INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Self-Supervised Learning

Tutorial 19-08-2021

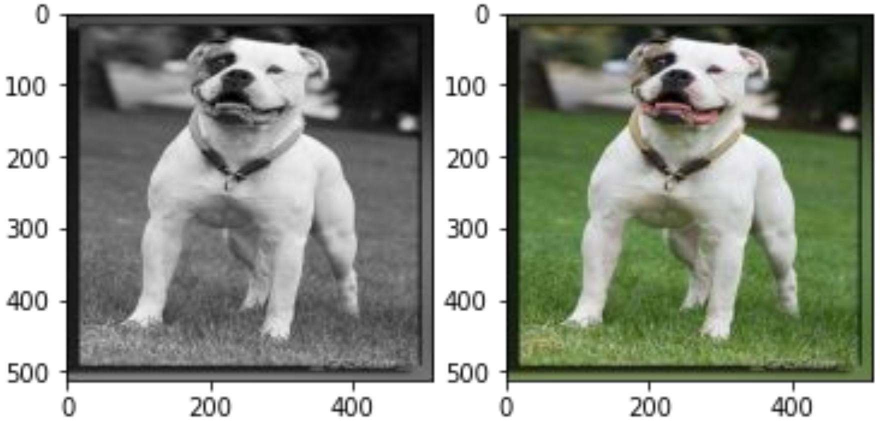
Supervised Learning

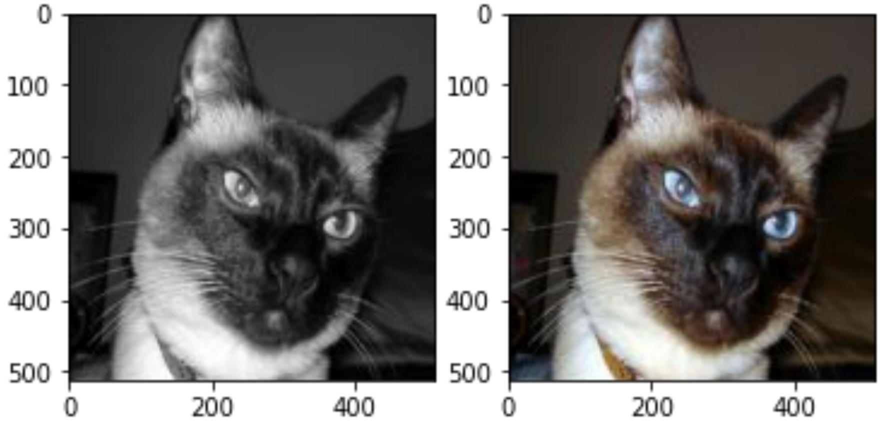
- The initial boost in the machine learning world came via the paradigm of supervised learning.
- In this setting, a model is trained for a specialized task for which the data is carefully labelled.
    - Bounding boxes for localization
    - Semantic maps for semantic segmentation, etc.
- Practically speaking, it’s impossible to label everything in the world.
- Unfortunately, this a limits how far the field of AI can go with supervised learning alone.

Basics of Self-Supervised Learning

- Data provides the supervision directly.
- In general, perturb the data and task a network to predict it back.
- We often may solve a proxy task using the network forcing it to learn meaningful semantics that can be used in downstream tasks.
- The proxy tasks are also often of great research importance in standalone form.
- Let us check some image level self-supervised learning tasks.

