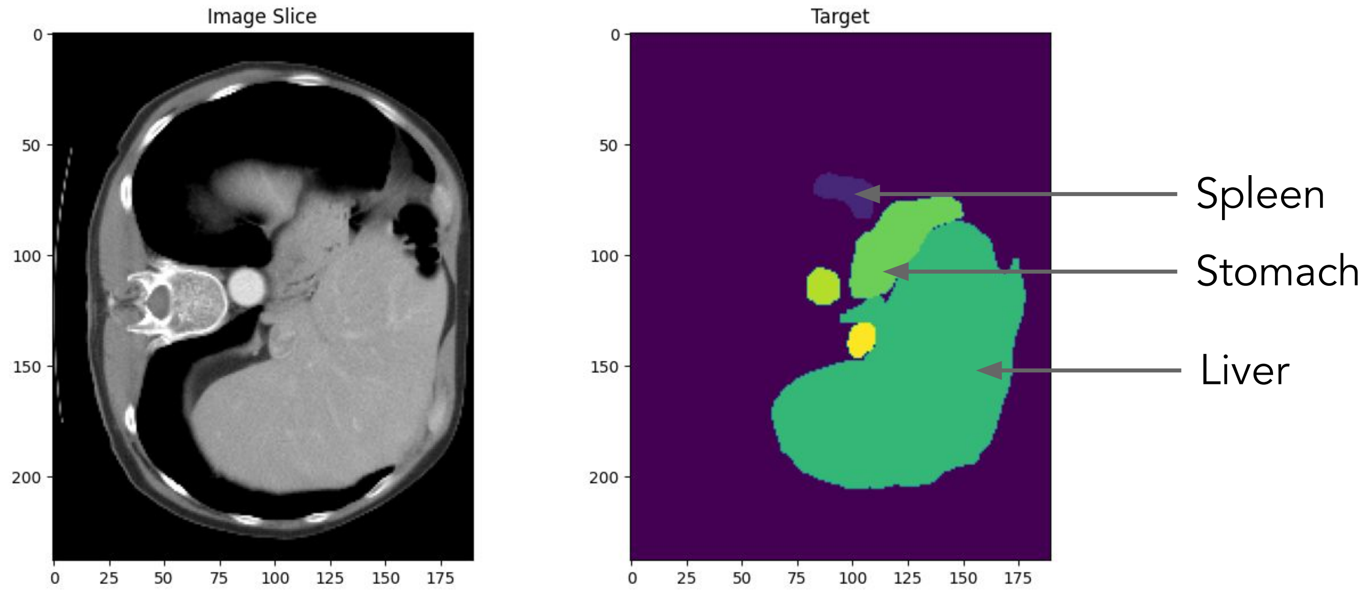


Note:
(1) Review - Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images, Sik-Ho Tsang.

Extensions of MAE: 3D Volumetric Medical Segmentation

Dataset

- Multi Atlas Labeling, Beyond the Cranial Vault (BTCV) dataset (1).
- 50 CT Scan Volumes, 30 of which are labelled.
- Aim to segment 13 different organs.



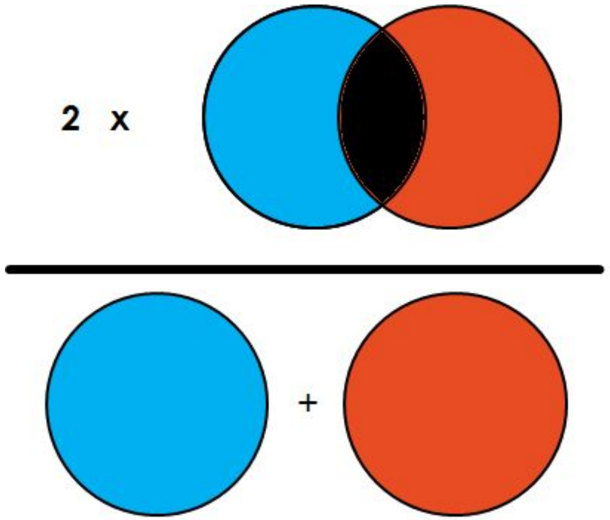
Note:
(1) https://www.synapse.org/#!Synapse:syn3193805

Extensions of MAE: 3D Volumetric Medical Segmentation

Evaluation Metric

- Dice Score
  - Measures the extent of overlap between target and predicted segmented results.

