

# RNN\_Captioning

December 8, 2021

```
[ ]: # This mounts your Google Drive to the Colab VM.  
from google.colab import drive  
drive.mount('/content/drive')  
  
# TODO: Enter the foldername in your Drive where you have saved the unzipped  
# assignment folder, e.g. 'cs231n/assignments/assignment1/'  
FOLDERNAME = 'CS682/assignment3'  
assert FOLDERNAME is not None, "(!) Enter the foldername."  
  
# Now that we've mounted your Drive, this ensures that  
# the Python interpreter of the Colab VM can load  
# python files from within it.  
import sys  
sys.path.append('/content/drive/My Drive/{}'.format(FOLDERNAME))  
  
# This downloads the CIFAR-10 dataset to your Drive  
# if it doesn't already exist.  
%cd /content/drive/My\ Drive/$FOLDERNAME/cs682/datasets/  
!bash get_datasets.sh  
%cd /content/drive/My\ Drive/$FOLDERNAME  
  
[ ]: !pip install scipy==1.1.0
```

## 1 Image Captioning with RNNs

In this exercise you will implement a vanilla recurrent neural networks and use them it to train a model that can generate novel captions for images.

```
[ ]: # As usual, a bit of setup  
import time, os, json  
import numpy as np  
import matplotlib.pyplot as plt  
  
from cs682.gradient_check import eval_numerical_gradient, □  
    →eval_numerical_gradient_array  
from cs682.rnn_layers import *  
from cs682.captioning_solver import CaptioningSolver
```

```

from cs682.classifiers.rnn import CaptioningRNN
from cs682.coco_utils import load_coco_data, sample_coco_minibatch, □
    →decode_captions
from cs682.image_utils import image_from_url

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/
→autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2

def rel_error(x, y):
    """ returns relative error """
    return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))

```

## 1.1 Install h5py

The COCO dataset we will be using is stored in HDF5 format. To load HDF5 files, we will need to install the h5py Python package. Check if h5py is already installed:

[4]: `import h5py`

If the modual is not found, you will need to install it now. From the command line, run: pip install h5py If you receive a permissions error, you may need to run the command as root: sudo pip install h5py

You can also run commands directly from the Jupyter notebook by prefixing the command with the "!" character:

[5]: `!pip install h5py`

## 2 Microsoft COCO

For this exercise we will use the 2014 release of the [Microsoft COCO dataset](#) which has become the standard testbed for image captioning. The dataset consists of 80,000 training images and 40,000 validation images, each annotated with 5 captions written by workers on Amazon Mechanical Turk.

You should have already downloaded the data by changing to the `cs682/datasets` directory and running the script `get_assignment3_data.sh`. If you haven't yet done so, run that script now. Warning: the COCO data download is ~1GB.

We have preprocessed the data and extracted features for you already. For all images we have extracted features from the fc7 layer of the VGG-16 network pretrained on ImageNet; these features are stored in the files `train2014_vgg16_fc7.h5` and `val2014_vgg16_fc7.h5` respectively. To cut down on processing time and memory requirements, we have reduced

the dimensionality of the features from 4096 to 512; these features can be found in the files `train2014_vgg16_fc7_pca.h5` and `val2014_vgg16_fc7_pca.h5`.

The raw images take up a lot of space (nearly 20GB) so we have not included them in the download. However all images are taken from Flickr, and URLs of the training and validation images are stored in the files `train2014_urls.txt` and `val2014_urls.txt` respectively. This allows you to download images on the fly for visualization. Since images are downloaded on-the-fly, **you must be connected to the internet to view images.**

Dealing with strings is inefficient, so we will work with an encoded version of the captions. Each word is assigned an integer ID, allowing us to represent a caption by a sequence of integers. The mapping between integer IDs and words is in the file `coco2014_vocab.json`, and you can use the function `decode_captions` from the file `cs682/coco_utils.py` to convert numpy arrays of integer IDs back into strings.

There are a couple special tokens that we add to the vocabulary. We prepend a special `<START>` token and append an `<END>` token to the beginning and end of each caption respectively. Rare words are replaced with a special `<UNK>` token (for "unknown"). In addition, since we want to train with minibatches containing captions of different lengths, we pad short captions with a special `<NULL>` token after the `<END>` token and don't compute loss or gradient for `<NULL>` tokens. Since they are a bit of a pain, we have taken care of all implementation details around special tokens for you.

You can load all of the MS-COCO data (captions, features, URLs, and vocabulary) using the `load_coco_data` function from the file `cs682/coco_utils.py`. Run the following cell to do so:

```
[ ]: %ls
[ ]: %cd cs682/datasets
!sh get_coco_captioning.sh
!sh get_squeezeNet_tf.sh
!sh get_imagenet_val.sh
%cd ..
%cd ..

[83]: # Load COCO data from disk; this returns a dictionary
# We'll work with dimensionality-reduced features for this notebook, but feel
# free to experiment with the original features by changing the flag below.
data = load_coco_data(pca_features=True)

# Print out all the keys and values from the data dictionary
for k, v in data.items():
    if type(v) == np.ndarray:
        print(k, type(v), v.shape, v.dtype)
    else:
        print(k, type(v), len(v))

trainCaptions <class 'numpy.ndarray'> (400135, 17) int32
trainImageIdxs <class 'numpy.ndarray'> (400135,) int32
valCaptions <class 'numpy.ndarray'> (195954, 17) int32
valImageIdxs <class 'numpy.ndarray'> (195954,) int32
trainFeatures <class 'numpy.ndarray'> (82783, 512) float32
valFeatures <class 'numpy.ndarray'> (40504, 512) float32
```

```
idx_to_word <class 'list'> 1004
word_to_idx <class 'dict'> 1004
train_urls <class 'numpy.ndarray'> (82783,) <U63
val_urls <class 'numpy.ndarray'> (40504,) <U63
```

## 2.1 Look at the data

It is always a good idea to look at examples from the dataset before working with it.

You can use the `sample_coco_minibatch` function from the file `cs682/coco_utils.py` to sample minibatches of data from the data structure returned from `load_coco_data`. Run the following to sample a small minibatch of training data and show the images and their captions. Running it multiple times and looking at the results helps you to get a sense of the dataset.

Note that we decode the captions using the `decode_captions` function and that we download the images on-the-fly using their Flickr URL, so **you must be connected to the internet to view images**.

```
[84]: # Sample a minibatch and show the images and captions
batch_size = 3

captions, features, urls = sample_coco_minibatch(data, batch_size=batch_size)
for i, (caption, url) in enumerate(zip(captions, urls)):
    plt.imshow(image_from_url(url))
    plt.axis('off')
    caption_str = decode_captions(caption, data['idx_to_word'])
    plt.title(caption_str)
    plt.show()
```

<START> a large <UNK> vase filled with white and yellow flowers <END>



<START> the vase has a pretty blue flower in it <END>



<START> the pizza is cooked and ready to be eaten <END>



### 3 Recurrent Neural Networks

As discussed in lecture, we will use recurrent neural network (RNN) language models for image captioning. The file `cs682/rnn_layers.py` contains implementations of different layer types that are needed for recurrent neural networks, and the file `cs682/classifiers/rnn.py` uses these layers to implement an image captioning model.

We will first implement different types of RNN layers in `cs682/rnn_layers.py`.

### 4 Vanilla RNN: step forward

Open the file `cs682/rnn_layers.py`. This file implements the forward and backward passes for different types of layers that are commonly used in recurrent neural networks.

First implement the function `rnn_step_forward` which implements the forward pass for a single timestep of a vanilla recurrent neural network. After doing so run the following to check your implementation. You should see errors on the order of e-8 or less.

[85]: `N, D, H = 3, 10, 4`

```
x = np.linspace(-0.4, 0.7, num=N*D).reshape(N, D)
prev_h = np.linspace(-0.2, 0.5, num=N*H).reshape(N, H)
Wx = np.linspace(-0.1, 0.9, num=D*H).reshape(D, H)
```

```

Wh = np.linspace(-0.3, 0.7, num=H*H).reshape(H, H)
b = np.linspace(-0.2, 0.4, num=H)

next_h, _ = rnn_step_forward(x, prev_h, Wx, Wh, b)
expected_next_h = np.asarray([
    [-0.58172089, -0.50182032, -0.41232771, -0.31410098],
    [ 0.66854692,  0.79562378,  0.87755553,  0.92795967],
    [ 0.97934501,  0.99144213,  0.99646691,  0.99854353]]))

print('next_h error: ', rel_error(expected_next_h, next_h))

```

next\_h error: 6.292421426471037e-09

## 5 Vanilla RNN: step backward

In the file cs682/rnn\_layers.py implement the rnn\_step\_backward function. After doing so run the following to numerically gradient check your implementation. You should see errors on the order of e-8 or less.

```

[86]: from cs682.rnn_layers import rnn_step_forward, rnn_step_backward
np.random.seed(231)
N, D, H = 4, 5, 6
x = np.random.randn(N, D)
h = np.random.randn(N, H)
Wx = np.random.randn(D, H)
Wh = np.random.randn(H, H)
b = np.random.randn(H)

out, cache = rnn_step_forward(x, h, Wx, Wh, b)

dnext_h = np.random.randn(*out.shape)

fx = lambda x: rnn_step_forward(x, h, Wx, Wh, b)[0]
fh = lambda prev_h: rnn_step_forward(x, h, Wx, Wh, b)[0]
fWx = lambda Wx: rnn_step_forward(x, h, Wx, Wh, b)[0]
fWh = lambda Wh: rnn_step_forward(x, h, Wx, Wh, b)[0]
fb = lambda b: rnn_step_forward(x, h, Wx, Wh, b)[0]

dx_num = eval_numerical_gradient_array(fx, x, dnext_h)
dprev_h_num = eval_numerical_gradient_array(fh, h, dnext_h)
dWx_num = eval_numerical_gradient_array(fWx, Wx, dnext_h)
dWh_num = eval_numerical_gradient_array(fWh, Wh, dnext_h)
db_num = eval_numerical_gradient_array(fb, b, dnext_h)

dx, dprev_h, dWx, dWh, db = rnn_step_backward(dnext_h, cache)

print('dx error: ', rel_error(dx_num, dx))

```

```

print('dprev_h error: ', rel_error(dprev_h_num, dprev_h))
print('dWx error: ', rel_error(dWx_num, dWx))
print('dWh error: ', rel_error(dWh_num, dWh))
print('db error: ', rel_error(db_num, db))

```

```

dx error: 4.0192769090159184e-10
dprev_h error: 2.5632975303201374e-10
dWx error: 8.820222259148609e-10
dWh error: 4.703287554560559e-10
db error: 7.30162216654e-11

```

## 6 Vanilla RNN: forward

Now that you have implemented the forward and backward passes for a single timestep of a vanilla RNN, you will combine these pieces to implement a RNN that processes an entire sequence of data.

In the file `cs682/rnn_layers.py`, implement the function `rnn_forward`. This should be implemented using the `rnn_step_forward` function that you defined above. After doing so run the following to check your implementation. You should see errors on the order of e-7 or less.

[87]: N, T, D, H = 2, 3, 4, 5

```

x = np.linspace(-0.1, 0.3, num=N*T*D).reshape(N, T, D)
h0 = np.linspace(-0.3, 0.1, num=N*H).reshape(N, H)
Wx = np.linspace(-0.2, 0.4, num=D*H).reshape(D, H)
Wh = np.linspace(-0.4, 0.1, num=H*H).reshape(H, H)
b = np.linspace(-0.7, 0.1, num=H)

h, _ = rnn_forward(x, h0, Wx, Wh, b)
expected_h = np.asarray([
    [
        [-0.42070749, -0.27279261, -0.11074945,  0.05740409,  0.22236251],
        [-0.39525808, -0.22554661, -0.0409454,   0.14649412,  0.32397316],
        [-0.42305111, -0.24223728, -0.04287027,  0.15997045,  0.35014525],
    ],
    [
        [-0.55857474, -0.39065825, -0.19198182,  0.02378408,  0.23735671],
        [-0.27150199, -0.07088804,  0.13562939,  0.33099728,  0.50158768],
        [-0.51014825, -0.30524429, -0.06755202,  0.17806392,  0.40333043]
    ]
])
print('h error: ', rel_error(expected_h, h))

```

```

h error: 7.728466158305164e-08

```

## 7 Vanilla RNN: backward

In the file `cs682/rnn_layers.py`, implement the backward pass for a vanilla RNN in the function `rnn_backward`. This should run back-propagation over the entire sequence, making calls to the

rnn\_step\_backward function that you defined earlier. You should see errors on the order of e-6 or less.

[88]: np.random.seed(231)

```
N, D, T, H = 2, 3, 10, 5

x = np.random.randn(N, T, D)
h0 = np.random.randn(N, H)
Wx = np.random.randn(D, H)
Wh = np.random.randn(H, H)
b = np.random.randn(H)

out, cache = rnn_forward(x, h0, Wx, Wh, b)

dout = np.random.randn(*out.shape)

dx, dh0, dWx, dWh, db = rnn_backward(dout, cache)

fx = lambda x: rnn_forward(x, h0, Wx, Wh, b)[0]
fh0 = lambda h0: rnn_forward(x, h0, Wx, Wh, b)[0]
fWx = lambda Wx: rnn_forward(x, h0, Wx, Wh, b)[0]
fWh = lambda Wh: rnn_forward(x, h0, Wx, Wh, b)[0]
fb = lambda b: rnn_forward(x, h0, Wx, Wh, b)[0]

dx_num = eval_numerical_gradient_array(fx, x, dout)
dh0_num = eval_numerical_gradient_array(fh0, h0, dout)
dWx_num = eval_numerical_gradient_array(fWx, Wx, dout)
dWh_num = eval_numerical_gradient_array(fWh, Wh, dout)
db_num = eval_numerical_gradient_array(fb, b, dout)

print('dx error: ', rel_error(dx_num, dx))
print('dh0 error: ', rel_error(dh0_num, dh0))
print('dWx error: ', rel_error(dWx_num, dWx))
print('dWh error: ', rel_error(dWh_num, dWh))
print('db error: ', rel_error(db_num, db))
```

```
dx error:  1.5382468491701097e-09
dh0 error:  3.3839681556240896e-09
dWx error:  7.150535245339328e-09
dWh error:  1.297338408201546e-07
db error:  1.4889022954777414e-10
```

## 8 Word embedding: forward

In deep learning systems, we commonly represent words using vectors. Each word of the vocabulary will be associated with a vector, and these vectors will be learned jointly with the rest of the system.

