Automatic Removal of Human Shapes on Pictures

Elise Fournier-Bidoz, Alexandre El Assad
Stanford University

Motivation

- The project aims at removing human shapes from landscape pictures. Today, this operation can be achieved through the commercial software Adobe® Photoshop® in about 20 min. This problem combines two interesting aspects which are object (in our case human) recognition and removal, as well as picture filling.

Removal of a human shape using Adobe® Photoshop®

- Our approach involves building a pipeline with two neural networks that will consecutively detect and remove human shape(s) from a picture, before filling in the removed pixels.

Related Work

- *Detectron* is a state-of-the art platform for object detection research developed by Facebook, built on Caffe2.

Detectron output

- *FCN-8* is a fully convolutional network for semantic segmentation. It adapts contemporary classification networks (AlexNet, VGG net and GoogLeNet) into fully convolutional networks and transfers their learned representations by fine-tuning to the segmentation task.
- *Semantic Image Inpainting with Deep Generative Models* reconstructs the missing content by conditioning on the available data, meaning that inference is independent of the missing content's structure. Existing methods which extract information from a single image generally don't perform as well due to their lack of high level context usage.

Datasets and Solving Approach

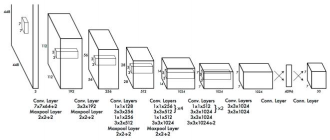
Datasets

We collected data for both neural networks using the same method:

- A web scrapping script was used to download Google images based on a query such as "landscapes with people".
- Data augmentation was performed on these images to increase the size of our dataset.

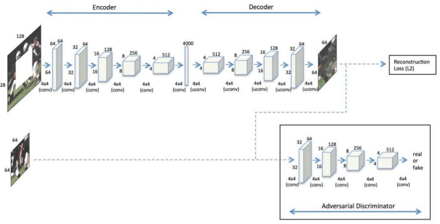
Combining two neural networks

- We use YOLO as our first neural network in order to detect humans on pictures. We modified YOLO such that in the output images :
  - only the class "person" is detected,
  - the pixels contained in the output bounding boxes are removed.



YOLO architecture

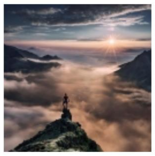
- To fill in the missing part of the picture, we use context encoders, a bootleneck convolutional neural network architecture, coupled to a GAN. In training, the bottleneck inputs the masked image and outputs the filled-in mask using L2 loss. The reconstructed image is then fed into the discriminator which compares it to the real image.



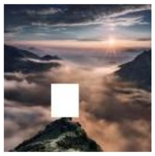
Context-encoder architecture

Results

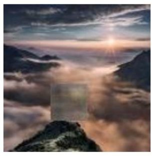
Human shape removal and replacement



Original image

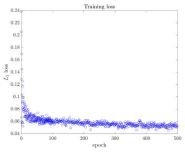


Masked image



Filled image

Filling neural network



Dataset	L_2 loss
Train	0.058
Dev	0.08
Test	0.15

- Good qualitative results on train and dev sets.
- Poor results on the test set ⇒ overfitting to the train and dev sets.

Conclusions and Future Work

- We have demonstrated the ability to automatically remove a human shape from a picture. However, better results can still be achieved.
- The test loss being considerably higher than the train and dev losses, we need to collect more training data to avoid overfitting.
- Performing the human detection and removal task with Detectron would reduce the size of removed areas from original images.
- Retraining the filling neural network on human or random shaped masks would then be necessary in order to obtain realistic image reconstructions.

References

[1] Pathak, Deepak et al. "Context Encoders: Feature Learning by Inpainting." CVPR 2016
[2]Redmon, Joseph and Farhadi, Ali. "YOLO9000: Better, Faster, Stronger" arXiv preprint arXiv:1612.08242, 2016
[3] Shelhamer, Evan et al. "Fully Convolutional Models for Semantic Segmentation." PAMI, arXiv:1605.06211, 2016 [4] Ulyanov, Dimitri et al. "Deep Image Prior", CoRR, 2017

CS 230: Deep Learning | Winter 2017

{efb, aelassad}@stanford.edu