$$\frac{2 \times |X \cap Y|}{|X| + |Y|}$$

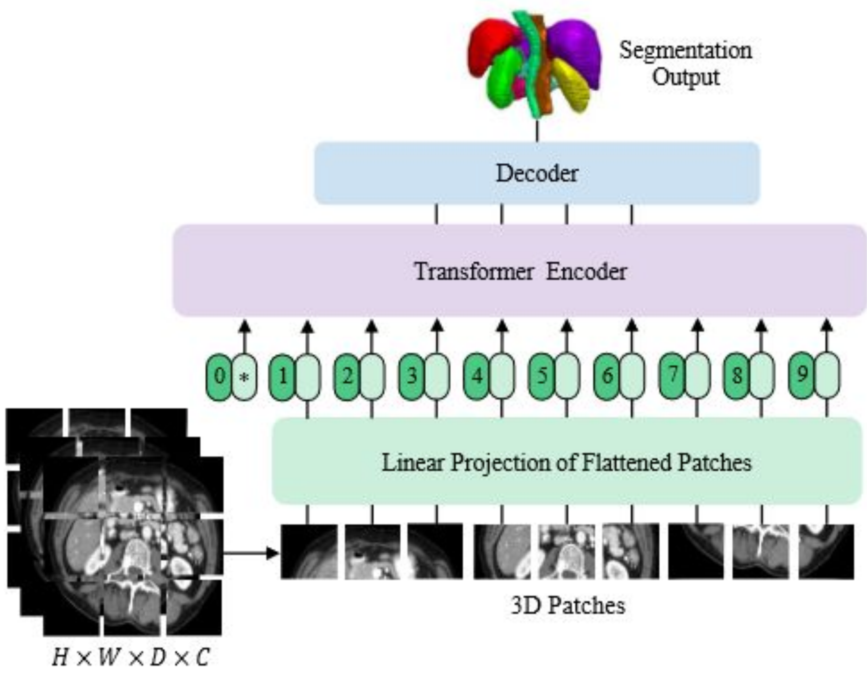


Note:
(1) https://www.kaggle.com/code/yerramvarun/understanding-dice-coefficient

Extensions of MAE: 3D Volumetric Medical Segmentation

Model

- UNEt TRansformers (UNETR) (1).
- Uses Vision Transformers to learn long-range dependencies and capture global context.



Vision Transformer
(Pretrain this using MAE)

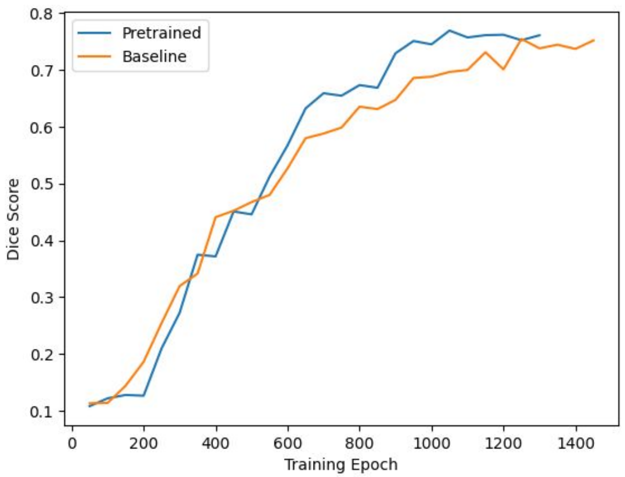
Note:
(1) UNETR: Transformers for 3D Medical Image Segmentation