In the file `cs682/rnn_layers.py`, implement the function `word_embedding_forward` to convert words (represented by integers) into vectors. Run the following to check your implementation. You should see an error on the order of e-8 or less.

[89]: N, T, V, D = 2, 4, 5, 3

```
x = np.asarray([[0, 3, 1, 2], [2, 1, 0, 3]])
W = np.linspace(0, 1, num=V*D).reshape(V, D)

out, _ = word_embedding_forward(x, W)
expected_out = np.asarray([
    [[0., 0.07142857, 0.14285714],
     [0.64285714, 0.71428571, 0.78571429],
     [0.21428571, 0.28571429, 0.35714286],
     [0.42857143, 0.5, 0.57142857]],
    [[0.42857143, 0.5, 0.57142857],
     [0.21428571, 0.28571429, 0.35714286],
     [0., 0.07142857, 0.14285714],
     [0.64285714, 0.71428571, 0.78571429]]])

print('out error: ', rel_error(expected_out, out))
```

out error: 1.000000094736443e-08

## 9 Word embedding: backward

Implement the backward pass for the word embedding function in the function `word_embedding_backward`. After doing so run the following to numerically gradient check your implementation. You should see an error on the order of e-11 or less.

[90]: np.random.seed(231)

```
N, T, V, D = 50, 3, 5, 6
x = np.random.randint(V, size=(N, T))
W = np.random.randn(V, D)

out, cache = word_embedding_forward(x, W)
dout = np.random.randn(*out.shape)
dW = word_embedding_backward(dout, cache)

f = lambda W: word_embedding_forward(x, W)[0]
dW_num = eval_numerical_gradient_array(f, W, dout)

print('dW error: ', rel_error(dW, dW_num))
```

dW error: 3.2774595693100364e-12

## 10 Temporal Affine layer

At every timestep we use an affine function to transform the RNN hidden vector at that timestep into scores for each word in the vocabulary. Because this is very similar to the affine layer that you implemented in assignment 2, we have provided this function for you in the `temporal_affine_forward` and `temporal_affine_backward` functions in the file `cs682/rnn_layers.py`. Run the following to perform numeric gradient checking on the implementation. You should see errors on the order of e-9 or less.

```
[91]: np.random.seed(231)

# Gradient check for temporal affine layer
N, T, D, M = 2, 3, 4, 5
x = np.random.randn(N, T, D)
w = np.random.randn(D, M)
b = np.random.randn(M)

out, cache = temporal_affine_forward(x, w, b)

dout = np.random.randn(*out.shape)

fx = lambda x: temporal_affine_forward(x, w, b)[0]
fw = lambda w: temporal_affine_forward(x, w, b)[0]
fb = lambda b: temporal_affine_forward(x, w, b)[0]

dx_num = eval_numerical_gradient_array(fx, x, dout)
dw_num = eval_numerical_gradient_array(fw, w, dout)
db_num = eval_numerical_gradient_array(fb, b, dout)

dx, dw, db = temporal_affine_backward(dout, cache)

print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))
```

```
dx error:  2.9215945034030545e-10
dw error:  1.5772088618663602e-10
db error:  3.252200556967514e-11
```

## 11 Temporal Softmax loss

In an RNN language model, at every timestep we produce a score for each word in the vocabulary. We know the ground-truth word at each timestep, so we use a softmax loss function to compute loss and gradient at each timestep. We sum the losses over time and average them over the minibatch.

However there is one wrinkle: since we operate over minibatches and different captions may have different lengths, we append <NULL> tokens to the end of each caption so they all have the

same length. We don't want these <NULL> tokens to count toward the loss or gradient, so in addition to scores and ground-truth labels our loss function also accepts a `mask` array that tells it which elements of the scores count towards the loss.

Since this is very similar to the softmax loss function you implemented in assignment 1, we have implemented this loss function for you; look at the `temporal_softmax_loss` function in the file `cs682/rnn_layers.py`.

Run the following cell to sanity check the loss and perform numeric gradient checking on the function. You should see an error for `dx` on the order of e-7 or less.

```
[92]: # Sanity check for temporal softmax loss
from cs682.rnn_layers import temporal_softmax_loss

N, T, V = 100, 1, 10

def check_loss(N, T, V, p):
    x = 0.001 * np.random.randn(N, T, V)
    y = np.random.randint(V, size=(N, T))
    mask = np.random.rand(N, T) <= p
    print(temporal_softmax_loss(x, y, mask)[0])

check_loss(100, 1, 10, 1.0)    # Should be about 2.3
check_loss(100, 10, 10, 1.0)   # Should be about 23
check_loss(5000, 10, 10, 0.1)  # Should be about 2.3

# Gradient check for temporal softmax loss
N, T, V = 7, 8, 9

x = np.random.randn(N, T, V)
y = np.random.randint(V, size=(N, T))
mask = (np.random.rand(N, T) > 0.5)

loss, dx = temporal_softmax_loss(x, y, mask, verbose=False)

dx_num = eval_numerical_gradient(lambda x: temporal_softmax_loss(x, y, mask)[0], x, verbose=False)

print('dx error: ', rel_error(dx, dx_num))
```

```
2.3027781774290146
23.025985953127226
2.2643611790293394
dx error:  2.583585303524283e-08
```

## 12 RNN for image captioning

Now that you have implemented the necessary layers, you can combine them to build an image captioning model. Open the file `cs682/classifiers/rnn.py` and look at the `CaptioningRNN` class.

Implement the forward and backward pass of the model in the `loss` function. For now you only need to implement the case where `cell_type='rnn'` for vanialla RNNs; you will implement the LSTM case later. After doing so, run the following to check your forward pass using a small test case; you should see error on the order of e-10 or less.

```
[93]: N, D, W, H = 10, 20, 30, 40
word_to_idx = {'<NULL>': 0, 'cat': 2, 'dog': 3}
V = len(word_to_idx)
T = 13

model = CaptioningRNN(word_to_idx,
                      input_dim=D,
                      wordvec_dim=W,
                      hidden_dim=H,
                      cell_type='rnn',
                      dtype=np.float64)

# Set all model parameters to fixed values
for k, v in model.params.items():
    model.params[k] = np.linspace(-1.4, 1.3, num=v.size).reshape(*v.shape)

features = np.linspace(-1.5, 0.3, num=(N * D)).reshape(N, D)
captions = (np.arange(N * T) % V).reshape(N, T)

loss, grads = model.loss(features, captions)
expected_loss = 9.83235591003

print('loss: ', loss)
print('expected loss: ', expected_loss)
print('difference: ', abs(loss - expected_loss))
```

```
loss: 9.832355910027387
expected loss: 9.83235591003
difference: 2.6130209107577684e-12
```

Run the following cell to perform numeric gradient checking on the `CaptioningRNN` class; you should see errors around the order of e-6 or less.

```
[94]: np.random.seed(231)

batch_size = 2
timesteps = 3
input_dim = 4
wordvec_dim = 5
hidden_dim = 6
word_to_idx = {'<NULL>': 0, 'cat': 2, 'dog': 3}
vocab_size = len(word_to_idx)

captions = np.random.randint(vocab_size, size=(batch_size, timesteps))
```

```

features = np.random.randn(batch_size, input_dim)

model = CaptioningRNN(word_to_idx,
    input_dim=input_dim,
    wordvec_dim=wordvec_dim,
    hidden_dim=hidden_dim,
    cell_type='rnn',
    dtype=np.float64,
)

loss, grads = model.loss(features, captions)

for param_name in sorted(grads):
    f = lambda _: model.loss(features, captions)[0]
    param_grad_num = eval_numerical_gradient(f, model.params[param_name], u
→verbose=False, h=1e-6)
    e = rel_error(param_grad_num, grads[param_name])
    print('%s relative error: %e' % (param_name, e))

```

```

W_embed relative error: 2.331070e-09
W_proj relative error: 1.112417e-08
W_vocab relative error: 4.274379e-09
Wh relative error: 5.858117e-09
Wx relative error: 1.590657e-06
b relative error: 9.727211e-10
b_proj relative error: 1.934807e-08
b_vocab relative error: 7.087097e-11

```

## 13 Overfit small data

Similar to the Solver class that we used to train image classification models on the previous assignment, on this assignment we use a CaptioningSolver class to train image captioning models. Open the file cs682/captioning\_solver.py and read through the CaptioningSolver class; it should look very familiar.

Once you have familiarized yourself with the API, run the following to make sure your model overfits a small sample of 100 training examples. You should see a final loss of less than 0.1.

[95]:

```

np.random.seed(231)

small_data = load_coco_data(max_train=50)

small_rnn_model = CaptioningRNN(
    cell_type='rnn',
    word_to_idx=data['word_to_idx'],
    input_dim=data['train_features'].shape[1],
    hidden_dim=512,
    wordvec_dim=256,
)

```

```

    )

small_rnn_solver = CaptioningSolver(small_rnn_model, small_data,
    update_rule='adam',
    num_epochs=50,
    batch_size=25,
    optim_config={
        'learning_rate': 5e-3,
    },
    lr_decay=0.95,
    verbose=True, print_every=10,
)
small_rnn_solver.train()

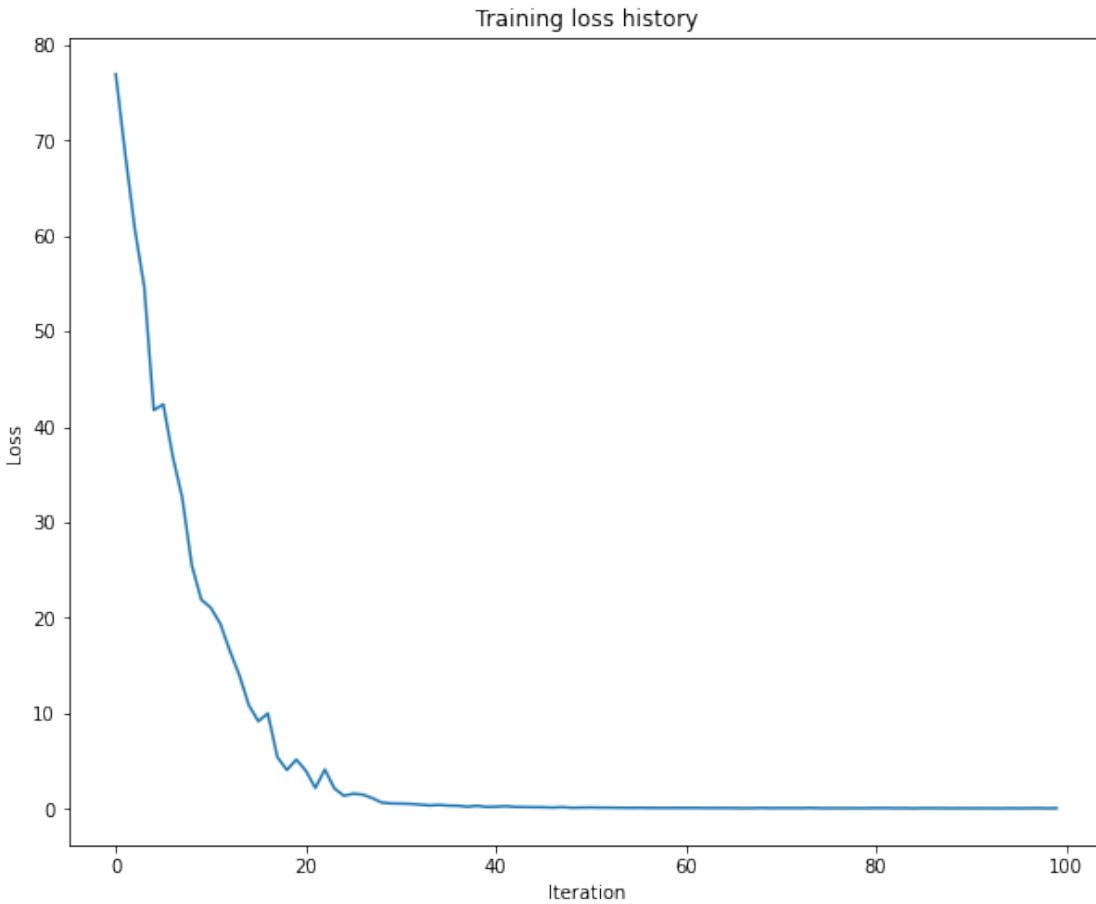
# Plot the training losses
plt.plot(small_rnn_solver.loss_history)
plt.xlabel('Iteration')
plt.ylabel('Loss')
plt.title('Training loss history')
plt.show()

```

```

(Iteration 1 / 100) loss: 76.913487
(Iteration 11 / 100) loss: 21.062729
(Iteration 21 / 100) loss: 4.016257
(Iteration 31 / 100) loss: 0.567182
(Iteration 41 / 100) loss: 0.239398
(Iteration 51 / 100) loss: 0.161974
(Iteration 61 / 100) loss: 0.111529
(Iteration 71 / 100) loss: 0.097580
(Iteration 81 / 100) loss: 0.099065
(Iteration 91 / 100) loss: 0.073967

```



## 14 Test-time sampling

Unlike classification models, image captioning models behave very differently at training time and at test time. At training time, we have access to the ground-truth caption, so we feed ground-truth words as input to the RNN at each timestep. At test time, we sample from the distribution over the vocabulary at each timestep, and feed the sample as input to the RNN at the next timestep.