An Example: Image Colorization



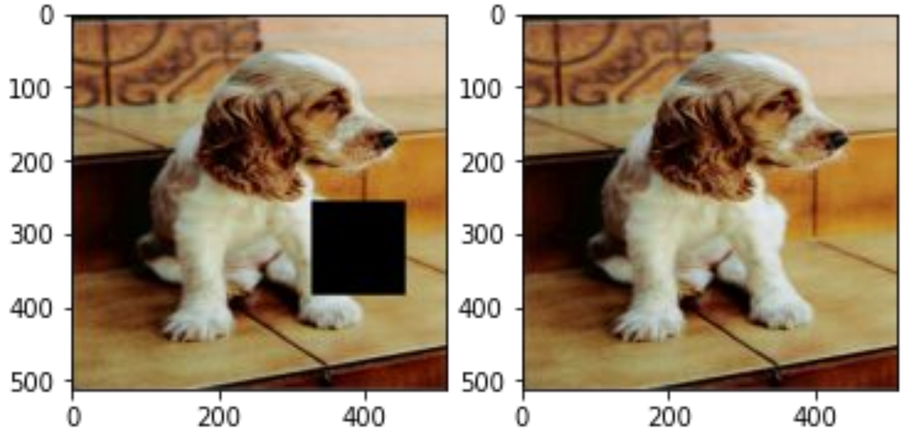


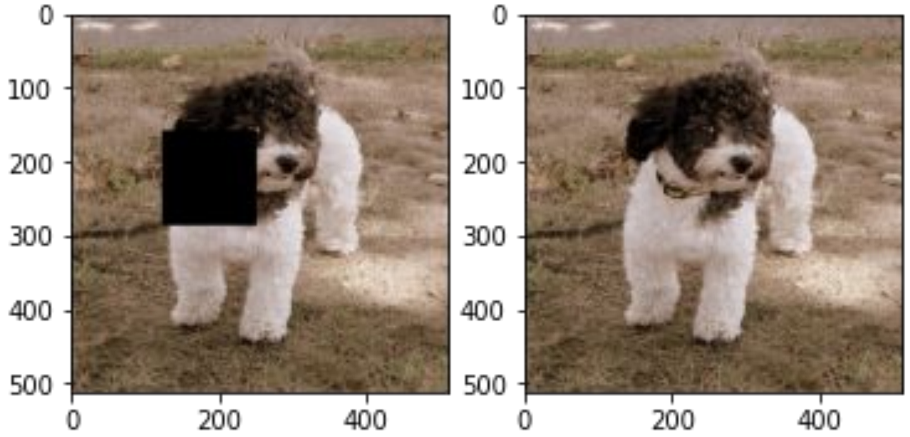
- Train a neural network to predict colours from a grayscale image.

- The network needs to inherently learn semantic boundaries present in the image, for example the shape of the foreground (dog), the background type etc.

- This semantic knowledge can be exploited in downstream tasks like semantic segmentation, a task which now needs a human to annotate every pixel present in an image.