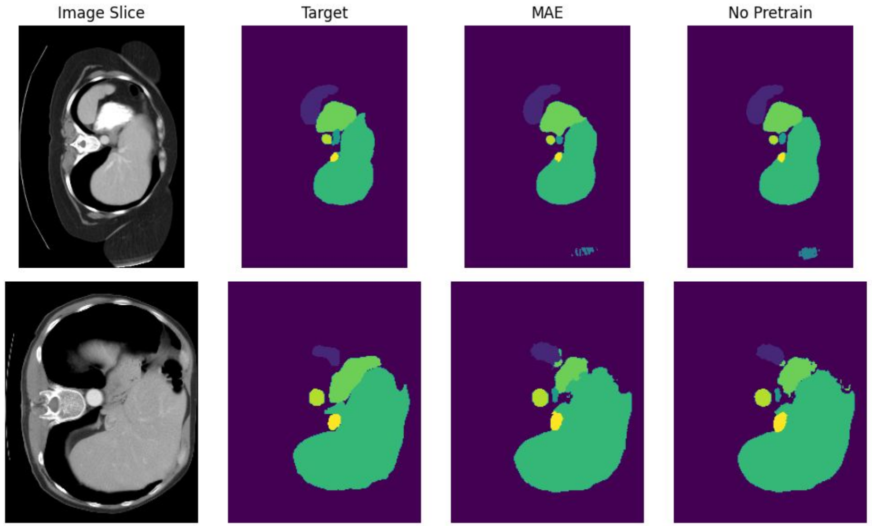
Extensions of MAE: 3D Volumetric Medical Segmentation

Results

- Pretraining with MAE allows for slight better results, in fewer number of epochs.
- Results can potentially be better if pretrained on a larger dataset of volumes.

	Dice Score
Baseline (No pretraining)	0.7548
Pretrained with MAE	0.7712





Extensions

Data Imputation

Extensions of MAE: Data Imputation

The imputation of missing data

- Data imputation is similar to the image reconstruction task
- Except with a different data modality: numerical/text data

The hypothesis

- If MAE can capture a hidden representation of data within images, it should work on other data modalities as well
- Therefore, MAE can recreate samples similar to the original dataset

Extensions of MAE: Data Imputation

The Dataset

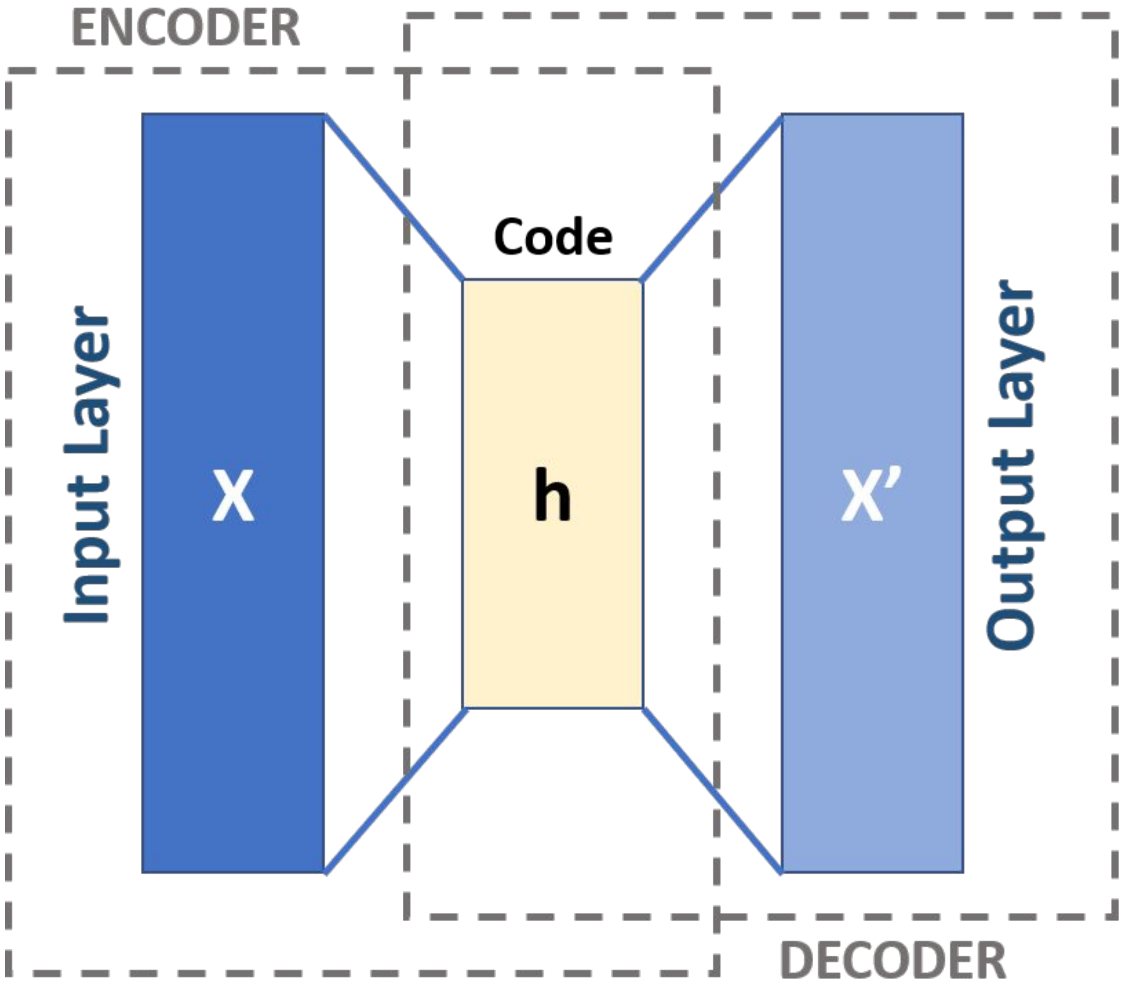
- California Housing Price, from the UCI Machine Learning repository
- The dataset contains information from the 1990 California census.
- There are 10 features, 1 of them is categorical variable while the rest are numerical variables.
- Data preprocessing includes encoding, scale normalization, manually simulating missing values, split of train and test datasets.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Extensions of MAE: Data Imputation