In the file `cs682/classifiers/rnn.py`, implement the `sample` method for test-time sampling. After doing so, run the following to sample from your overfitted model on both training and validation data. The samples on training data should be very good; the samples on validation data probably won't make sense.

```
[96]: for split in ['train', 'val']:
    minibatch = sample_coco_minibatch(small_data, split=split, batch_size=2)
    gt_captions, features, urls = minibatch
    gt_captions = decode_captions(gt_captions, data['idx_to_word'])

    sample_captions = small_rnn_model.sample(features)
    sample_captions = decode_captions(sample_captions, data['idx_to_word'])
```

```
for gt_caption, sample_caption, url in zip(gtCaptions, sampleCaptions, urls):
    plt.imshow(image_from_url(url))
    plt.title('%s\n%s\nGT:%s' % (split, sampleCaption, gtCaption))
    plt.axis('off')
    plt.show()
```

train

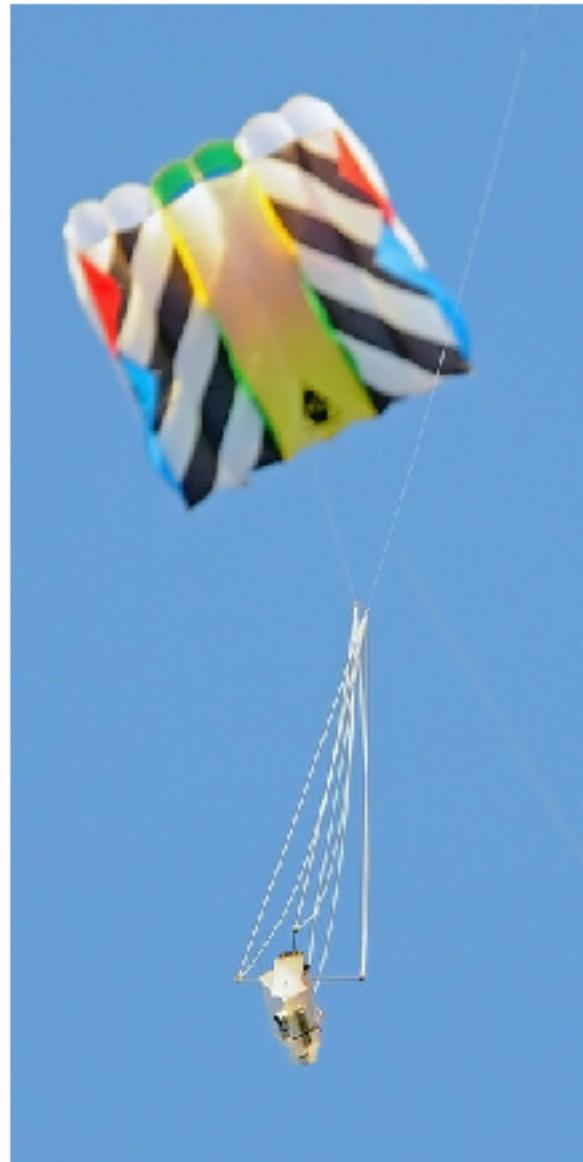
a boy sitting with <UNK> on with a donut in his hand <END>  
GT:<START> a boy sitting with <UNK> on with a donut in his hand <END>



train

a man <UNK> with a bright colorful kite <END>

GT:<START> a man <UNK> with a bright colorful kite <END>



val

two of <UNK> woman of a while in sun <UNK> <END>  
GT:<START> a red and white light tower on a hill near the ocean <END>



val  
 to tracks with out of a <END>  
 GT:<START> a table filled with many assorted food items <END>



## 15 INLINE QUESTION 1

In our current image captioning setup, our RNN language model produces a word at every timestep as its output. However, an alternate way to pose the problem is to train the network to operate over *characters* (e.g. 'a', 'b', etc.) as opposed to words, so that at every timestep, it receives the previous character as input and tries to predict the next character in the sequence. For example, the network might generate a caption like

'A', ' ', 'c', 'a', 't', ' ', 'o', ' ', 'n', ' ', 'a', ' ', 'b', ' ', 'e', ' ', 'd'

Can you describe one advantage of an image-captioning model that uses a character-level RNN? Can you also describe one disadvantage? HINT: there are several valid answers, but it might be useful to compare the parameter space of word-level and character-level models.

**Advantage:** The number of objects in the vocabulary may be lower. When using word-level RNN the number of words in our vocabulary may be over a million. At the same time the number of characters when using character-level RNN will be way lower most of the times. Therefore, a character-level RNN may use less memory than a word-level RNN for the vocabulary.

**Disadvantage:** The character-level RNN may produce words that don't make sense. Meaning it may group characters together to form a word that doesn't make sense rationally.

# Network Visualization-PyTorch

December 8, 2021

```
[ ]: # This mounts your Google Drive to the Colab VM.
from google.colab import drive
drive.mount('/content/drive')

# TODO: Enter the foldername in your Drive where you have saved the unzipped
# assignment folder, e.g. 'cs231n/assignments/assignment1/'
FOLDERNAME = 'CS682/assignment3'
assert FOLDERNAME is not None, "[!] Enter the foldername."

# Now that we've mounted your Drive, this ensures that
# the Python interpreter of the Colab VM can load
# python files from within it.
import sys
sys.path.append('/content/drive/My\ Drive/{}'.format(FOLDERNAME))

# This downloads the CIFAR-10 dataset to your Drive
# if it doesn't already exist.
%cd /content/drive/My\ Drive/$FOLDERNAME/cs682/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME
```

```
[ ]: pip install scipy==1.1.0
```

## 1 Network Visualization (PyTorch)

In this notebook we will explore the use of *image gradients* for generating new images.

When training a model, we define a loss function which measures our current unhappiness with the model's performance; we then use backpropagation to compute the gradient of the loss with respect to the model parameters, and perform gradient descent on the model parameters to minimize the loss.

Here we will do something slightly different. We will start from a convolutional neural network model which has been pretrained to perform image classification on the ImageNet dataset. We will use this model to define a loss function which quantifies our current unhappiness with our image, then use backpropagation to compute the gradient of this loss with respect to the pixels of the image. We will then keep the model fixed, and perform gradient descent *on the image* to synthesize a new image which minimizes the loss.

In this notebook we will explore three techniques for image generation:

1. **Saliency Maps**: Saliency maps are a quick way to tell which part of the image influenced the classification decision made by the network.
2. **Fooling Images**: We can perturb an input image so that it appears the same to humans, but will be misclassified by the pretrained network.
3. **Class Visualization**: We can synthesize an image to maximize the classification score of a particular class; this can give us some sense of what the network is looking for when it classifies images of that class.

This notebook uses **PyTorch**; we have provided another notebook which explores the same concepts in TensorFlow. You only need to complete one of these two notebooks.

```
[2]: import torch
import torchvision
import torchvision.transforms as T
import random
import numpy as np
from scipy.ndimage.filters import gaussian_filter1d
import matplotlib.pyplot as plt
from cs682.image_utils import SQUEEZENET_MEAN, SQUEEZENET_STD
from PIL import Image

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'
```

### 1.0.1 Helper Functions

Our pretrained model was trained on images that had been preprocessed by subtracting the per-color mean and dividing by the per-color standard deviation. We define a few helper functions for performing and undoing this preprocessing. You don't need to do anything in this cell.

```
[3]: def preprocess(img, size=224):
    transform = T.Compose([
        T.Resize(size),
        T.ToTensor(),
        T.Normalize(mean=SQUEEZENET_MEAN.tolist(),
                   std=SQUEEZENET_STD.tolist()),
        T.Lambda(lambda x: x[None]),
    ])
    return transform(img)

def deprocess(img, should_rescale=True):
    transform = T.Compose([
        T.Lambda(lambda x: x[0]),
        T.Normalize(mean=[0, 0, 0], std=(1.0 / SQUEEZENET_STD).tolist()),
        T.Normalize(mean=(-SQUEEZENET_MEAN).tolist(), std=[1, 1, 1]),
```

```

        T.Lambda(rescale) if should_rescale else T.Lambda(lambda x: x),
        T.ToPILImage(),
    ])
return transform(img)

def rescale(x):
    low, high = x.min(), x.max()
    x_rescaled = (x - low) / (high - low)
    return x_rescaled

def blur_image(X, sigma=1):
    X_np = X.cpu().clone().numpy()
    X_np = gaussian_filter1d(X_np, sigma, axis=2)
    X_np = gaussian_filter1d(X_np, sigma, axis=3)
    X.copy_(torch.Tensor(X_np).type_as(X))
    return X

```

## 2 Pretrained Model

For all of our image generation experiments, we will start with a convolutional neural network which was pretrained to perform image classification on ImageNet. We can use any model here, but for the purposes of this assignment we will use SqueezeNet [1], which achieves accuracies comparable to AlexNet but with a significantly reduced parameter count and computational complexity.

Using SqueezeNet rather than AlexNet or VGG or ResNet means that we can easily perform all image generation experiments on CPU.

[1] Iandola et al, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5MB model size", arXiv 2016

```

[ ]: # Download and load the pretrained SqueezeNet model.
model = torchvision.models.squeezenet1_1(pretrained=True)

# We don't want to train the model, so tell PyTorch not to compute gradients
# with respect to model parameters.
for param in model.parameters():
    param.requires_grad = False

# you may see warning regarding initialization deprecated, that's fine, please ↵
# continue to next steps

```

### 2.1 Load some ImageNet images

We have provided a few example images from the validation set of the ImageNet ILSVRC 2012 Classification dataset. To download these images, descend into cs682/datasets/ and run `get_imagenet_val.sh`.

Since they come from the validation set, our pretrained model did not see these images during training.

Run the following cell to visualize some of these images, along with their ground-truth labels.

```
[5]: from cs682.data_utils import load_imagenet_val
X, y, class_names = load_imagenet_val(num=5)

plt.figure(figsize=(12, 6))
for i in range(5):
    plt.subplot(1, 5, i + 1)
    plt.imshow(X[i])
    plt.title(class_names[y[i]])
    plt.axis('off')
plt.gcf().tight_layout()
```



## 3 Saliency Maps

Using this pretrained model, we will compute class saliency maps as described in Section 3.1 of [2].

A **saliency map** tells us the degree to which each pixel in the image affects the classification score for that image. To compute it, we compute the gradient of the unnormalized score corresponding to the correct class (which is a scalar) with respect to the pixels of the image. If the image has shape  $(3, H, W)$  then this gradient will also have shape  $(3, H, W)$ ; for each pixel in the image, this gradient tells us the amount by which the classification score will change if the pixel changes by a small amount. To compute the saliency map, we take the absolute value of this gradient, then take the maximum value over the 3 input channels; the final saliency map thus has shape  $(H, W)$  and all entries are nonnegative.

[2] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.

### 3.0.1 Hint: PyTorch gather method

Recall in Assignment 1 you needed to select one element from each row of a matrix; if  $s$  is an numpy array of shape  $(N, C)$  and  $y$  is a numpy array of shape  $(N,)$  containing integers  $0 \leq y[i] < C$ , then  $s[np.arange(N), y]$  is a numpy array of shape  $(N,)$  which selects one element from each element in  $s$  using the indices in  $y$ .

In PyTorch you can perform the same operation using the `gather()` method. If  $s$  is a PyTorch Tensor of shape  $(N, C)$  and  $y$  is a PyTorch Tensor of shape  $(N,)$  containing longs in the range  $0 \leq y[i] < C$ , then

```
s.gather(1, y.view(-1, 1)).squeeze()
```

will be a PyTorch Tensor of shape  $(N,)$  containing one entry from each row of  $s$ , selected according to the indices in  $y$ .

run the following cell to see an example.

You can also read the documentation for [the gather method](#) and [the squeeze method](#).

```
[6]: # Example of using gather to select one entry from each row in PyTorch
def gather_example():
    N, C = 4, 5
    s = torch.randn(N, C)
    y = torch.LongTensor([1, 2, 1, 3])
    print(s)
    print(y)
    print(s.gather(1, y.view(-1, 1)).squeeze())
gather_example()
```

```
tensor([[ 2.0176, -1.0338,  1.2604, -1.5676, -0.6635],
        [ 0.3531, -0.7488, -1.0099, -1.2633,  0.8685],
        [ 0.2895, -0.2216, -0.7186, -2.4654,  0.9633],
        [-2.3723,  0.4200, -1.0631, -0.1528,  1.3277]])
tensor([1, 2, 1, 3])
tensor([-1.0338, -1.0099, -0.2216, -0.1528])
```

```
[11]: def compute_saliency_maps(X, y, model):
    """
    Compute a class saliency map using the model for images X and labels y.

    Input:
    - X: Input images; Tensor of shape (N, 3, H, W)
    - y: Labels for X; LongTensor of shape (N,)
    - model: A pretrained CNN that will be used to compute the saliency map.

    Returns:
    - saliency: A Tensor of shape (N, H, W) giving the saliency maps for the
      input
      images.
    """
    # Make sure the model is in "test" mode
    model.eval()

    # Make input tensor require gradient
    X.requires_grad_()

    saliency = None
    #############################################################################
    # TODO: Implement this function. Perform a forward and backward pass through #
    # the model and return the gradients of the class saliency map with respect #
    # to the input.                                                             #
    #############################################################################
    # Replace "None" with your computed saliency map.                          #
    # saliency = None                                                          #
    #############################################################################
    # *****START OF YOUR CODE***** #
    # *****END OF YOUR CODE***** #

    # The returned saliency map needs to be a 2D tensor of shape (N, H, W).
    # You can implement this with a call to the .view() or .reshape() #
    # method. If you are not familiar with these may want to look up #
    # what they do in the PyTorch documentation.                         #
    # saliency = ...                                                       #
    # *****START OF YOUR CODE***** #
    # *****END OF YOUR CODE***** #

    return saliency
```

```
# the model to compute the gradient of the correct class score with respect to
# each input image. You first want to compute the loss over the correct
# scores (we'll combine losses across a batch by summing), and then compute
# the gradients with a backward pass.
#
#####
scores = model(X).gather(1, y.view(-1, 1)).squeeze()
scores.backward(torch.ones(scores.size()))
saliency = torch.max(X.grad.abs(), dim=1)[0]

#
#                                     END OF YOUR CODE
#
#####
return saliency
```

Once you have completed the implementation in the cell above, run the following to visualize some class saliency maps on our example images from the ImageNet validation set:

```
[12]: def show_saliency_maps(X, y):
    # Convert X and y from numpy arrays to Torch Tensors
    X_tensor = torch.cat([preprocess(Image.fromarray(x)) for x in X], dim=0)
    y_tensor = torch.LongTensor(y)

    # Compute saliency maps for images in X
    saliency = compute_saliency_maps(X_tensor, y_tensor, model)

    # Convert the saliency map from Torch Tensor to numpy array and show images
    # and saliency maps together.
    saliency = saliency.numpy()
    N = X.shape[0]
    for i in range(N):
        plt.subplot(2, N, i + 1)
        plt.imshow(X[i])
        plt.axis('off')
        plt.title(class_names[y[i]])
        plt.subplot(2, N, N + i + 1)
        plt.imshow(saliency[i], cmap=plt.cm.hot)
        plt.axis('off')
        plt.gcf().set_size_inches(12, 5)
    plt.show()

show_saliency_maps(X, y)
```



## 4 INLINE QUESTION

A friend of yours suggests that in order to find an image that maximizes the correct score, we can perform gradient ascent on the input image, but instead of the gradient we can actually use the saliency map in each step to update the image. Is this assertion true? Why or why not?

