An Analysis of Synthetic Dataset Creation Using Deep Generative Models

By William Lee, Zach Rothenberg, and Jimmy Shah

Motivation

The meteoric rise of deep learning to the mainstream has spelled the start of an age of data

We propose the use of modern deep generative models for the purpose of generating new synthetic datasets from existing datasets

Central Question

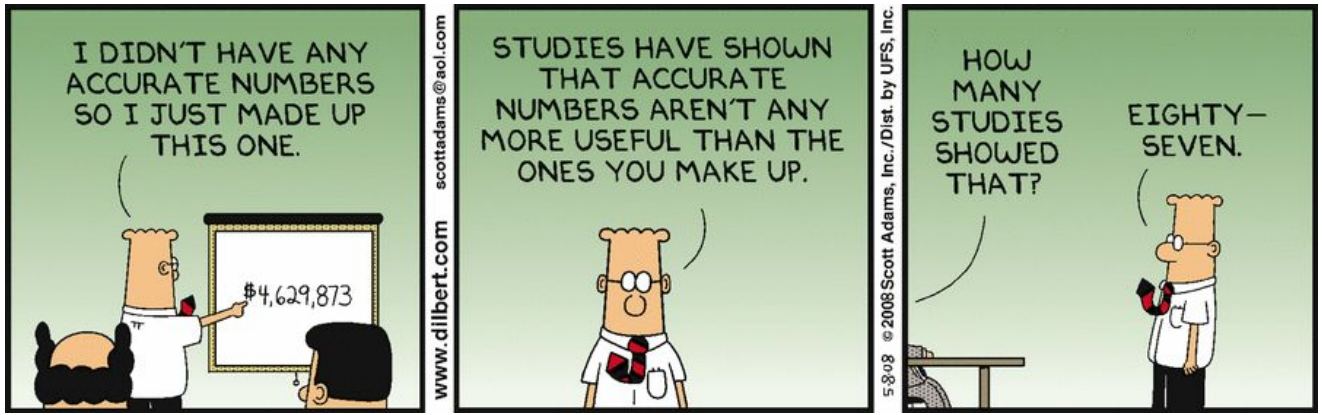
Can synthetic data augment or replace real datasets?

(in the image domain).

Applications in other domains

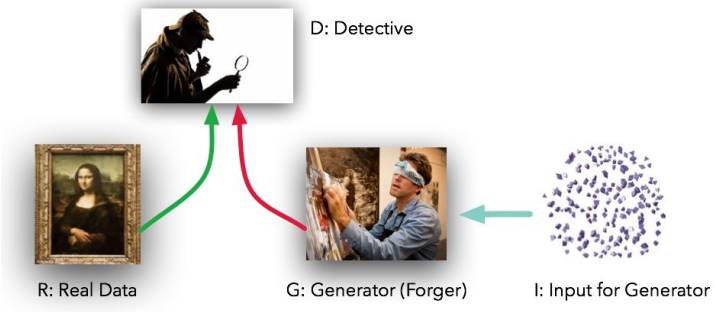
Good Synthetic Data

- What makes “good” synthetic data “good”?
  - "obviously synthetic"
    - different from the original data.
  - "offer strong privacy guarantees"



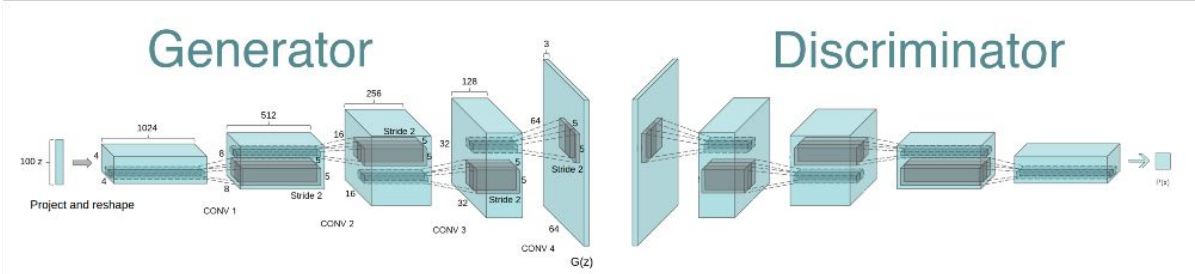
Generative Adversarial Networks (GANs)