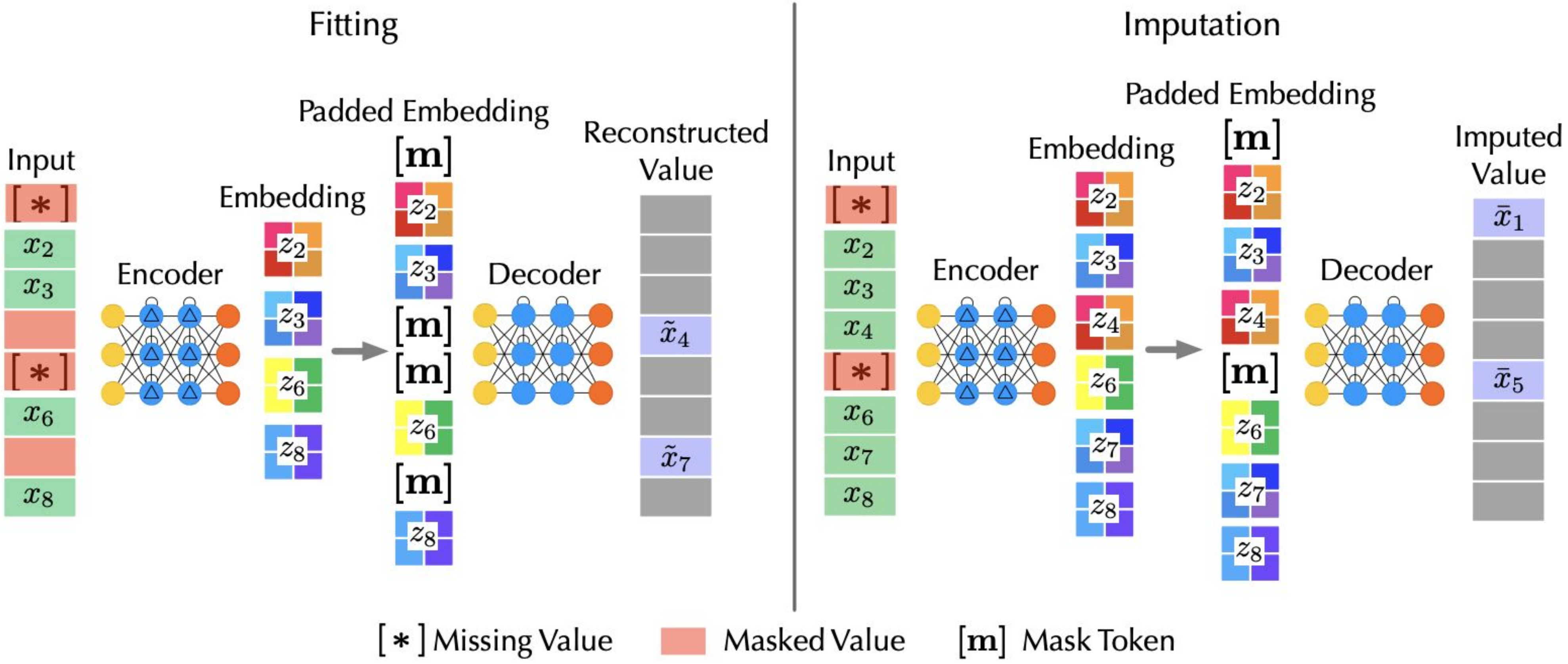
MAE Model Design

- Encoder:
  - takes in the concatenated input and mask
  - compresses the input into a dense representation
- Decoder:
  - expands the encoded representation
  - maps the output of the dropout layer back to the original input dimension
- Mask:
  - concatenate with the input along the feature dimension, forming a single input tensor that doubles the feature space
  - allows the model to learn not just from the data but also from the structure of its availability



Extensions of MAE: Data Imputation

Model Design



https://arxiv.org/pdf/2309.13793.pdf

Extensions of MAE: Data Imputation

Evaluation Metric

- MSE
    - the average squared difference between the estimated values and the actual value

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

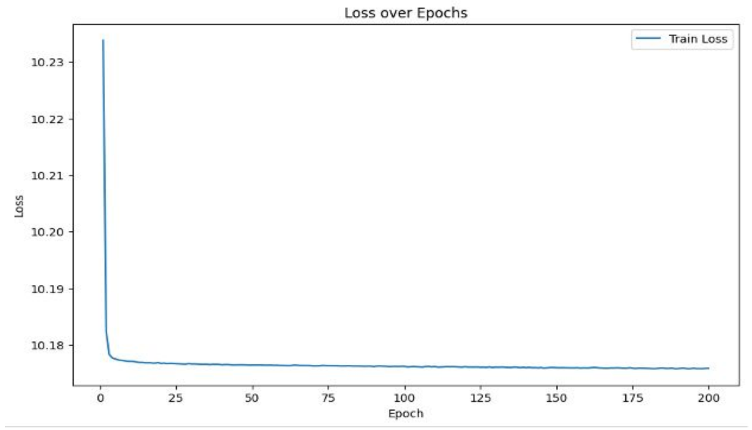
- Wasserstein distance
    - known as the Earth Mover's distance (EMD), is a measure of the distance between two probability distributions over a given metric space

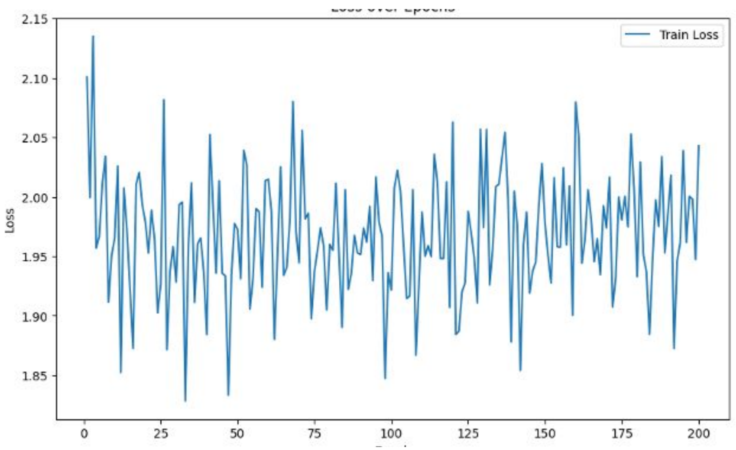
$$W_p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \left(\mathbf{E}_{(x,y)\sim\gamma}\, d(x,y)^p\right)^{1/p}$$

Extensions of MAE: Data Imputation

Results

- The training loss of baseline model decreases after a few number of epochs then converge while there are significant fluctuations of loss throughout the training process for MAE model without systematic stabilization.





- Under same number of epochs, baseline imputer model turns out to give better result, with lower MSE loss and Wasserstein distance

Evaluation metric	Baseline	MAE
MSE loss	0.0093	0.5171
Wasserstein distance	0.2482	0.3098

Extensions of MAE: Data Imputation

Interesting insights

- Data preprocessing: MinMax Scaler outperforms Standard Scaler
    - Preservation of sparse structure
    - Dominance of outlier sensitivity
    - More suitable for neural networks and gradient descent

- Hypothesis for ineffective application of MAE on imputation of missing values
    - Data structure – housing prices
    - Masking strategy
    - Overemphasis on masking

