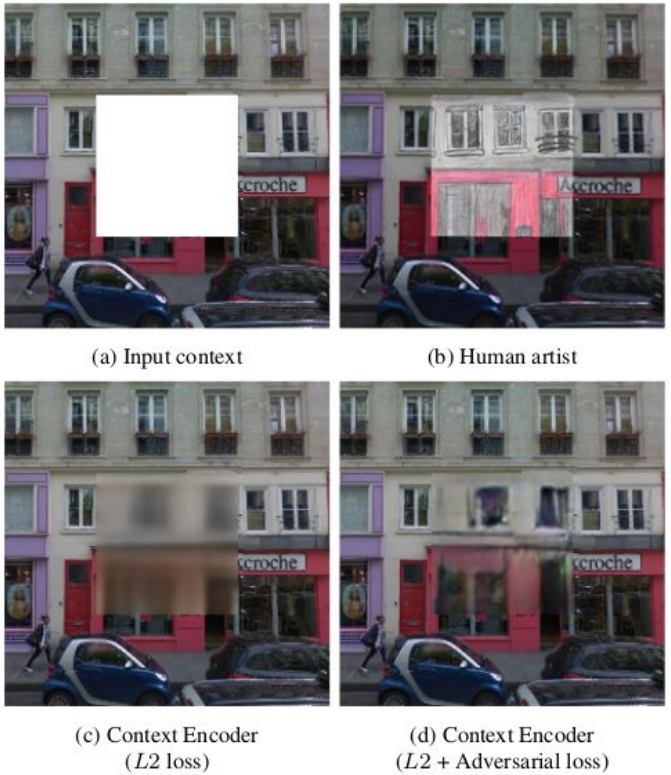
Context Encoders: Feature Learning by Inpainting

Team Name : Jaathirathnalu

Abhishek Reddy - 2018101028
Sreeharsha Paruchuri - 2018102002
Rahul Kashyap - 2018102037
Nihar Potturu - 2018102039

Main Objective

- The main objective of this paper is to reconstruct the missing part of an image, using Context Encoders – a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings.



(a) Input context

(b) Human artist

(c) Context Encoder ($L2$ loss)

(d) Context Encoder ($L2$ + Adversarial loss)

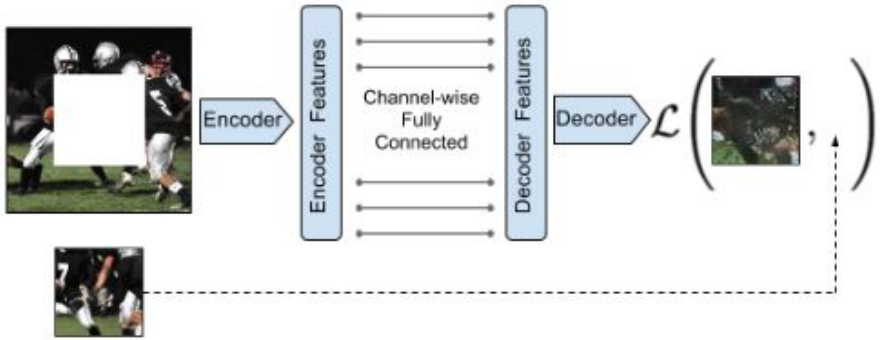
Related Previous Works

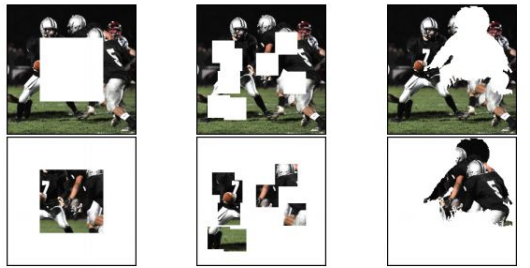
- The part of the image missing is too large for **classical inpaintings.**
- **Scene completion** can’t fill arbitrary holes in the image
- Previous implementations use **hand-crafted distance metric,** like Gist

We use learned distance metric which is superior to hand crafted distance metric.

Method overview :

- We use an encoder-decoder architecture.
- The output of the context-encoders is a feature vector.
- The feature vector is channel-wise connected to the decoder
- The decoder fills in realistic image content using the encoded feature vector
- This architecture is based on the renowned AlexNet Architecture





(a) Central region (b) Random block (c) Random region

Method overview :

The pipeline consists of three major stages:

- **Context Encoder**: Our model is not trained for ImageNet classification; rather, the network is trained for context prediction “from scratch” with randomly initialized weights.
- **Channel-wise fully connected layer**.
- **Decoder**: It generates pixels of the image using the encoder features.

Loss function:

- **Reconstruction Loss:** We use a normalized masked L2 as our reconstruction loss function
- **Adversarial Loss:** Loss based on GAN’s