Answer: This assertion is not true. It is because for saliency takes the absolute over each channel prior to finding the maximum over them. Due to this, if we use saliency map instead of the gradients, updates during gradient ascent can only increase pixel values and not decrease them.

## 5 Fooling Images

We can also use image gradients to generate "fooling images" as discussed in [3]. Given an image and a target class, we can perform gradient **ascent** over the image to maximize the target class, stopping when the network classifies the image as the target class. Implement the following function to generate fooling images.

[3] Szegedy et al, "Intriguing properties of neural networks", ICLR 2014

```
[15]: def make_fooling_image(X, target_y, model):
    """
    Generate a fooling image that is close to X, but that the model classifies
    as target_y.

    Inputs:
    - X: Input image; Tensor of shape (1, 3, 224, 224)
    - target_y: An integer in the range [0, 1000)
    - model: A pretrained CNN

    Returns:
    - X_fooling: An image that is close to X, but that is classified as target_y
    by the model.
    """

    # Your code here
    pass
```

```

"""
# Initialize our fooling image to the input image, and make it require gradient
X_fooling = X.clone()
X_fooling.requires_grad_()

learning_rate = 1

#####
# TODO: Generate a fooling image X_fooling that the model will classify as the # target class target_y. You should perform gradient ascent on the score of the # target class, stopping when the model is fooled.

# When computing an update step, first normalize the gradient:
# dX = learning_rate * g / ||g||_2

# You should write a training loop.

# Hint: For most examples, you should be able to generate a fooling image in fewer than 100 iterations of gradient ascent.

# You can print your progress over iterations to check your algorithm.

#####

scores = model(X_fooling)
class_y = torch.argmax(scores, dim=1)[0]
while(class_y!=target_y):
    scores[:, target_y].backward()
    g = X_fooling.grad.data
    dX = learning_rate * g/g.norm()
    X_fooling.data += dX.data
    X_fooling.grad.data.zero_()
    scores = model(X_fooling)
    class_y = torch.argmax(scores, dim=1)[0]

#####

```

```

#                                     END OF YOUR CODE
→ #
└
→#####
return X_fooling

```

Run the following cell to generate a fooling image. You should ideally see at first glance no major difference between the original and fooling images, and the network should now make an incorrect prediction on the fooling one. However you should see a bit of random noise if you look at the 10x magnified difference between the original and fooling images. Feel free to change the idx variable to explore other images.

```
[19]: idx = 0
target_y = 6

X_tensor = torch.cat([preprocess(Image.fromarray(x)) for x in X], dim=0)
X_fooling = make_fooling_image(X_tensor[idx:idx+1], target_y, model)

scores = model(X_fooling)
assert target_y == scores.data.max(1)[0].item(), 'The model is not fooled!'
```

After generating a fooling image, run the following cell to visualize the original image, the fooling image, as well as the difference between them.

```
[20]: X_fooling_np = deprocess(X_fooling.clone())
X_fooling_np = np.asarray(X_fooling_np).astype(np.uint8)

plt.subplot(1, 4, 1)
plt.imshow(X[idx])
plt.title(class_names[y[idx]])
plt.axis('off')

plt.subplot(1, 4, 2)
plt.imshow(X_fooling_np)
plt.title(class_names[target_y])
plt.axis('off')

plt.subplot(1, 4, 3)
X_pre = preprocess(Image.fromarray(X[idx]))
diff = np.asarray(deprocess(X_fooling - X_pre, should_rescale=False))
plt.imshow(diff)
plt.title('Difference')
plt.axis('off')

plt.subplot(1, 4, 4)
diff = np.asarray(deprocess(10 * (X_fooling - X_pre), should_rescale=False))
plt.imshow(diff)
plt.title('Magnified difference (10x)')
plt.axis('off')
```

```
plt.gcf().set_size_inches(12, 5)
plt.show()
```



## 6 Class visualization

By starting with a random noise image and performing gradient ascent on a target class, we can generate an image that the network will recognize as the target class. This idea was first presented in [2]; [3] extended this idea by suggesting several regularization techniques that can improve the quality of the generated image.

Concretely, let  $I$  be an image and let  $y$  be a target class. Let  $s_y(I)$  be the score that a convolutional network assigns to the image  $I$  for class  $y$ ; note that these are raw unnormalized scores, not class probabilities. We wish to generate an image  $I^*$  that achieves a high score for the class  $y$  by solving the problem

$$I^* = \arg \max_I (s_y(I) - R(I))$$

where  $R$  is a (possibly implicit) regularizer (note the sign of  $R(I)$  in the argmax: we want to minimize this regularization term). We can solve this optimization problem using gradient ascent, computing gradients with respect to the generated image. We will use (explicit) L2 regularization of the form

$$R(I) = \lambda \|I\|_2^2$$

and implicit regularization as suggested by [3] by periodically blurring the generated image. We can solve this problem using gradient ascent on the generated image.

In the cell below, complete the implementation of the `create_class_visualization` function.

[2] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.

[3] Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML 2015 Deep Learning Workshop

```
[21]: def jitter(X, ox, oy):
    """
    Helper function to randomly jitter an image.
    
```

*Inputs*

```
- X: PyTorch Tensor of shape (N, C, H, W)
- ox, oy: Integers giving number of pixels to jitter along W and H axes
```

Returns: A new PyTorch Tensor of shape (N, C, H, W)

```
"""
```

```
if ox != 0:
    left = X[:, :, :, :-ox]
    right = X[:, :, :, -ox:]
    X = torch.cat([right, left], dim=3)
if oy != 0:
    top = X[:, :, :-oy]
    bottom = X[:, :, -oy:]
    X = torch.cat([bottom, top], dim=2)
return X
```

[22]: `def create_class_visualization(target_y, model, dtype, **kwargs):`

```
"""
```

Generate an image to maximize the score of target\_y under a pretrained  $\rightarrow$ model.

Inputs:

```
- target_y: Integer in the range [0, 1000) giving the index of the class
- model: A pretrained CNN that will be used to generate the image
- dtype: Torch datatype to use for computations
```

Keyword arguments:

```
- l2_reg: Strength of L2 regularization on the image
- learning_rate: How big of a step to take
- num_iterations: How many iterations to use
- blur_every: How often to blur the image as an implicit regularizer
- max_jitter: How much to jitter the image as an implicit regularizer
- show_every: How often to show the intermediate result
"""
```

```
model.type(dtype)
```

```
l2_reg = kwargs.pop('l2_reg', 1e-3)
learning_rate = kwargs.pop('learning_rate', 25)
num_iterations = kwargs.pop('num_iterations', 100)
blur_every = kwargs.pop('blur_every', 10)
max_jitter = kwargs.pop('max_jitter', 16)
show_every = kwargs.pop('show_every', 25)
```

# Randomly initialize the image as a PyTorch Tensor, and make it requires  $\rightarrow$ gradient.

```
img = torch.randn(1, 3, 224, 224).mul_(1.0).type(dtype).requires_grad_()
```

```
for t in range(num_iterations):
```

# Randomly jitter the image a bit; this gives slightly nicer results

```

    ox, oy = random.randint(0, max_jitter), random.randint(0, max_jitter)
    img.data.copy_(jitter(img.data, ox, oy))

    # TODO: Use the model to compute the gradient of the score for the
    # class target_y with respect to the pixels of the image, and make a
    # gradient step on the image using the learning rate. Don't forget the
    # L2 regularization term!
    # Be very careful about the signs of elements in your code.

    scores = model(img)
    scores[:, target_y].backward()
    g = img.grad.data - 2*l2_reg*img.data
    img.data += learning_rate*g/g.norm()
    img.grad.zero_()

    # END OF YOUR CODE

    # Undo the random jitter
    img.data.copy_(jitter(img.data, -ox, -oy))

    # As regularizer, clamp and periodically blur the image
    for c in range(3):
        lo = float(-SQUEEZENET_MEAN[c] / SQUEEZENET_STD[c])
        hi = float((1.0 - SQUEEZENET_MEAN[c]) / SQUEEZENET_STD[c])
        img.data[:, c].clamp_(min=lo, max=hi)
    if t % blur_every == 0:
        blur_image(img.data, sigma=0.5)

    # Periodically show the image
    if t == 0 or (t + 1) % show_every == 0 or t == num_iterations - 1:
        plt.imshow(deprocess(img.data.clone().cpu()))
        class_name = class_names[target_y]
        plt.title('%s\nIteration %d / %d' % (class_name, t + 1,
                                          num_iterations))

```

```
plt.gcf().set_size_inches(4, 4)
plt.axis('off')
plt.show()

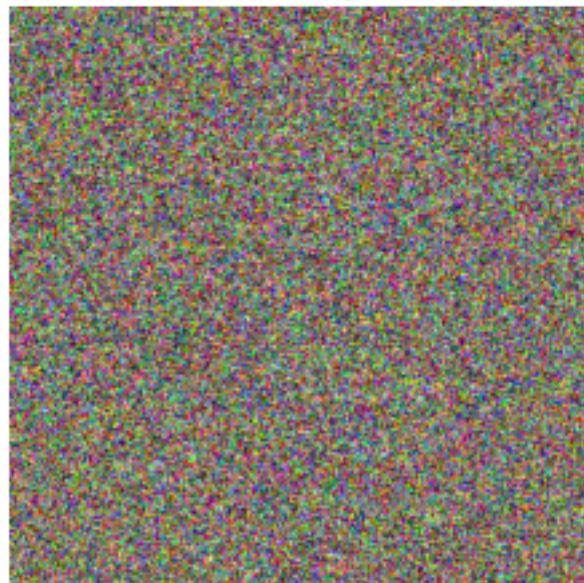
return deprocess(img.data.cpu())
```

Once you have completed the implementation in the cell above, run the following cell to generate an image of a Tarantula:

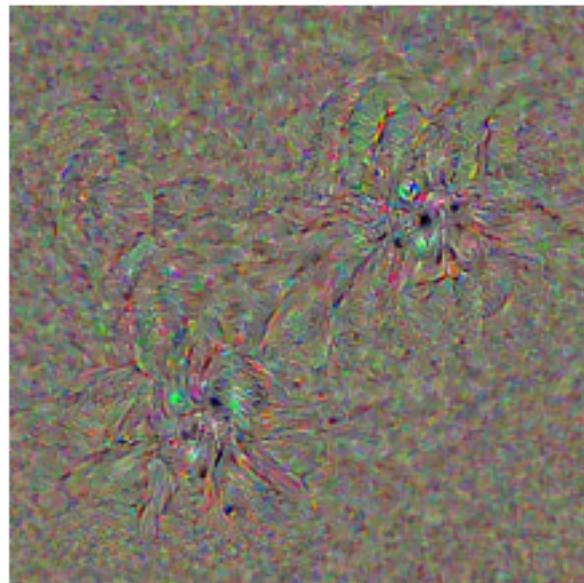
```
[23]: dtype = torch.FloatTensor
# dtype = torch.cuda.FloatTensor # Uncomment this to use GPU
model.type(dtype)

target_y = 76 # Tarantula
# target_y = 78 # Tick
# target_y = 187 # Yorkshire Terrier
# target_y = 683 # Oboe
# target_y = 366 # Gorilla
# target_y = 604 # Hourglass
out = create_class_visualization(target_y, model, dtype)
```