Conclusion

Conclusion

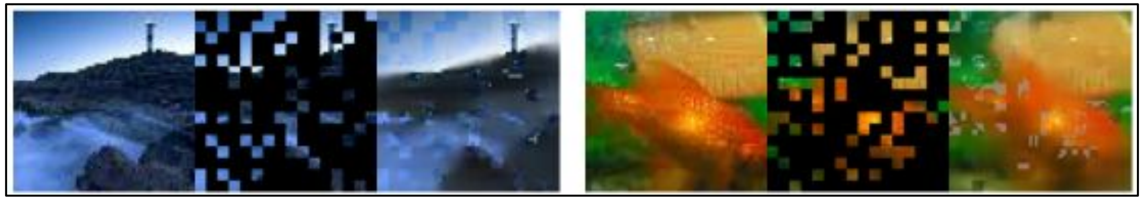
Masked Autoencoders

- In the paper “Masked Autoencoders are Scalable Vision Learners”
    - Hypothesized to capture a rich hidden representation of data
    - By masking out input datapoints

Conclusion

Masked Autoencoders

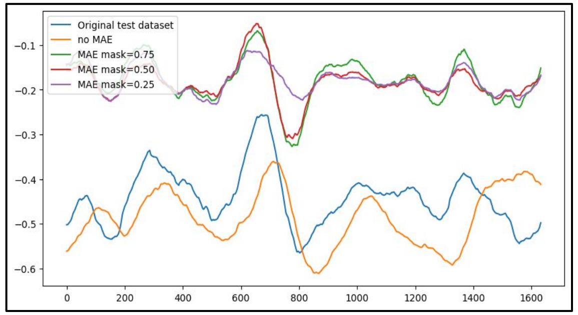
- In the paper “Masked Autoencoders are Scalable Vision Learners”
  - Hypothesized to capture a rich hidden representation of data
  - By masking out input datapoints
- Reproduction on the TinyImageNet
  - Able to recreate images from masked input

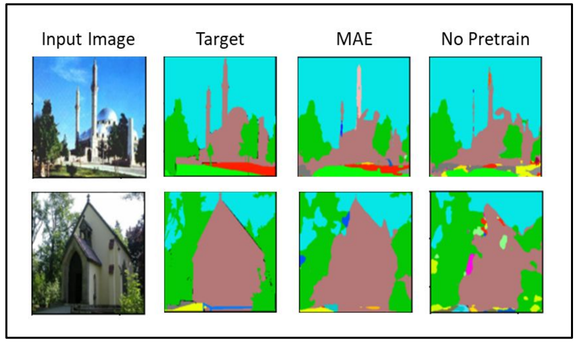


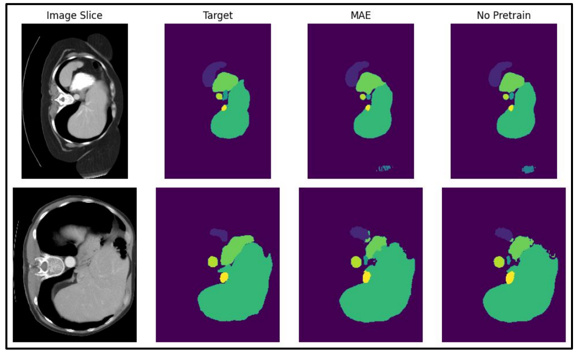
Conclusion

Masked Autoencoders

- In the paper “Masked Autoencoders are Scalable Vision Learners”
  - Hypothesized to capture a rich hidden representation of data
  - By masking out input datapoints
- Reproduction on the TinyImageNet
  - Able to recreate images from masked input
- Extension of MAE:
  - Semantic segmentation of images
  - 3D medical segmentation
  - Time series forecasting
  - Data imputation







Conclusion

Masked Autoencoders

- In the paper "Masked Autoencoders are Scalable Vision Learners"
  - Hypothesized to capture a rich hidden representation of data
  - By masking out input datapoints
- Reproduction on the TinyImageNet
  - Able to recreate images from masked input
- Extension of MAE:
  - Semantic segmentation of images
  - 3D medical segmentation
  - Time series forecasting
  - Data imputation
- Differences could be due to:
  - Favours image data and adverse to other modalities of data (text/time-series)
  - MAE may be able to capture more complex, spatial relationships between pixels
  - Other data modalities contains lesser spatial relationships which can be modelled using simpler models

End