Significance of AlexNet

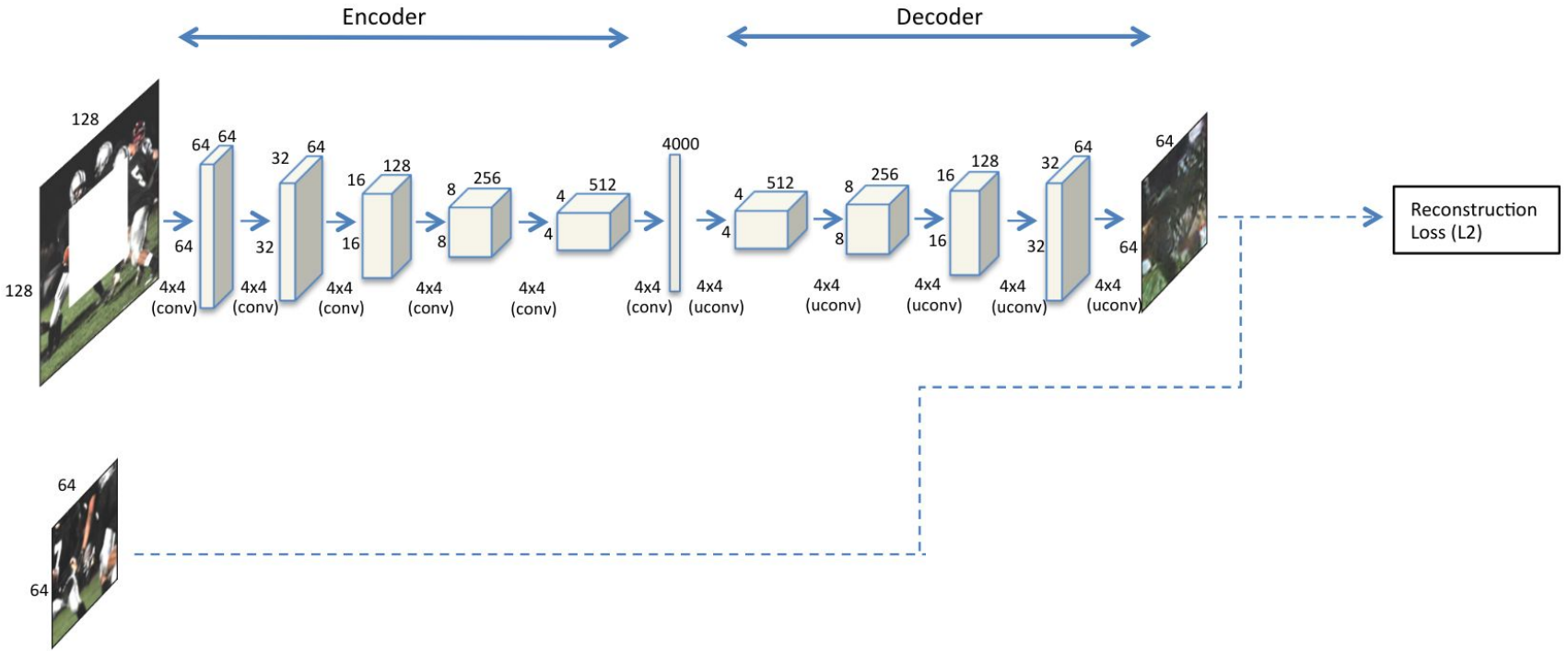
Alex net is a Deep Learning Architecture that consists of :

- 5 Convolutional Layers
- 3 Fully Connected Layers

What makes this architecture special is its outstanding performance on the ImageNet image classification dataset.

This architecture shows good performance in other tasks involving extracting features from scenes

Unlike the actual AlexNet, our modal isn't trained for the ImageNet dataset.


Encoder
Decoder
128
128
64
64
64
64
32
64
32
16
128
16
8
256
8
4
512
4
4000
4
512
4
8
256
8
16
128
16
32
64
32
64
64
4x4
(conv)
4x4
(uconv)
Reconstruction
Loss (L2)

What has been implemented

- The alexnet architecture has been incorporated to create the encoder.
- Decoders is also made on a similar line
- Only L2 loss has been used while training the network

The dataset that we used is a set of images of the paris street view

The number of images are 6392 each of 227x227x3 size

Left to be implemented

- Create a Discriminator network which can enable us to implement adversarial calculations
- Train this discriminator on the real images and the fake images to make the discriminator better
- Incorporate the Adversarial loss into the loss of the network

Results

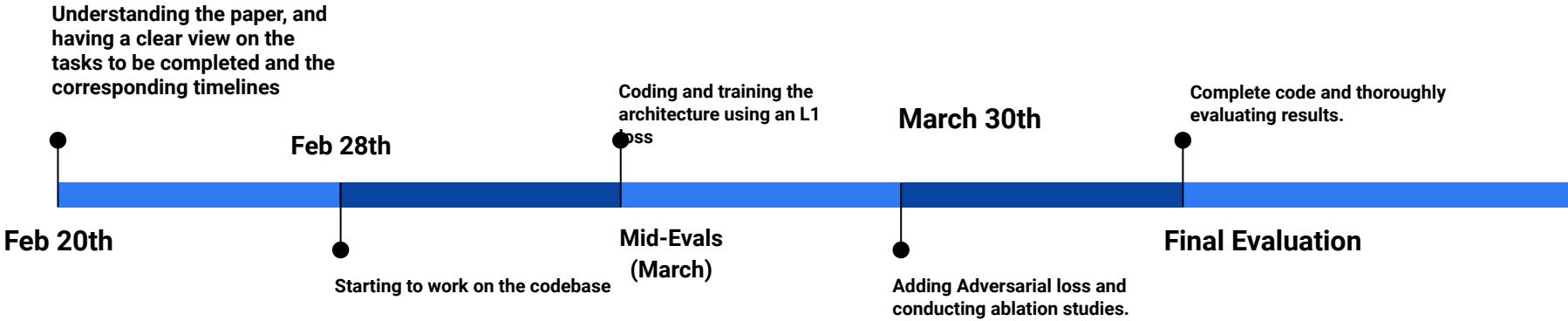
Cropped Image

Real Image

Output after 50 epochs

Paper implementation
(L2 loss)

Timeline


Understanding the paper, and having a clear view on the tasks to be completed and the corresponding timelines
Feb 20th
Feb 28th
Starting to work on the codebase
Coding and training the architecture using an L1 loss
Mid-Evals (March)
March 30th
Adding Adversarial loss and conducting ablation studies.
Complete code and thoroughly evaluating results.
Final Evaluation