- Based on a two player zero-sum game
  - Two Neural Networks compete against one another
    - Generator
    - Discriminator
  - Just like real life forgery
- Both networks get better in tandem



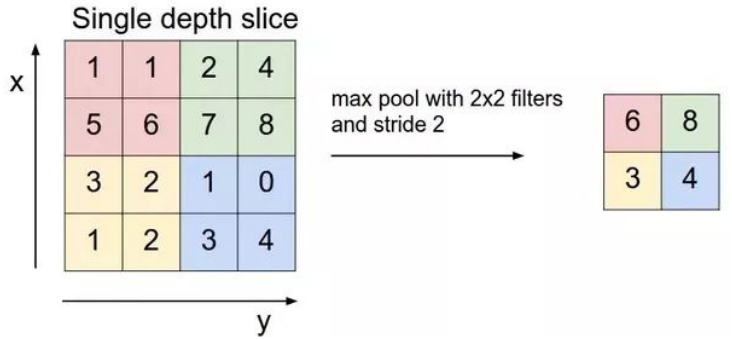
Deep Convolutional Generative Adversarial Networks (DCGANs)

- Use Convolutional Neural Nets (CNNs) as the two neural networks
  - Shared parameters across the network allow for performance on high dimensional feature vectors (Images)
- Attempt to solve GAN instability (i.e. mode collapse, G/D imbalance)
  - **High Stride convolutions**
  - No fully connected layers
  - **Batch Normalization**
  - Leaky ReLU activation



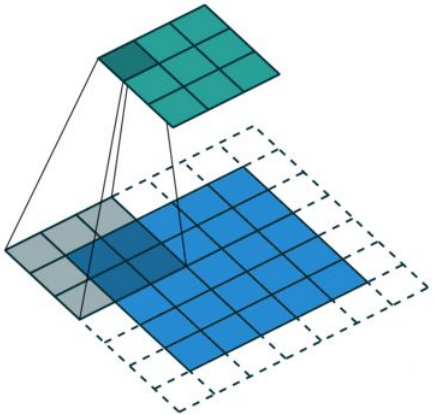
High Stride Convolution

Max Pooling / Average Pooling



High Stride Length

Allow networks to learn their own down/up sampling

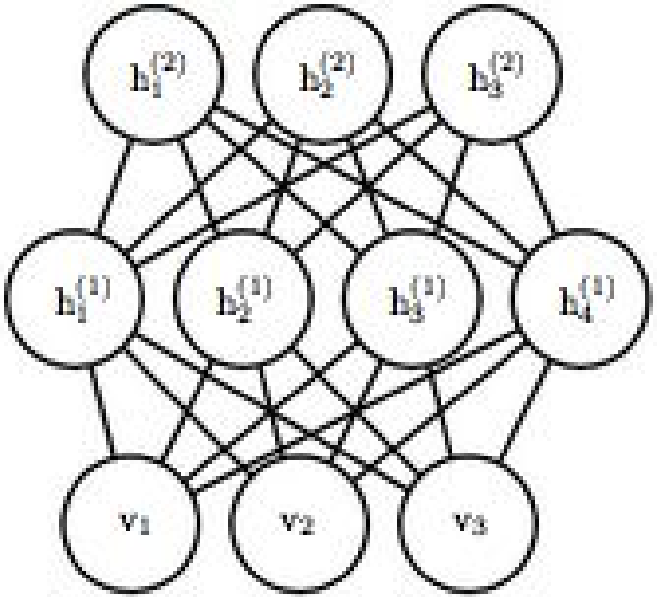


Batch Normalization

- Discriminator network expects array of images
  - Can discriminate based on entropy of Generation
  - Prevents Generator from collapsing on a single representation
- Normalize inputs after each activation
  - Helps prevent vanishing gradients
  - Faster Learning
  - More accuracy