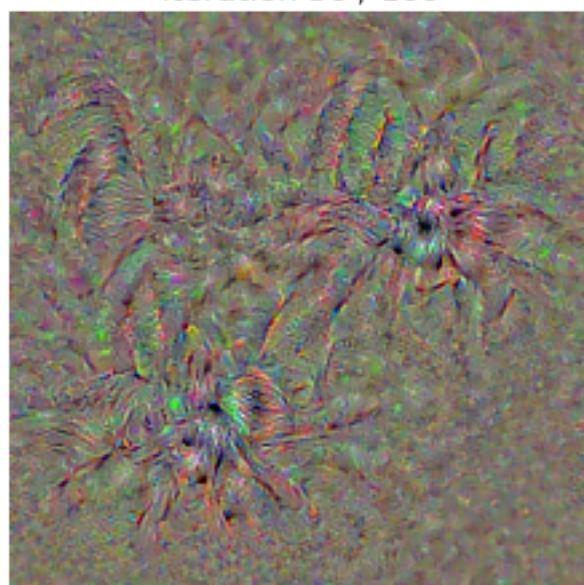
tarantula  
Iteration 1 / 100



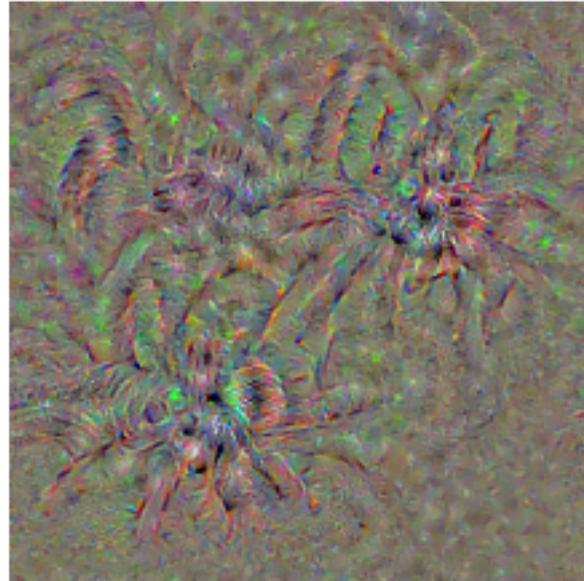
tarantula  
Iteration 25 / 100



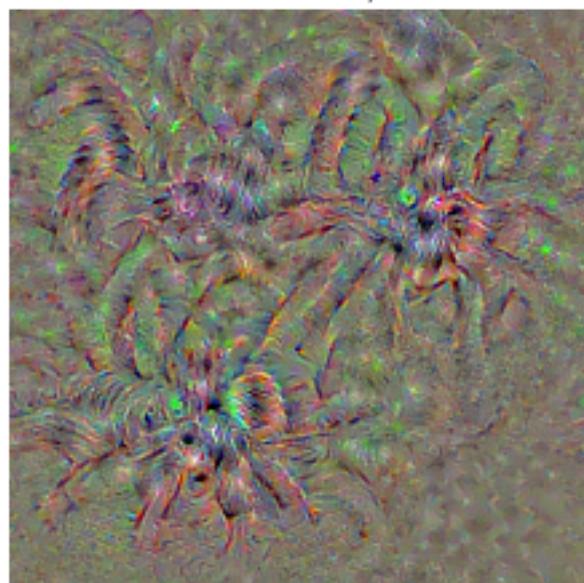
tarantula  
Iteration 50 / 100



tarantula  
Iteration 75 / 100



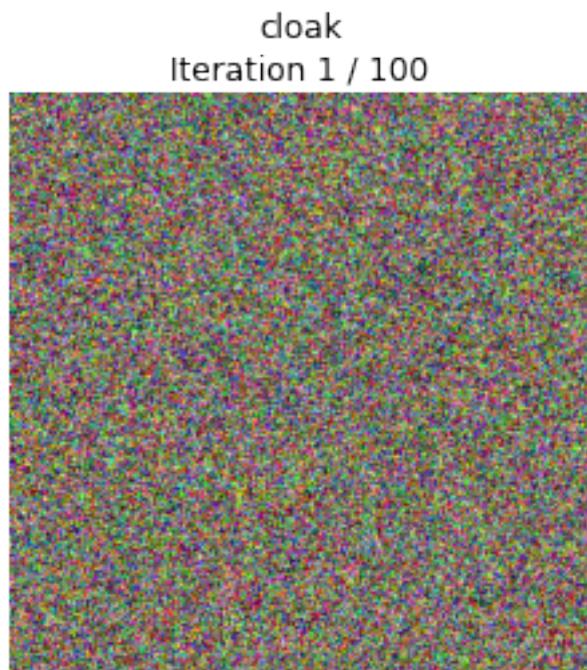
tarantula  
Iteration 100 / 100



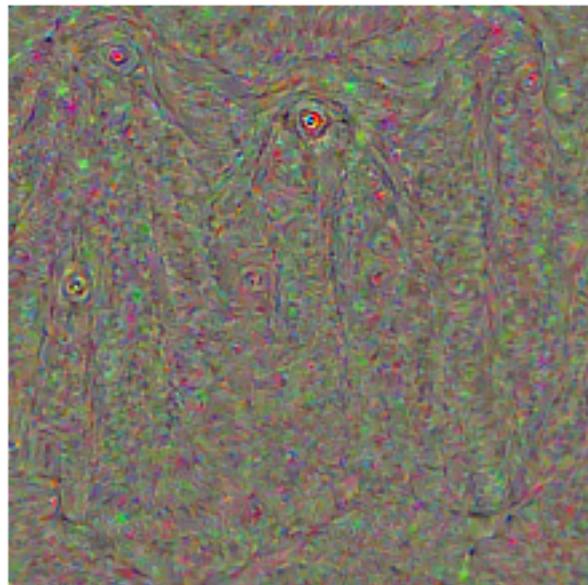
Try out your class visualization on other classes! You should also feel free to play with various hyperparameters to try and improve the quality of the generated image, but this is not required.

```
[26]: # target_y = 78 # Tick
# target_y = 187 # Yorkshire Terrier
# target_y = 683 # Oboe
target_y = 366 # Gorilla
# target_y = 604 # Hourglass
target_y = np.random.randint(1000)
print(class_names[target_y])
X = create_class_visualization(target_y, model, dtype)
```

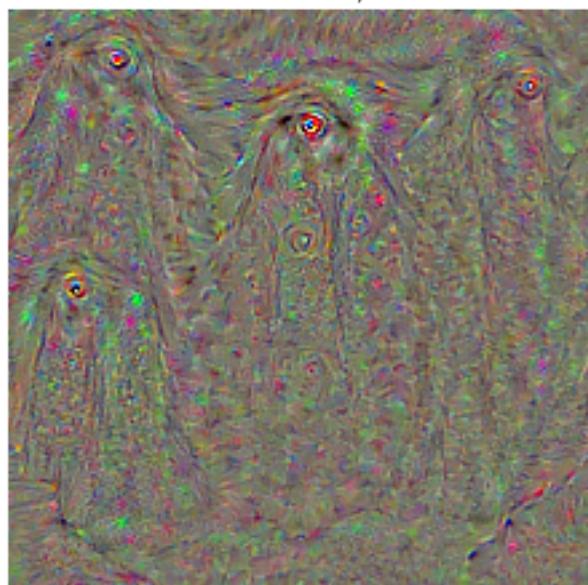
cloak



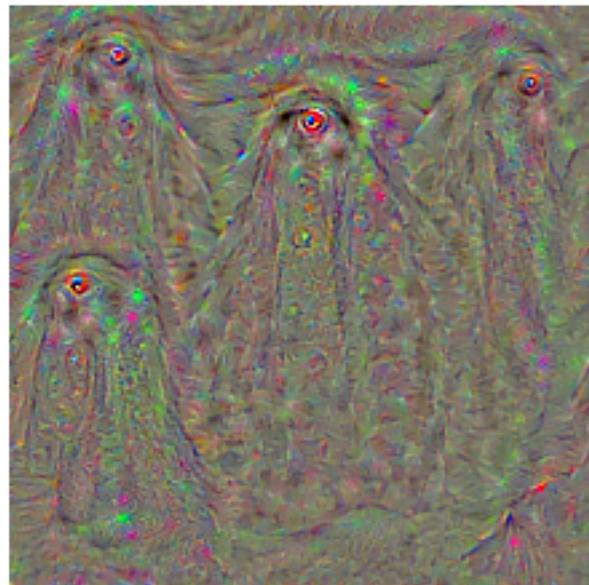
cloak  
Iteration 25 / 100



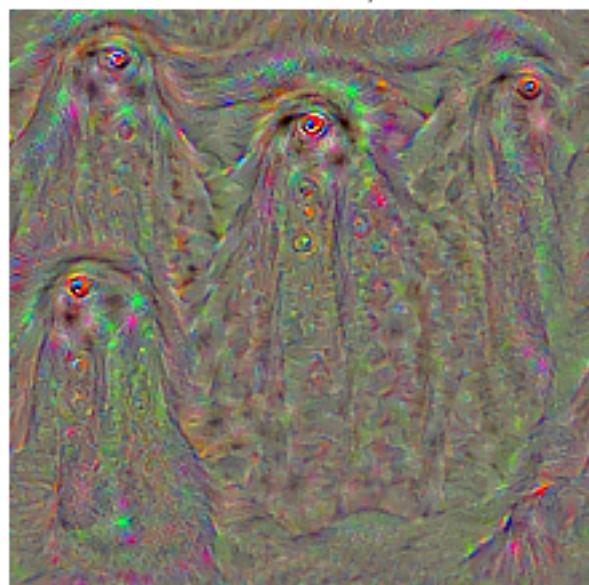
cloak  
Iteration 50 / 100



cloak  
Iteration 75 / 100



cloak  
Iteration 100 / 100



# GANs-PyTorch

December 8, 2021

```
[ ]: # This mounts your Google Drive to the Colab VM.
from google.colab import drive
drive.mount('/content/drive')

# TODO: Enter the foldername in your Drive where you have saved the unzipped
# assignment folder, e.g. 'cs231n/assignments/assignment1/'
FOLDERNAME = 'CS682/assignment3'
assert FOLDERNAME is not None, "[!] Enter the foldername."

# Now that we've mounted your Drive, this ensures that
# the Python interpreter of the Colab VM can load
# python files from within it.
import sys
sys.path.append('/content/drive/My Drive/{}'.format(FOLDERNAME))

# This downloads the CIFAR-10 dataset to your Drive
# if it doesn't already exist.
%cd /content/drive/My\ Drive/$FOLDERNAME/cs682/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME

[ ]: pip install scipy==1.1.0
```

## 1 Generative Adversarial Networks (GANs)

So far in cs682, all the applications of neural networks that we have explored have been **discriminative models** that take an input and are trained to produce a labeled output. This has ranged from straightforward classification of image categories to sentence generation (which was still phrased as a classification problem, our labels were in vocabulary space and we'd learned a recurrence to capture multi-word labels). In this notebook, we will expand our repertoire, and build **generative models** using neural networks. Specifically, we will learn how to build models which generate novel images that resemble a set of training images.

### 1.0.1 What is a GAN?

In 2014, [Goodfellow et al.](#) presented a method for training generative models called Generative Adversarial Networks (GANs for short). In a GAN, we build two different neural networks. Our

first network is a traditional classification network, called the **discriminator**. We will train the discriminator to take images, and classify them as being real (belonging to the training set) or fake (not present in the training set). Our other network, called the **generator**, will take random noise as input and transform it using a neural network to produce images. The goal of the generator is to fool the discriminator into thinking the images it produced are real.

We can think of this back and forth process of the generator ( $G$ ) trying to fool the discriminator ( $D$ ), and the discriminator trying to correctly classify real vs. fake as a minimax game:

$$\underset{G}{\text{minimize}} \underset{D}{\text{maximize}} \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))]$$

where  $z \sim p(z)$  are the random noise samples,  $G(z)$  are the generated images using the neural network generator  $G$ , and  $D$  is the output of the discriminator, specifying the probability of an input being real. In [Goodfellow et al.](#), they analyze this minimax game and show how it relates to minimizing the Jensen-Shannon divergence between the training data distribution and the generated samples from  $G$ .

To optimize this minimax game, we will alternate between taking gradient *descent* steps on the objective for  $G$ , and gradient *ascent* steps on the objective for  $D$ : 1. update the **generator** ( $G$ ) to minimize the probability of the **discriminator making the correct choice**. 2. update the **discriminator** ( $D$ ) to maximize the probability of the **discriminator making the correct choice**.

While these updates are useful for analysis, they do not perform well in practice. Instead, we will use a different objective when we update the generator: maximize the probability of the **discriminator making the incorrect choice**. This small change helps to alleviate problems with the generator gradient vanishing when the discriminator is confident. This is the standard update used in most GAN papers, and was used in the original paper from [Goodfellow et al.](#).

In this assignment, we will alternate the following updates: 1. Update the generator ( $G$ ) to maximize the probability of the discriminator making the incorrect choice on generated data:

$$\underset{G}{\text{maximize}} \mathbb{E}_{z \sim p(z)} [\log D(G(z))]$$

2. Update the discriminator ( $D$ ), to maximize the probability of the discriminator making the correct choice on real and generated data:

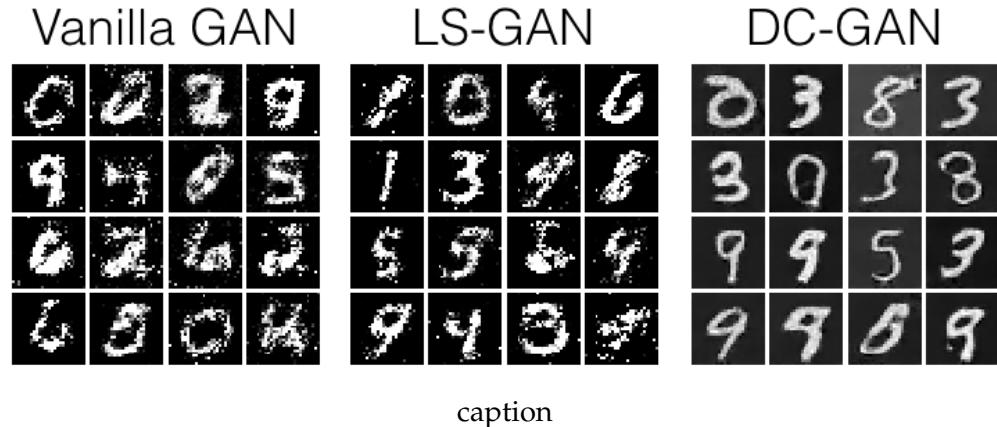
$$\underset{D}{\text{maximize}} \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))]$$

### 1.0.2 What else is there?

Since 2014, GANs have exploded into a huge research area, with massive [workshops](#), and [hundreds of new papers](#). Compared to other approaches for generative models, they often produce the highest quality samples but are some of the most difficult and finicky models to train (see [this github repo](#) that contains a set of 17 hacks that are useful for getting models working). Improving the stability and robustness of GAN training is an open research question, with new papers coming out every day! For a more recent tutorial on GANs, see [here](#). There is also some even more recent exciting work that changes the objective function to Wasserstein distance and yields much more stable results across model architectures: [WGAN](#), [WGAN-GP](#).

GANs are not the only way to train a generative model! For other approaches to generative modeling check out the [deep generative model chapter](#) of the Deep Learning [book](#). Another popular way of training neural networks as generative models is Variational Autoencoders (co-discovered [here](#) and [here](#)). Variational autoencoders combine neural networks with variational inference to train deep generative models. These models tend to be far more stable and easier to train but currently don't produce samples that are as pretty as GANs.

Here's an example of what your outputs from the 3 different models you're going to train should look like... note that GANs are sometimes finicky, so your outputs might not look exactly like this... this is just meant to be a *rough* guideline of the kind of quality you can expect:



## 1.1 Setup

```
[2]: import torch
import torch.nn as nn
from torch.nn import init
import torchvision
import torchvision.transforms as T
import torch.optim as optim
from torch.utils.data import DataLoader
from torch.utils.data import sampler
import torchvision.datasets as dset

import numpy as np

import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

def show_images(images):
    images = np.reshape(images, [images.shape[0], -1]) # images reshape to (batch_size, D)
    sq rtn = int(np.ceil(np.sqrt(images.shape[0])))
    sq rtimg = int(np.ceil(np.sqrt(images.shape[1])))

    fig = plt.figure(figsize=(sq rtn, sq rtn))
    gs = gridspec.GridSpec(sq rtn, sq rtn)
```

```

gs.update(wspace=0.05, hspace=0.05)

for i, img in enumerate(images):
    ax = plt.subplot(gs[i])
    plt.axis('off')
    ax.set_xticklabels([])
    ax.set_yticklabels([])
    ax.set_aspect('equal')
    plt.imshow(img.reshape([sqrtimg,sqrtimg]))
return

def preprocess_img(x):
    return 2 * x - 1.0

def deprocess_img(x):
    return (x + 1.0) / 2.0

def rel_error(x,y):
    return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y)))))

def count_params(model):
    """Count the number of parameters in the current TensorFlow graph """
    param_count = np.sum([np.prod(p.size()) for p in model.parameters()])
    return param_count

answers = dict(np.load('gan-checks-tf.npz'))

```

## 1.2 Dataset

GANs are notoriously finicky with hyperparameters, and also require many training epochs. In order to make this assignment approachable without a GPU, we will be working on the MNIST dataset, which is 60,000 training and 10,000 test images. Each picture contains a centered image of white digit on black background (0 through 9). This was one of the first datasets used to train convolutional neural networks and it is fairly easy -- a standard CNN model can easily exceed 99% accuracy.