An Example: Image Inpainting

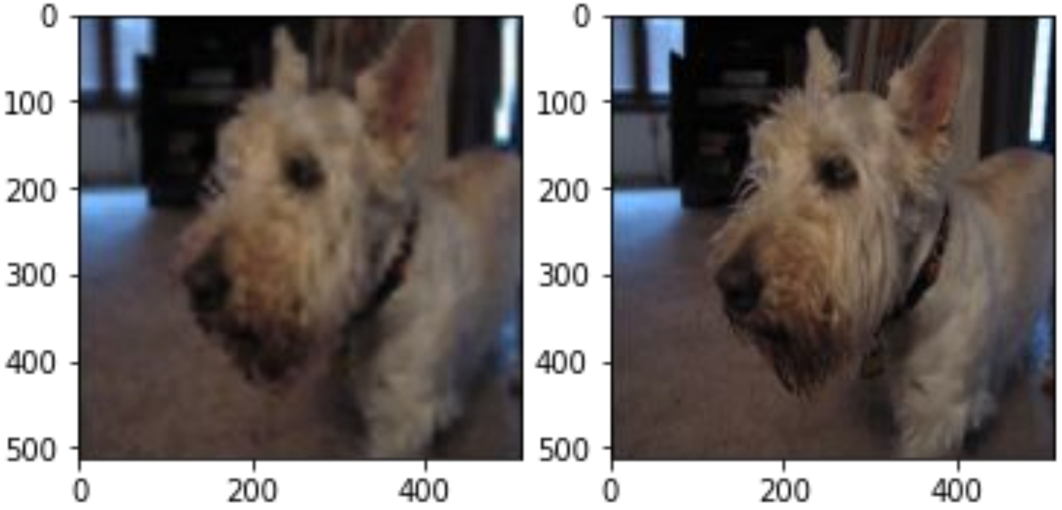


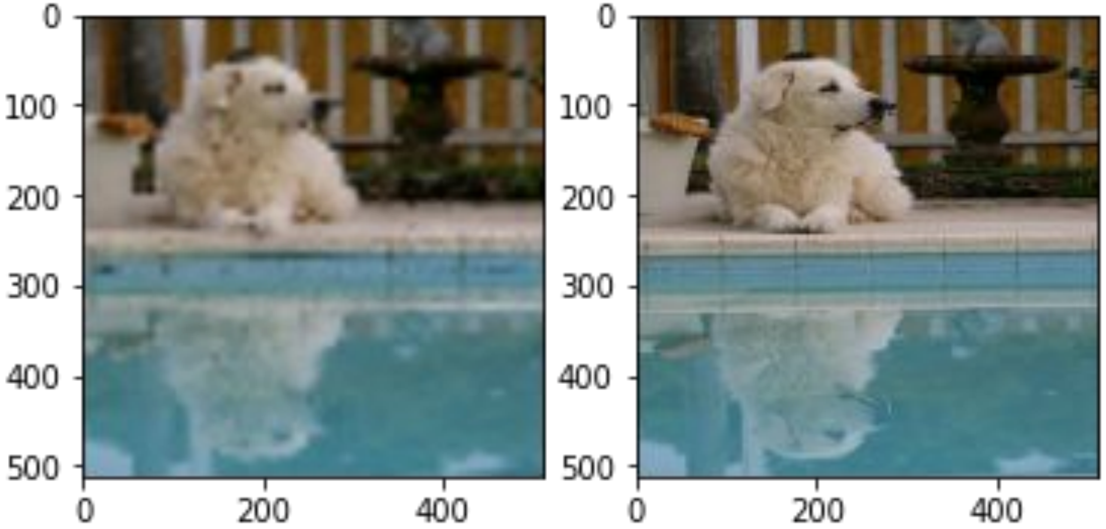


- Remove a particular area of an image randomly and ask a NN to predict it back.

- The network requires to understand the structure of objects present in the image to inpaint the required region

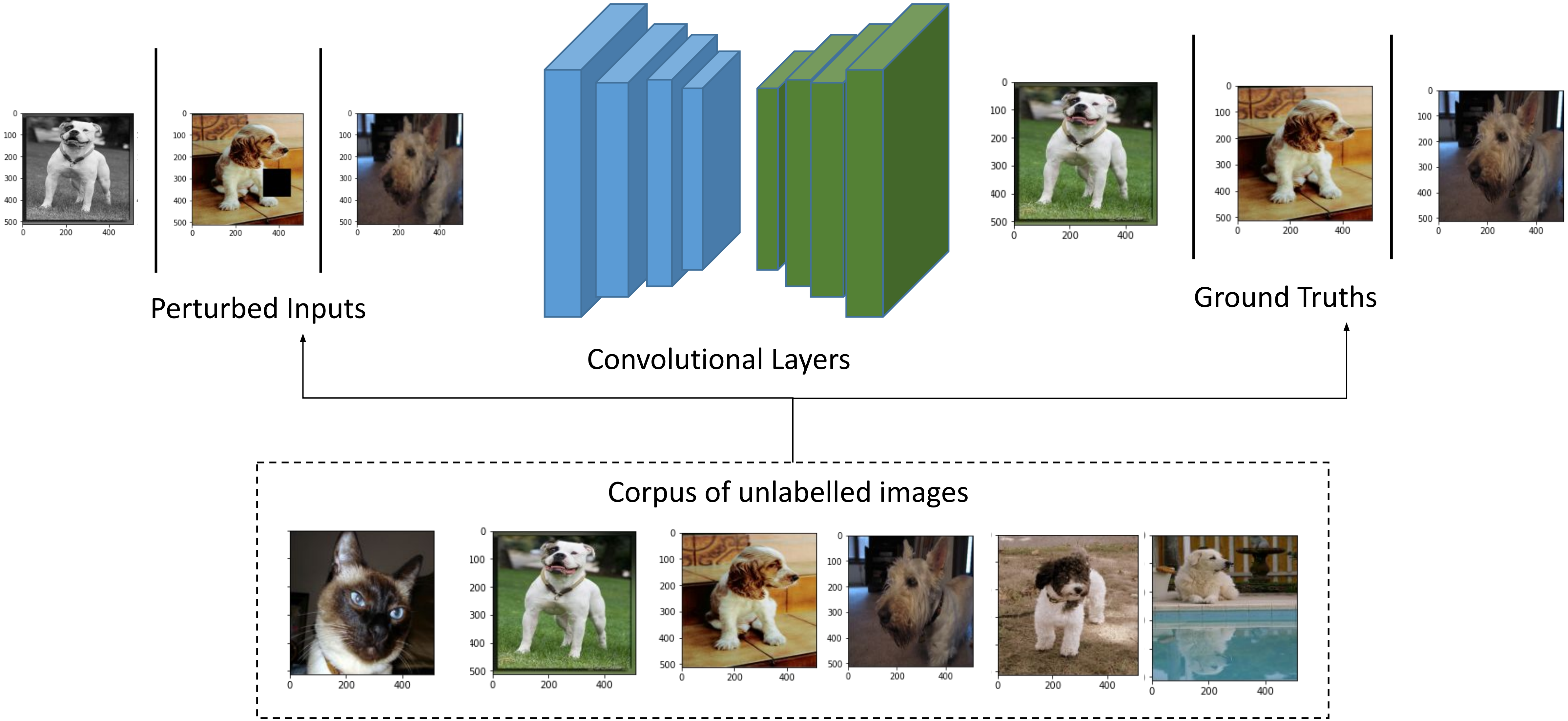
An Example: Image Super-resolution





- Predicting a higher resolution image from a lower input.
- We will be showing a code walk through for this particular topic in today’s session.