Deep Boltzmann Machines

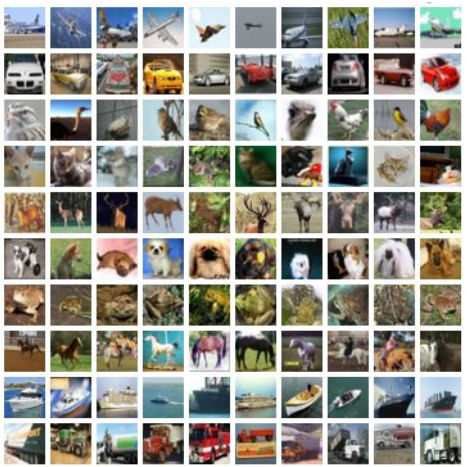
- Undirected Graphical Model
  - Markov Random Field
  - Deep architecture
- Gibbs sampling
  - Random initialization
  - Alternate between even and odd layers



Datasets

MNIST

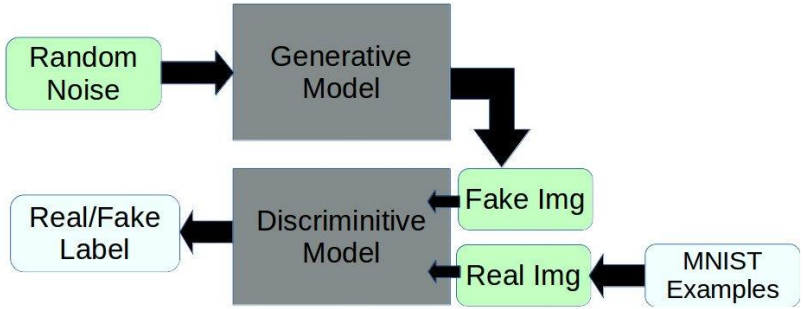
CIFAR-10



GAN Setup

GAN:

- Trained for 7 epochs on MNIST, with 64 image batches
- Fed generator Gaussian noise samples and labels set
- Produced 100 images for a given label, for 1000 total images



DBM Issues

DBM:

- Generated fuzzy images, as shown on CIFAR-10 sample
- DBM on MNIST resulted in better images than CIFAR-10, but 100 images limit made it tough as a dataset.
- Trouble with TensorFlow setup. Code was not modular.

Evaluation

- MNIST Dataset
- Two models:
  - CNN
    - Unstable model
    - Commonly used for modern image processing
  - SVM
    - Stable model
    - Less state of the art for images

Evaluation (Continued)

- **Experiment #1:** Models trained on 100 images (10 each class)
  - Test set of 10,000 real images
  - Performance when trained on 100 real images vs 100 generated
  - Does the Synthetic dataset lead to a model generalized on the real dataset?
- **Experiment #2:** Mixed dataset
  - Starting seed of 10 real images (1 each class)
  - Progressively add 10 images to the training set
  - Plot training curves
  - Does the Synthetic data add the same progressive benefit that the real data adds?

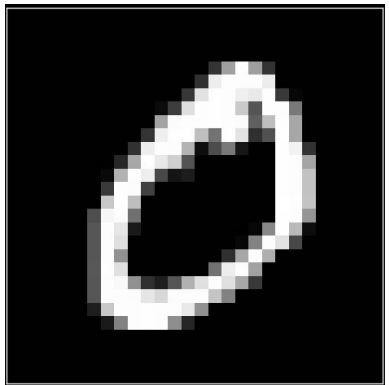
Results! (Experiment #1)

- CNN
  - Synthetic data is just as good as real data
- SVM
  - Synthetic data is better than real data?
  - The Synthetic data is “fuzzier”
  - “Fuzzy” data prevents overfitting

Table 1. Real vs Generated MNIST Performance

MODEL	REAL	GENERATED
SVM	0.653 ± 0.048	0.742 ± 0.021
CNN	0.610 ± 0.048	0.604 ± 0.038

Real



Synthetic

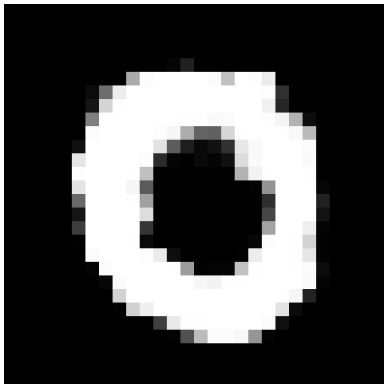


Table 2. Real vs Generated MNIST Performance (1000 training examples)

MODEL	REAL	GENERATED
SVM	0.903 ± 0.008	0.899

Once the model is given adequate examples to not overfit, performance equalizes

CNN Performance on Real Data (Experiment #2)