To simplify our code here, we will use the PyTorch MNIST wrapper, which downloads and loads the MNIST dataset. See the [documentation](#) for more information about the interface. The default parameters will take 5,000 of the training examples and place them into a validation dataset. The data will be saved into a folder called `MNIST_data`.

```
[4]: class ChunkSampler(sampler.Sampler):
    """Samples elements sequentially from some offset.
    Arguments:
        num_samples: # of desired datapoints
        start: offset where we should start selecting from
    """
    def __init__(self, num_samples, start=0):
```

```

    self.num_samples = num_samples
    self.start = start

    def __iter__(self):
        return iter(range(self.start, self.start + self.num_samples))

    def __len__(self):
        return self.num_samples

NUM_TRAIN = 50000
NUM_VAL = 5000

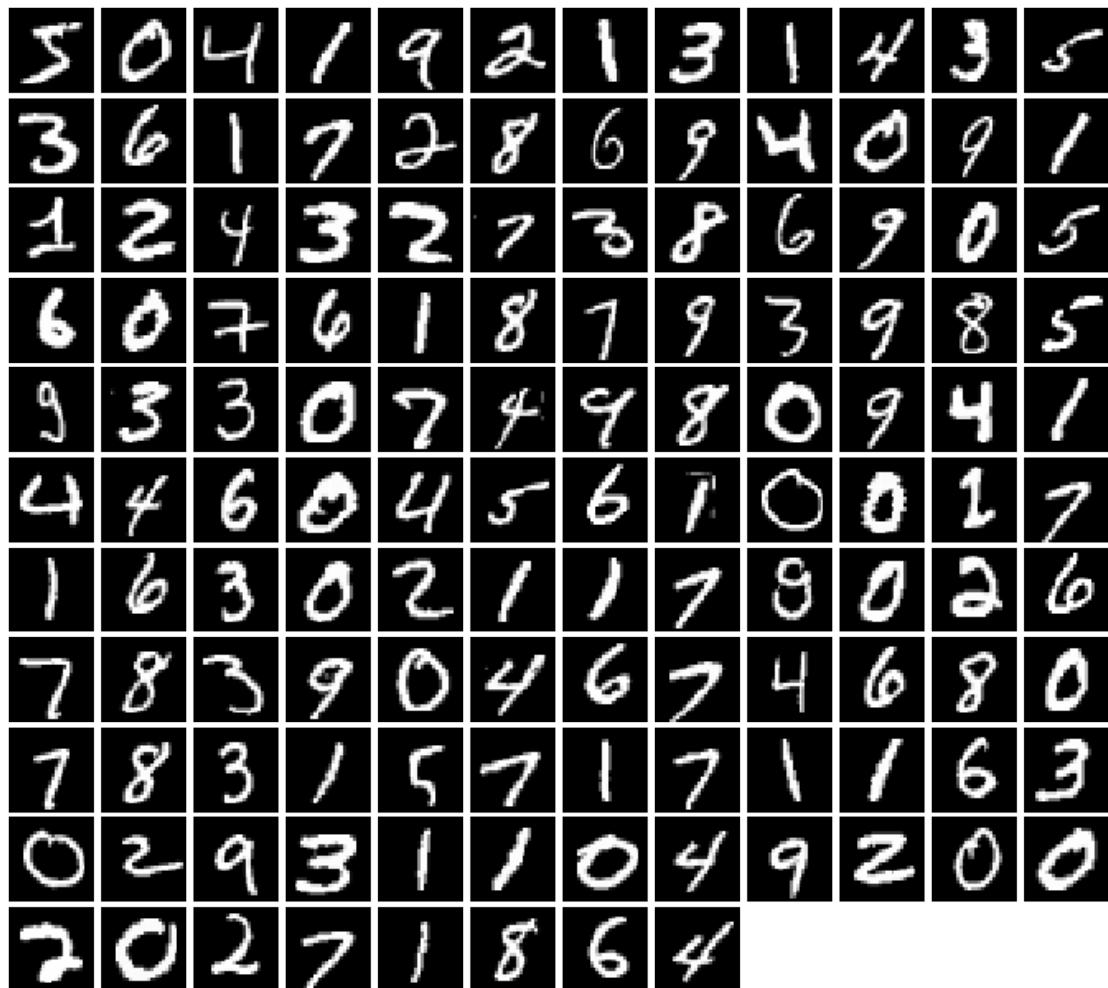
NOISE_DIM = 96
batch_size = 128

mnist_train = dset.MNIST('./cs682/datasets/MNIST_data', train=True,
    download=True,
    transform=T.ToTensor())
loader_train = DataLoader(mnist_train, batch_size=batch_size,
    sampler=ChunkSampler(NUM_TRAIN, 0))

mnist_val = dset.MNIST('./cs682/datasets/MNIST_data', train=True, download=True,
    transform=T.ToTensor())
loader_val = DataLoader(mnist_val, batch_size=batch_size,
    sampler=ChunkSampler(NUM_VAL, NUM_TRAIN))

imgs = loader_train.__iter__().next()[0].view(batch_size, 784).numpy().squeeze()
show_images(imgs)

```



### 1.3 Random Noise

Generate uniform noise from -1 to 1 with shape [batch\_size, dim].

Hint: use `torch.rand`.

```
[5]: def sample_noise(batch_size, dim):  
    """
```

*Generate a PyTorch Tensor of uniform random noise.*

*Input:*

- *batch\_size*: Integer giving the batch size of noise to generate.
- *dim*: Integer giving the dimension of noise to generate.

*Output:*

- A PyTorch Tensor of shape (batch\_size, dim) containing uniform random noise in the range (-1, 1).

```
    return (torch.rand(batch_size, dim)*2)-1
```

Make sure noise is the correct shape and type:

```
[6]: def test_sample_noise():
    batch_size = 3
    dim = 4
    torch.manual_seed(231)
    z = sample_noise(batch_size, dim)
    np_z = z.cpu().numpy()
    assert np_z.shape == (batch_size, dim)
    assert torch.is_tensor(z)
    assert np.all(np_z >= -1.0) and np.all(np_z <= 1.0)
    assert np.any(np_z < 0.0) and np.any(np_z > 0.0)
    print('All tests passed!')

test_sample_noise()
```

All tests passed!

## 1.4 Flatten

Recall our Flatten operation from previous notebooks... this time we also provide an Unflatten, which you might want to use when implementing the convolutional generator. We also provide a weight initializer (and call it for you) that uses Xavier initialization instead of PyTorch's uniform default.

```
[7]: class Flatten(nn.Module):
    def forward(self, x):
        N, C, H, W = x.size() # read in N, C, H, W
        return x.view(N, -1) # "flatten" the C * H * W values into a single vector per image

class Unflatten(nn.Module):
    """
    An Unflatten module receives an input of shape (N, C*H*W) and reshapes it
    to produce an output of shape (N, C, H, W).
    """
    def __init__(self, N=-1, C=128, H=7, W=7):
        super(Unflatten, self).__init__()
        self.N = N
        self.C = C
        self.H = H
        self.W = W
    def forward(self, x):
        return x.view(self.N, self.C, self.H, self.W)

def initialize_weights(m):
    if isinstance(m, nn.Linear) or isinstance(m, nn.ConvTranspose2d):
```

```
init.xavier_uniform_(m.weight.data)
```

## 1.5 CPU / GPU

By default all code will run on CPU. GPUs are not needed for this assignment, but will help you to train your models faster. If you do want to run the code on a GPU, then change the `dtype` variable in the following cell.

```
[8]: dtype = torch.FloatTensor  
#dtype = torch.cuda.FloatTensor ## UNCOMMENT THIS LINE IF YOU'RE ON A GPU!
```

## 2 Discriminator

Our first step is to build a discriminator. Fill in the architecture as part of the `nn.Sequential` constructor in the function below. All fully connected layers should include bias terms. The architecture is:

- \* Fully connected layer with input size 784 and output size 256
- \* LeakyReLU with alpha 0.01
- \* Fully connected layer with input\_size 256 and output size 256
- \* LeakyReLU with alpha 0.01
- \* Fully connected layer with input size 256 and output size 1

Recall that the Leaky ReLU nonlinearity computes  $f(x) = \max(\alpha x, x)$  for some fixed constant  $\alpha$ ; for the LeakyReLU nonlinearities in the architecture above we set  $\alpha = 0.01$ .

The output of the discriminator should have shape `[batch_size, 1]`, and contain real numbers corresponding to the scores that each of the `batch_size` inputs is a real image.

```
[15]: def discriminator():  
    """  
        Build and return a PyTorch model implementing the architecture above.  
    """  
  
    model = nn.Sequential(  
        Flatten(),  
        nn.Linear(784, 256),  
        nn.LeakyReLU(0.01),  
        nn.Linear(256, 256),  
        nn.LeakyReLU(0.01),  
        nn.Linear(256, 1)  
    )  
    return model
```

Test to make sure the number of parameters in the discriminator is correct:

```
[16]: def test_discriminator(true_count=267009):  
    model = discriminator()  
    cur_count = count_params(model)  
    if cur_count != true_count:  
        print('Incorrect number of parameters in discriminator. Check your  
        →architecture.')  
    else:  
        print('Correct number of parameters in discriminator.')
```

```
test_discriminator()
```

Correct number of parameters in discriminator.

### 3 Generator

Now to build the generator network:

- \* Fully connected layer from noise\_dim to 1024
- \* ReLU
- \* Fully connected layer with size 1024
- \* ReLU
- \* Fully connected layer with size 784
- \* TanH (to clip the image to be in the range of [-1,1])

```
[17]: def generator(noise_dim=NOISE_DIM):  
    """  
        Build and return a PyTorch model implementing the architecture above.  
    """  
  
    model = nn.Sequential(  
        nn.Linear(noise_dim, 1024),  
        nn.ReLU(),  
        nn.Linear(1024, 1024),  
        nn.ReLU(),  
        nn.Linear(1024, 784),  
        nn.Tanh()  
    )  
  
    return model
```

Test to make sure the number of parameters in the generator is correct:

```
[18]: def test_generator(true_count=1858320):  
    model = generator(4)  
    cur_count = count_params(model)  
    if cur_count != true_count:  
        print('Incorrect number of parameters in generator. Check your  
→architecture.')  
    else:  
        print('Correct number of parameters in generator.')  
  
test_generator()
```

Correct number of parameters in generator.

### 4 GAN Loss

Compute the generator and discriminator loss. The generator loss is:

$$\ell_G = -\mathbb{E}_{z \sim p(z)} [\log D(G(z))]$$

and the discriminator loss is:

$$\ell_D = -\mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] - \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))]$$

Note that these are negated from the equations presented earlier as we will be *minimizing* these losses.

**HINTS:** You should use the `bce_loss` function defined below to compute the binary cross entropy loss which is needed to compute the log probability of the true label given the logits output from the discriminator. Given a score  $s \in \mathbb{R}$  and a label  $y \in \{0, 1\}$ , the binary cross entropy loss is

$$bce(s, y) = -y * \log(s) - (1 - y) * \log(1 - s)$$

A naive implementation of this formula can be numerically unstable, so we have provided a numerically stable implementation for you below.

You will also need to compute labels corresponding to real or fake and use the `logit` arguments to determine their size. Make sure you cast these labels to the correct data type using the global `dtype` variable, for example:

```
true_labels = torch.ones(size).type(dtype)
```

Instead of computing the expectation of  $\log D(G(z))$ ,  $\log D(x)$  and  $\log(1 - D(G(z)))$ , we will be averaging over elements of the minibatch, so make sure to combine the loss by averaging instead of summing.

```
[19]: def bce_loss(input, target):
```

```
    """
```

*Numerically stable version of the binary cross-entropy loss function.*

*As per <https://github.com/pytorch/pytorch/issues/751>*

*See the TensorFlow docs for a derivation of this formula:*

*[https://www.tensorflow.org/api\\_docs/python/tf/nn/sigmoid\\_cross\\_entropy\\_with\\_logits](https://www.tensorflow.org/api_docs/python/tf/nn/sigmoid_cross_entropy_with_logits)*

*Inputs:*

- `input`: PyTorch Tensor of shape  $(N, )$  giving scores.
- `target`: PyTorch Tensor of shape  $(N, )$  containing 0 and 1 giving targets.

*Returns:*

*- A PyTorch Tensor containing the mean BCE loss over the minibatch of input data.*

```
    """
```

```
    neg_abs = - input.abs()
```

```
    loss = input.clamp(min=0) - input * target + (1 + neg_abs.exp()).log()
```

```
    return loss.mean()
```

```
[26]: def discriminator_loss(logits_real, logits_fake):
```

```
    """
```

*Computes the discriminator loss described above.*

*Inputs:*

- `logits_real`: PyTorch Tensor of shape  $(N, )$  giving scores for the real data.
- `logits_fake`: PyTorch Tensor of shape  $(N, )$  giving scores for the fake data.

