Project Proposal - ECE 176

Ethan Morales
Cognitive Science ML
A17909364

Jiseon Kim
Electrical Engineering(Machine Learning and Controls)
A18088614

Abstract

We propose implementing an advanced image inpainting model inspired by Context Encoders: Feature Learning by Inpainting. Our focus is on training a convolutional neural network (CNN) to reconstruct missing regions of human face images using contextual cues. Instead of AlexNet, which was used in the original paper, we will adopt a more modern architecture, such as ResNet, to enhance feature extraction and generation quality. We will use the VOC2012 dataset to generate random masks, ensuring diverse and challenging inpainting tasks. This project aims to explore the effectiveness of contextual feature learning for human face restoration while leveraging deep learning techniques for image reconstruction.

1 Problem Definition

The problem we aim to solve is reconstructing missing regions in images, specifically focusing on human faces. The ability to predict and restore missing visual information is crucial for applications in computer vision, medical imaging, and digital forensics. Our approach is based on the context encoder framework, which utilizes an encoder-decoder architecture to generate plausible image reconstructions.

- Motivation: Image inpainting has significant real-world applications, including photo restoration, video frame interpolation, and medical image recovery.
- Key Components: We will train a deep learning model to predict and restore missing portions of face images by learning contextual features from surrounding pixels. Instead of using AlexNet, we will using more modern model like ResNet for better feature extraction and reconstruction quality.
- Understanding the Problem: The challenge lies in generating semantically meaningful content while maintaining coherence with the given context. To improve model robustness, we will use random masks derived from the VOC2012 dataset. Our implementation will aim to improve upon the baseline results of the original paper by leveraging a more powerful architecture and optimizing the loss function to balance realism and accuracy in image inpainting.

2 Tentative Method

Our approach follows a deep learning-based image inpainting framework inspired by Context Encoders: Feature Learning by Inpainting [1] , with key modifications to enhance performance. The method consists of the following components:

- Dataset: We will use a dataset consisting of human face images. To create diverse occlusions for training, we will apply random masks generated from the VOC2012 dataset [2].
- Model Architecture: Instead of AlexNet, we will use modern architecture like ResNet [3] as our encoder network. The encoder will extract high-level semantic features from the

unmasked regions of the image. The decoder will then generate the missing region based on the learned context.

- Training Strategy: Our training objective will include a combination of reconstruction loss and adversarial loss to produce realistic and coherent inpainting results.
- Evaluation: The model will be evaluated based on qualitative results (visual inspection) and quantitative metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM).

Reason for Choosing This Method:

- Improved Feature Representation: ResNet is known for its strong feature extraction capabilities compared to AlexNet, allowing for better semantic understanding and more realistic inpainting results.
- Robust Masking Strategy: By using VOC2012 segmentation masks, we ensure diverse occlusions that improve the generalizability of our model beyond fixed central square masks.
- Effective Loss Function: Combining L2 reconstruction loss with adversarial loss enables the model to balance overall structure reconstruction with fine-grained details, preventing overly blurry inpainted regions.

Strength of the Chosen Method:

- Modern and Efficient Architecture: ResNet allows deeper feature learning, leading to more realistic and contextually accurate reconstructions.
- Generalization Capability: Using VOC2012 masks ensures that the model learns to fill in arbitrary missing regions rather than just fixed areas, making it more adaptable to real-world applications.
- Combination of Losses for Better Performance: The joint use of reconstruction and adversarial losses ensures that the generated content is both visually convincing and structurally consistent with the surrounding image.

This methodology aims to improve upon existing inpainting techniques and provide a robust approach to restoring human face images with missing regions.

3 Experiments

- **Data:**

  As the task for which we are trying to solve is more specific than the initial research paper, we chose an intial dataset that contains around 7.2k images of faces [4]. This dataset is specifically made for ML training and contains a variety of different faces to work with. The data format of all images in the dataset are RGB JPEG images. We will change the format of these images to be pytorch tensors after the pre-processing steps, which the size of these pytorch tensors will be determined after seeing which size is best during pre-processing.

  This dataset, while being expansive and made for general ML models, isn't specifically made for CNN training and it isn't made for inpainting tasks either. As such, the images are not rectangular nor do they have artifacts taken out of the image, leading to them needing pre-processing on our end. We will need to use a basic face detection model to detect the faces in our dataset and crop around it. After this, we will use a combination of the masks of VOC2012 dataset for randomized shapes and normal square masks to set pixels 0, creating the dropout artifacts mentioned in the paper. If this dataset doesn't work under these pre-processing steps, we will try to find a new dataset that may not be as expansive, but has images better suited to the type of model we are trying to train.

- **Planned Experiments:**

  The planned experiments for this model are as followed:

  Experiment 1: Fill in the missing artifacts for a subset of the data, specifically not used during the training and validation processes, that has some missing artifacts the shape of the squares and others the shape of the deformed VOC2012 masks. This is to ensure that data similar to what we trained on can be tested and verified to fill correctly.

  Experiment 2: Fill in the missing artifacts for pictures of new faces that aren't part of the dataset (either taken by us or taken from a brand new dataset), with missing randomized artifacts taken out of these images. View and evaluate the results of said context encoding filling.

  Experiment 3: Compare the performance of our revised context encoder for face inpainting with the model created in [5], which is specifically created for resolving missing artifacts in faces using a deep generative model.

References

[1] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, & Alexei A. Efros. (2016). Context Encoders: Feature Learning by Inpainting.

[2] PASCAL VOC 2012 DATASET. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/gopalbhattrai/pascal-voc-2012-dataset

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. (2015). Deep Residual Learning for Image Recognition.

[4] Gupta, A. (2020). Human Faces. Kaggle.com. https://www.kaggle.com/datasets/ashwingupta3012/human-faces/data

[5] Yijun Li, Sifei Liu, Jimei Yang, & Ming-Hsuan Yang. (2017). Generative Face Completion.