How are the networks trained?



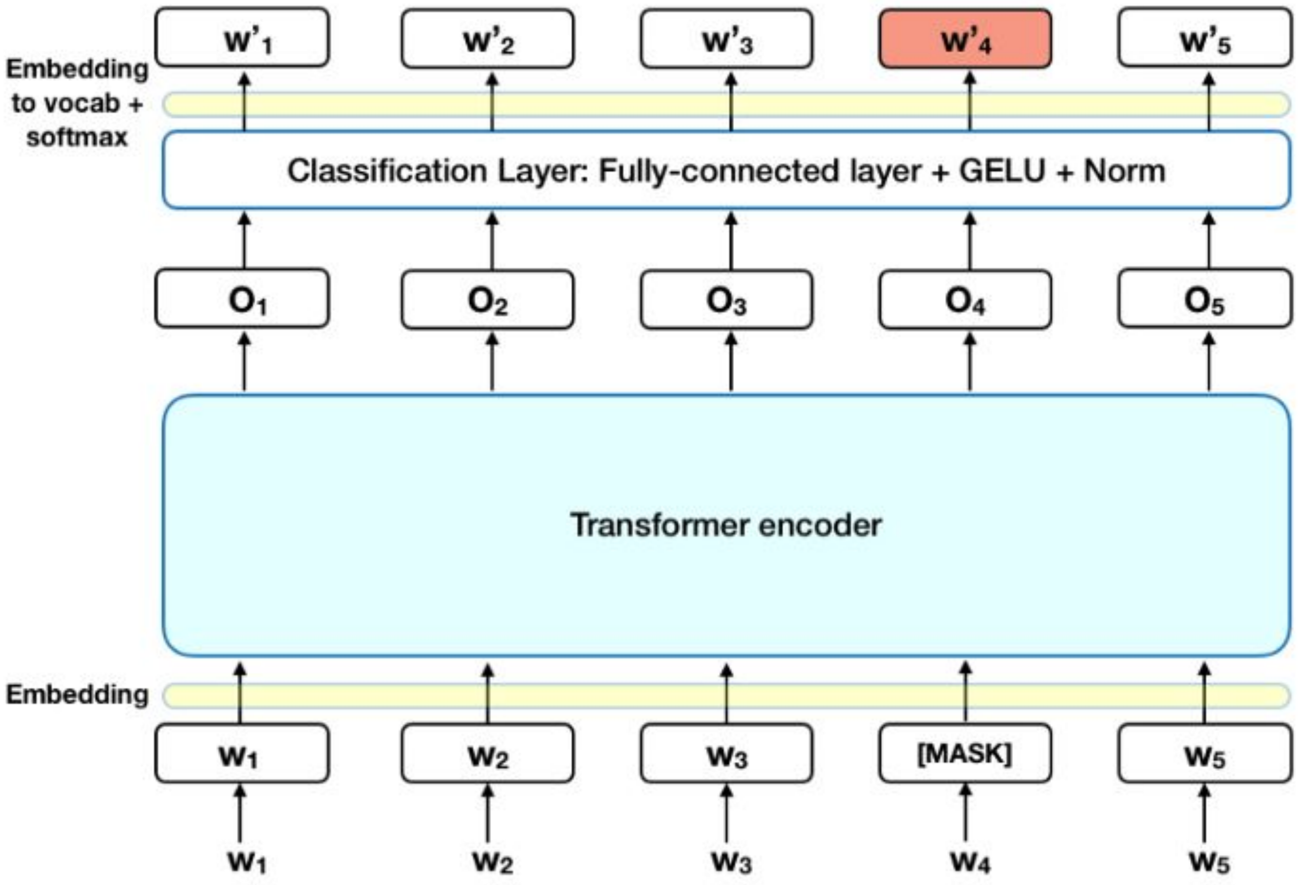
Perturbed Inputs

Convolutional Layers

Ground Truths

Corpus of unlabelled images

Similar approach in other fields:

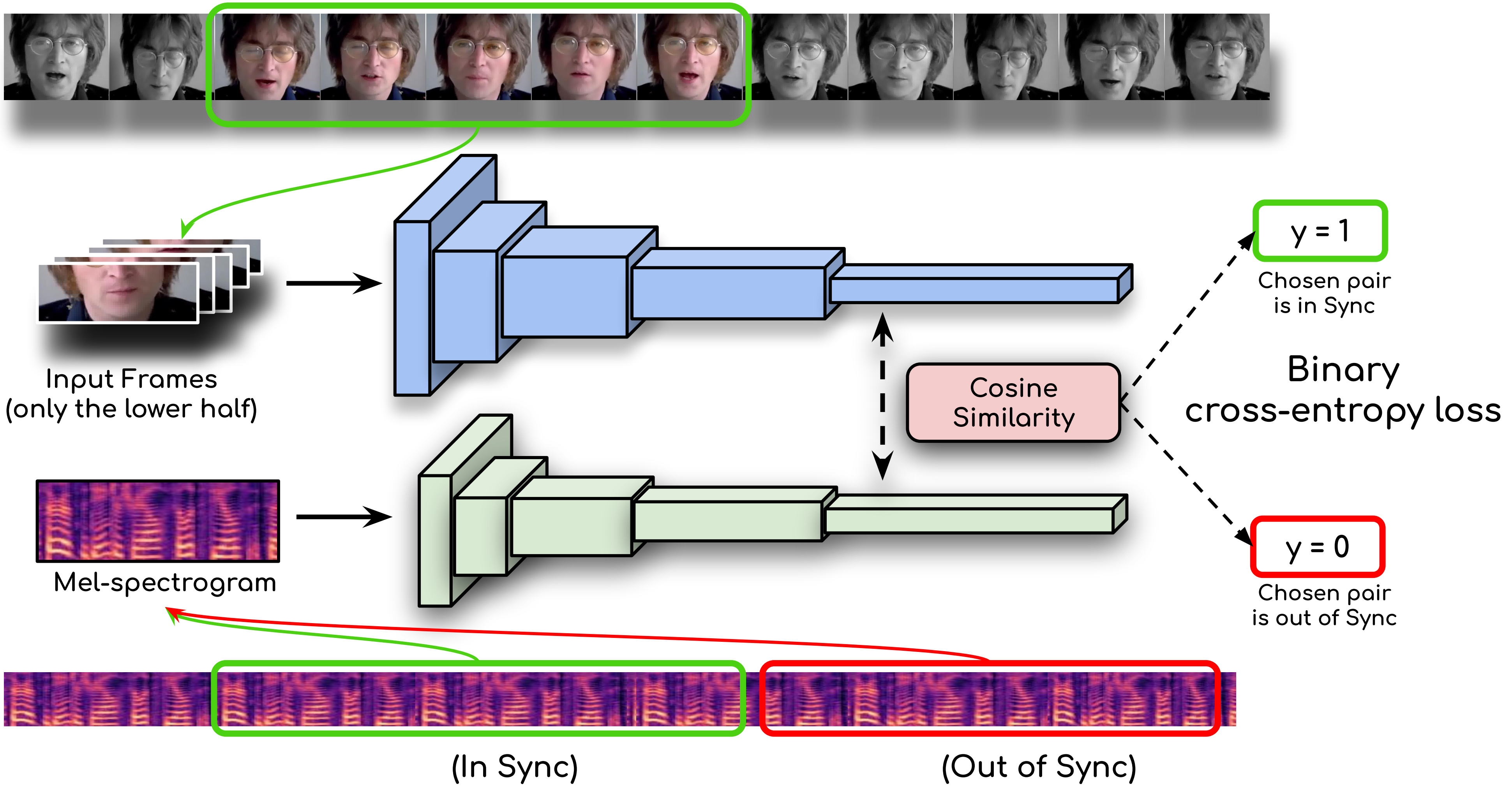


W'1
W'2
W'3
W'4
W'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
W1
W2
W3
[MASK]
W5
W1
W2
W3
W4
W5

- The BERT language model is also trained on a similar concept.
- Instead of image patch, we mask a word from a random sentence.
- A transformer based network is tasked to predict the masked word back learning rich semantics present in language.
- Wav2Vec also follows a similar principal for speech.

Image credit: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Audio-Visual Self Supervised Learning



Let us now go through a SR code:

- Please go to this repository: https://github.com/Rudrabha/SS2021-19-08-2021

- Open this notebook for the code walk through: https://github.com/Rudrabha/SS2021-19-08-2021/blob/main/Image_Super_Resolve_Tutorial.ipynb

- There are other two notebooks containing codes for Image inpainting and Image Colorization.

- Please note that these codes are for basic introduction and not meant for State-of-the-art uses in any of these problems. However, the building blocks of the network can be used to train much more complex networks.

- Please check Prof. Andrew Zisserman's slides: https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf for more insights.

Thank You!