*Returns:*

```

- loss: PyTorch Tensor containing (scalar) the loss for the discriminator.
"""

N = logits_real.size()[0]
loss = bce_loss(logits_real, torch.ones(N))+bce_loss(logits_fake, torch.
→zeros(N))
return loss

def generator_loss(logits_fake):
"""
Computes the generator loss described above.

Inputs:
- logits_fake: PyTorch Tensor of shape (N,) giving scores for the fake data.

Returns:
- loss: PyTorch Tensor containing the (scalar) loss for the generator.
"""

N = logits_fake.size()[0]
loss = bce_loss(logits_fake, torch.ones(N))
return loss

```

Test your generator and discriminator loss. You should see errors < 1e-7.

```
[23]: def test_discriminator_loss(logits_real, logits_fake, d_loss_true):
    d_loss = discriminator_loss(torch.Tensor(logits_real).type(dtype),
                                torch.Tensor(logits_fake).type(dtype)).cpu().
→numpy()
    print("Maximum error in d_loss: %g"%rel_error(d_loss_true, d_loss))

test_discriminator_loss(answers['logits_real'], answers['logits_fake'],
                        answers['d_loss_true'])
```

Maximum error in d\_loss: 3.97058e-09

```
[27]: def test_generator_loss(logits_fake, g_loss_true):
    g_loss = generator_loss(torch.Tensor(logits_fake).type(dtype)).cpu().numpy()
    print("Maximum error in g_loss: %g"%rel_error(g_loss_true, g_loss))

test_generator_loss(answers['logits_fake'], answers['g_loss_true'])
```

Maximum error in g\_loss: 4.4518e-09

## 5 Optimizing our loss

Make a function that returns an `optim.Adam` optimizer for the given model with a 1e-3 learning rate, `beta1=0.5`, `beta2=0.999`. You'll use this to construct optimizers for the generators and discriminators for the rest of the notebook.

```
[28]: def get_optimizer(model):
    """
    Construct and return an Adam optimizer for the model with learning rate ↪
    →1e-3,
    beta1=0.5, and beta2=0.999.

    Input:
    - model: A PyTorch model that we want to optimize.

    Returns:
    - An Adam optimizer for the model with the desired hyperparameters.
    """
optimizer = optim.Adam(model.parameters(), lr=1e-3, betas=(0.5, 0.999))
return optimizer
```

## 6 Training a GAN!

We provide you the main training loop... you won't need to change this function, but we encourage you to read through and understand it.

```
[29]: def run_a_gan(D, G, D_solver, G_solver, discriminator_loss, generator_loss, ↪
    →show_every=250,
    batch_size=128, noise_size=96, num_epochs=10):
    """
    Train a GAN!

    Inputs:
    - D, G: PyTorch models for the discriminator and generator
    - D_solver, G_solver: torch.optim Optimizers to use for training the
      discriminator and generator.
    - discriminator_loss, generator_loss: Functions to use for computing the
      generator and
      discriminator loss, respectively.
    - show_every: Show samples after every show_every iterations.
    - batch_size: Batch size to use for training.
    - noise_size: Dimension of the noise to use as input to the generator.
    - num_epochs: Number of epochs over the training dataset to use for
      training.
    """
    iter_count = 0
    for epoch in range(num_epochs):
        for x, _ in loader_train:
            if len(x) != batch_size:
                continue
            D_solver.zero_grad()
            real_data = x.type(dtype)
```

```

logits_real = D(2* (real_data - 0.5)).type(dtype)

g_fake_seed = sample_noise(batch_size, noise_size).type(dtype)
fake_images = G(g_fake_seed).detach()
logits_fake = D(fake_images.view(batch_size, 1, 28, 28))

d_total_error = discriminator_loss(logits_real, logits_fake)
d_total_error.backward()
D_solver.step()

G_solver.zero_grad()
g_fake_seed = sample_noise(batch_size, noise_size).type(dtype)
fake_images = G(g_fake_seed)

gen_logits_fake = D(fake_images.view(batch_size, 1, 28, 28))
g_error = generator_loss(gen_logits_fake)
g_error.backward()
G_solver.step()

if (iter_count % show_every == 0):
    print('Iter: {}, D: {:.4}, G:{:.4}'.format(iter_count,d_total_error.item(),g_error.item()))
    imgs_numpy = fake_images.data.cpu().numpy()
    show_images(imgs_numpy[0:16])
    plt.show()
    print()
    iter_count += 1

```

[30]: # Make the discriminator

```

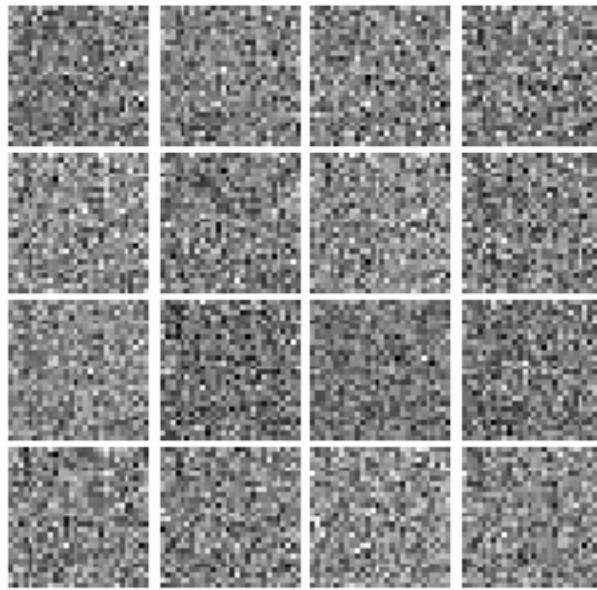
D = discriminator().type(dtype)

# Make the generator
G = generator().type(dtype)

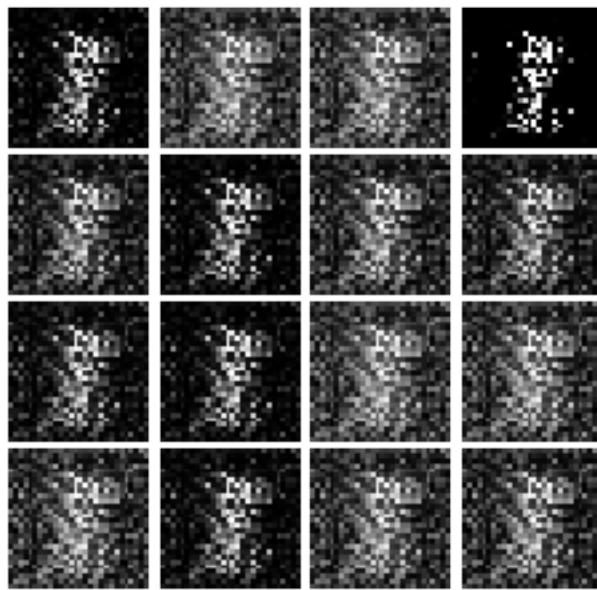
# Use the function you wrote earlier to get optimizers for the Discriminator and the Generator
D_solver = get_optimizer(D)
G_solver = get_optimizer(G)
# Run it!
run_a_gan(D, G, D_solver, G_solver, discriminator_loss, generator_loss)