- As we increased the sample of real images, CNN performance increased.
- Peak accuracy of 64.77%

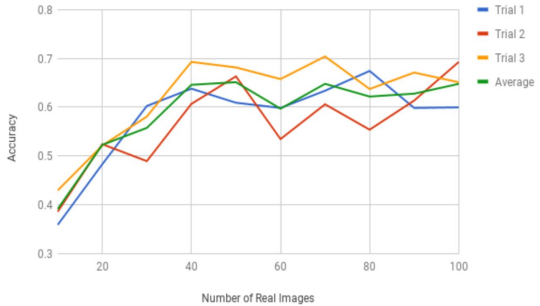


Figure 4. CNN Performance on Varied Sample of Real Images

CNN Performance on Generated Data

- With Real images fixed to 10, CNN performance on generated also increased.
- Peak accuracy of 61.52%

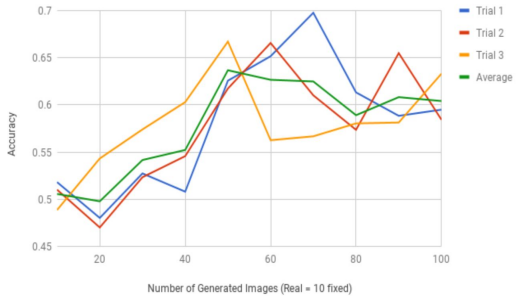


Figure 5. CNN Performance on Varied Generated Data

Comparing Learning Curves

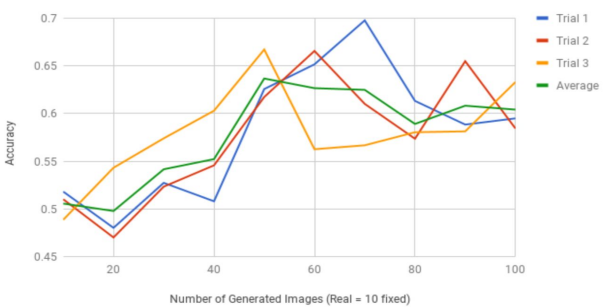


Figure 5. CNN Performance on Varied Generated Data

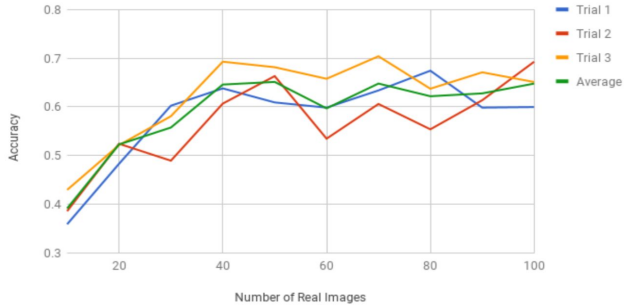


Figure 4. CNN Performance on Varied Sample of Real Images

More Trials

Over 10 trials, real and generated data performed similarly (n=100 for both).

This suggests that our synthetic data worked well?

Howe et al. suggested that similar correlations among variables should exist for both datasets, and we find this to be true

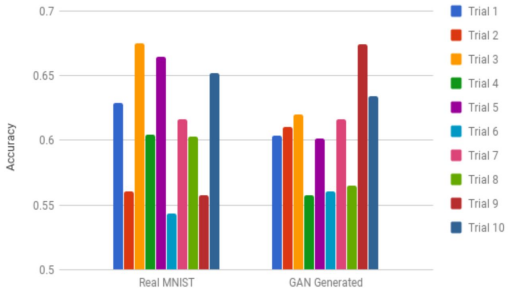


Figure 6. Performance Comparison over 10 Trials

Conclusions

- Creation of Synthetic Datasets in Privacy sensitive fields
  - Medicine
  - Childcare
  - Education
  - Personal consumer records
    - Energy
- Allows for academic cooperation
  - ImageNet

Future Work

- Generalizing on more complex datasets
- TFGAN
- DBM

Google Research Blog

The latest news from Research at Google

TFGAN: A Lightweight Library for Generative Adversarial Networks

Tuesday, December 12, 2017

Posted by Joel Shor, Senior Software Engineer, Machine Perception