```

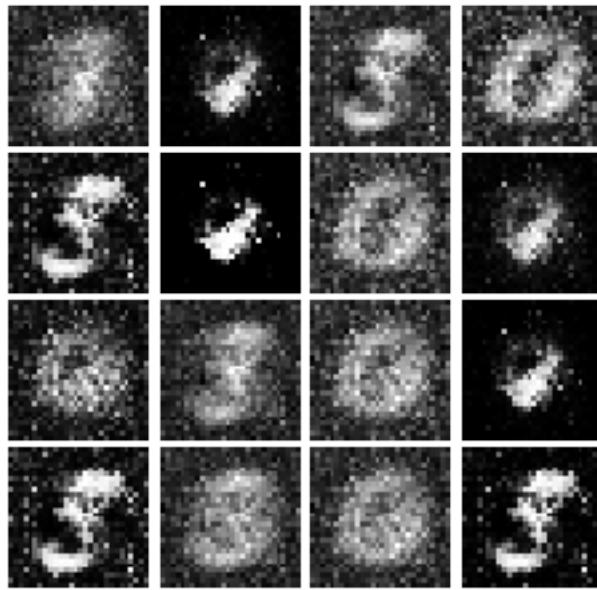
Iter: 0, D: 1.395, G:0.6751



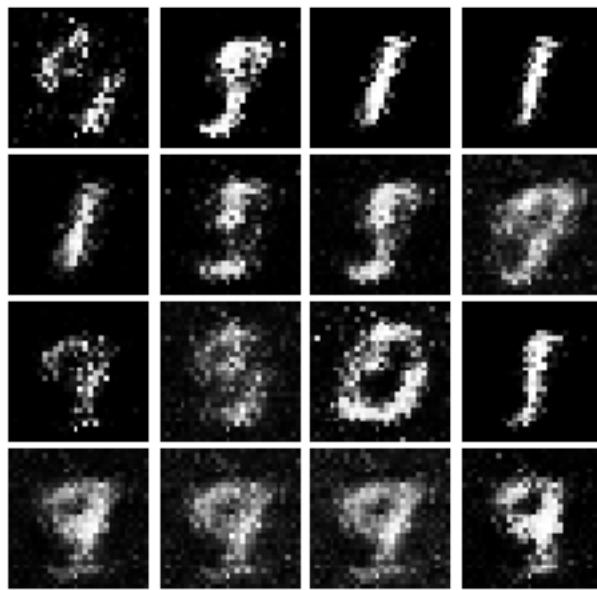
Iter: 250, D: 1.023, G:1.266



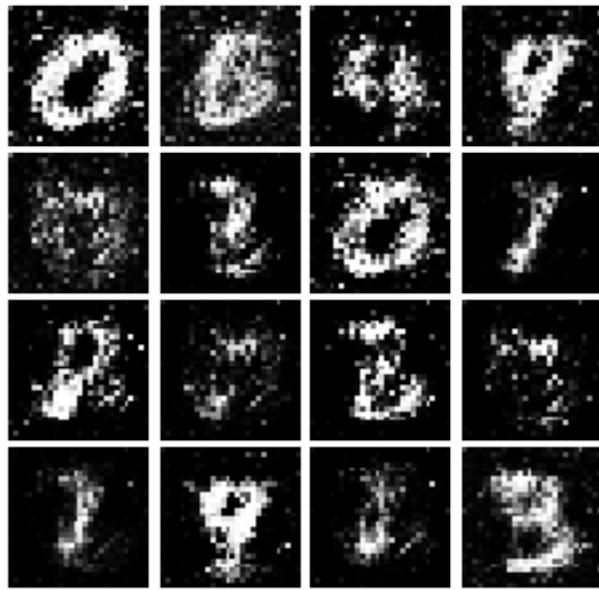
Iter: 500, D: 0.8688, G:1.291



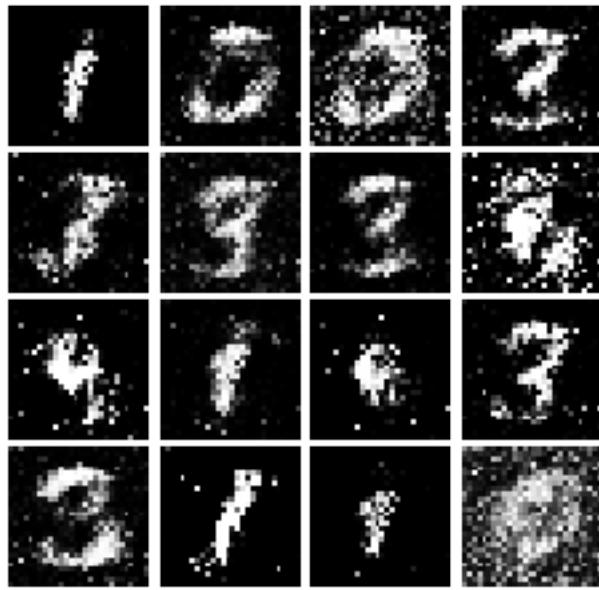
Iter: 750, D: 1.522, G:0.7612



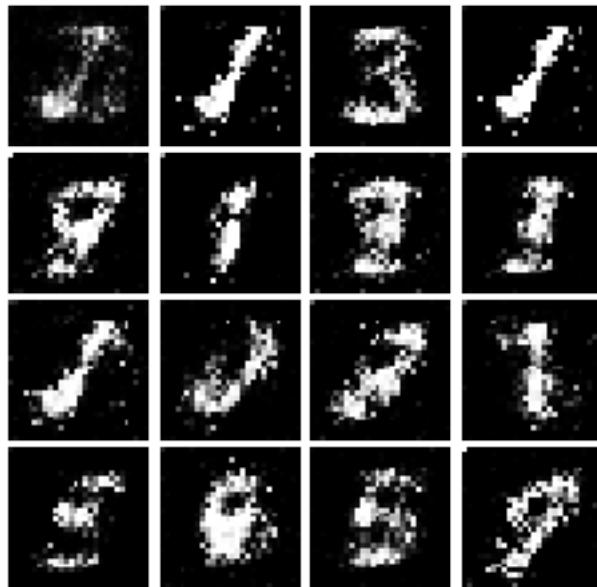
Iter: 1000, D: 1.091, G:1.414



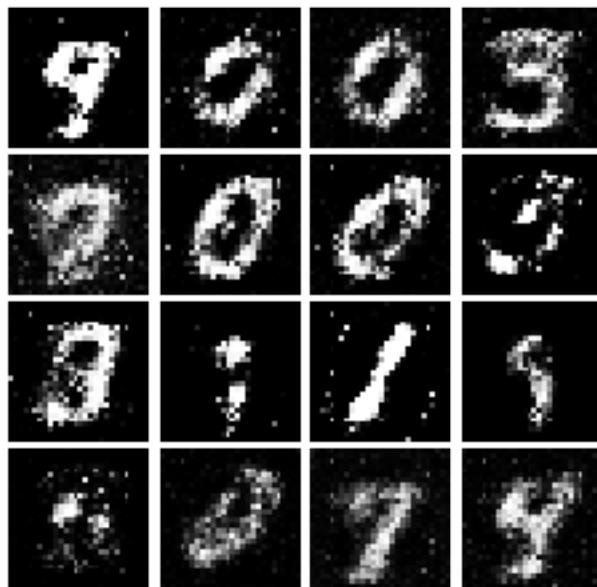
Iter: 1250, D: 1.347, G:0.5707



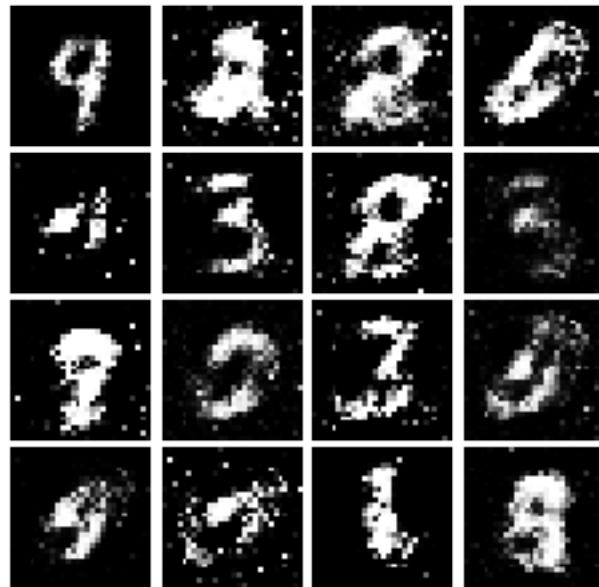
Iter: 1500, D: 1.269, G:0.8972



Iter: 1750, D: 1.369, G: 1.021



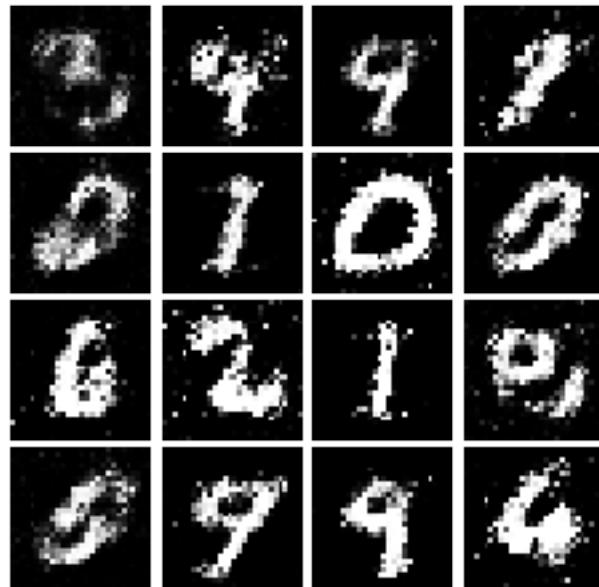
Iter: 2000, D: 1.246, G: 0.8893



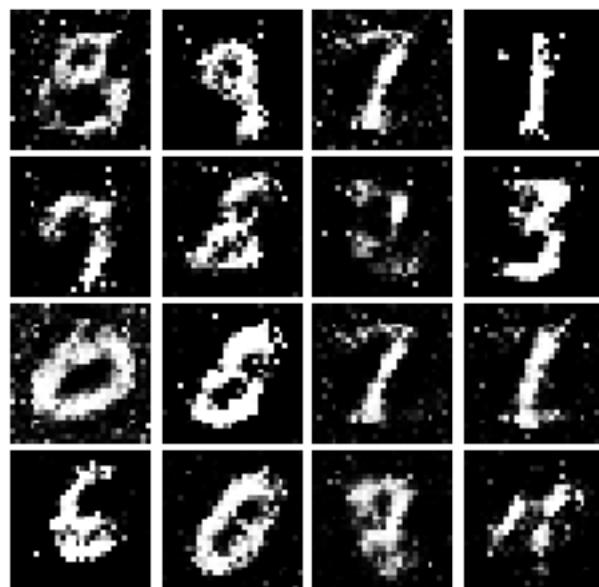
Iter: 2250, D: 1.307, G:0.9956



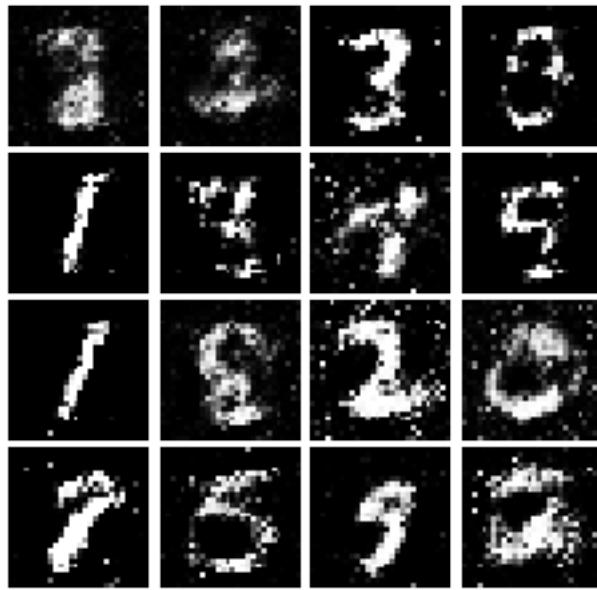
Iter: 2500, D: 1.399, G:0.9371



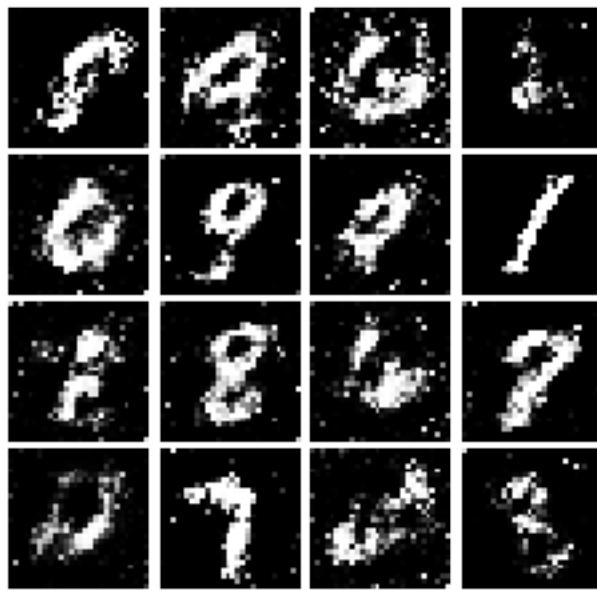
Iter: 2750, D: 1.295, G:0.9527



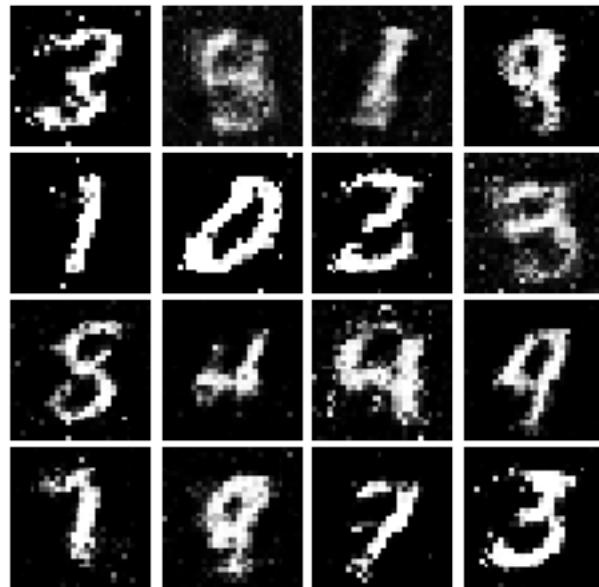
Iter: 3000, D: 1.384, G:0.8431



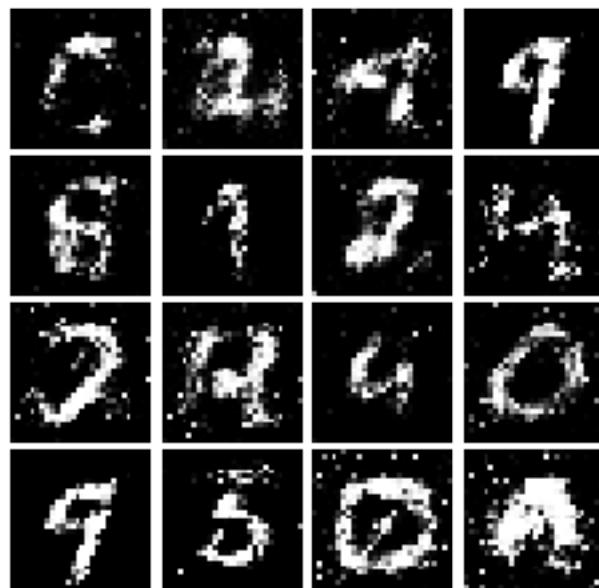
Iter: 3250, D: 1.244, G:0.8524



Iter: 3500, D: 1.294, G:0.7984



Iter: 3750, D: 1.376, G:0.8686



Well that wasn't so hard, was it? In the iterations in the low 100s you should see black backgrounds, fuzzy shapes as you approach iteration 1000, and decent shapes, about half of which will be sharp and clearly recognizable as we pass 3000.

## 7 Least Squares GAN

We'll now look at [Least Squares GAN](#), a newer, more stable alternative to the original GAN loss function. For this part, all we have to do is change the loss function and retrain the model. We'll implement equation (9) in the paper, with the generator loss:

$$\ell_G = \frac{1}{2} \mathbb{E}_{z \sim p(z)} [(D(G(z)) - 1)^2]$$

and the discriminator loss:

$$\ell_D = \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}} [(D(x) - 1)^2] + \frac{1}{2} \mathbb{E}_{z \sim p(z)} [(D(G(z)))^2]$$

**HINTS:** Instead of computing the expectation, we will be averaging over elements of the mini-batch, so make sure to combine the loss by averaging instead of summing. When plugging in for  $D(x)$  and  $D(G(z))$  use the direct output from the discriminator (`scores_real` and `scores_fake`).

```
[31]: def ls_discriminator_loss(scores_real, scores_fake):
    """
    Compute the Least-Squares GAN loss for the discriminator.

    Inputs:
    - scores_real: PyTorch Tensor of shape (N,) giving scores for the real data.
    - scores_fake: PyTorch Tensor of shape (N,) giving scores for the fake data.

    Outputs:
    - loss: A PyTorch Tensor containing the loss.
    """
    loss = (1/2)*torch.mean((scores_real-1)**2) + (1/2)*torch.
    →mean(scores_fake**2)
    return loss

def ls_generator_loss(scores_fake):
    """
    Computes the Least-Squares GAN loss for the generator.

    Inputs:
    - scores_fake: PyTorch Tensor of shape (N,) giving scores for the fake data.

    Outputs:
    - loss: A PyTorch Tensor containing the loss.
    """
    loss = (1/2)*torch.mean((scores_fake-1)**2)
    return loss
```

Before running a GAN with our new loss function, let's check it:

```
[32]: def test_lsgan_loss(score_real, score_fake, d_loss_true, g_loss_true):
    score_real = torch.Tensor(score_real).type(dtype)
    score_fake = torch.Tensor(score_fake).type(dtype)
    d_loss = ls_discriminator_loss(score_real, score_fake).cpu().numpy()
```

```

g_loss = ls_generator_loss(score_fake).cpu().numpy()
print("Maximum error in d_loss: %g"%rel_error(d_loss_true, d_loss))
print("Maximum error in g_loss: %g"%rel_error(g_loss_true, g_loss))

test_lsgan_loss(answers['logits_real'], answers['logits_fake'],
                 answers['d_loss_lsgan_true'], answers['g_loss_lsgan_true'])

```

Maximum error in d\_loss: 1.53171e-08  
 Maximum error in g\_loss: 3.36961e-08

Run the following cell to train your model!

```

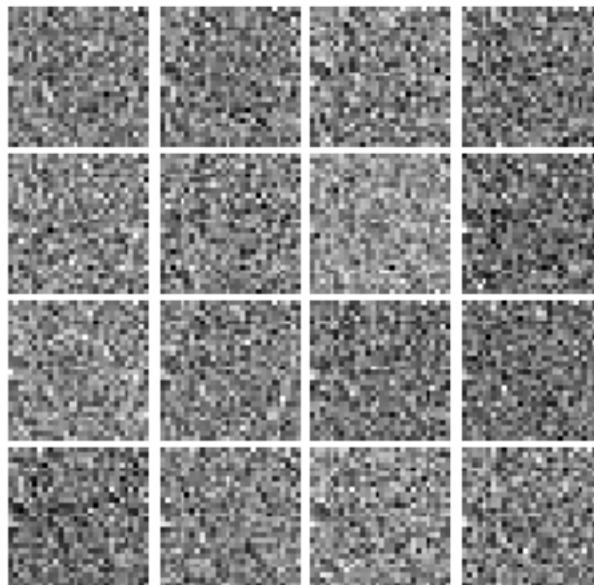
[33]: D_LS = discriminator().type(dtype)
G_LS = generator().type(dtype)

D_LS_solver = get_optimizer(D_LS)
G_LS_solver = get_optimizer(G_LS)

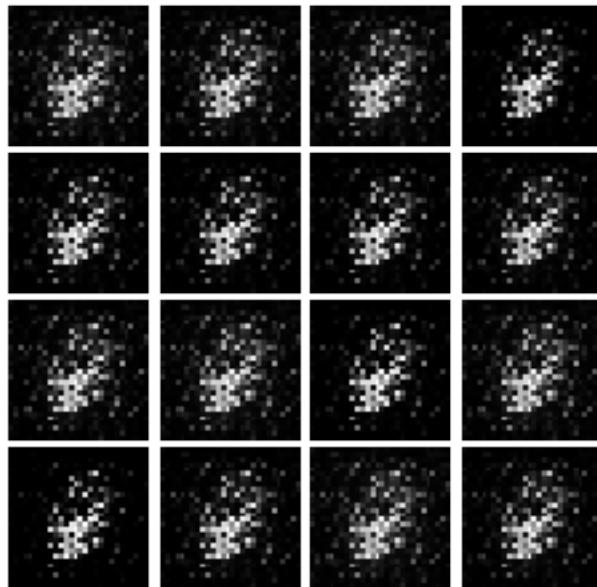
run_a_gan(D_LS, G_LS, D_LS_solver, G_LS_solver, ls_discriminator_loss,
          ↪ls_generator_loss)

```

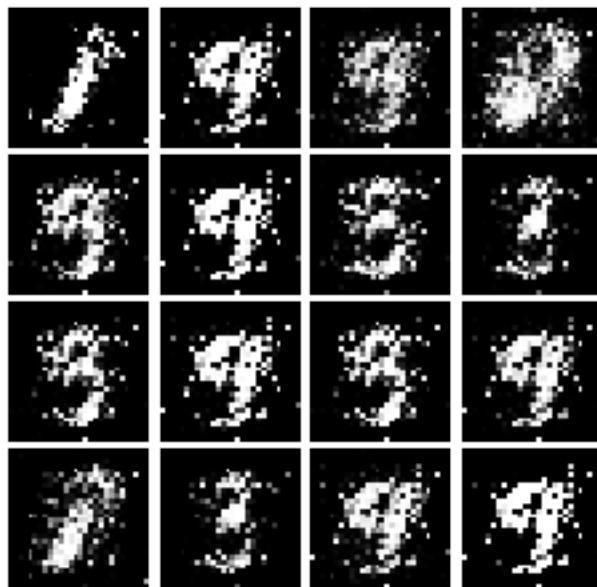
Iter: 0, D: 0.5925, G:0.5071



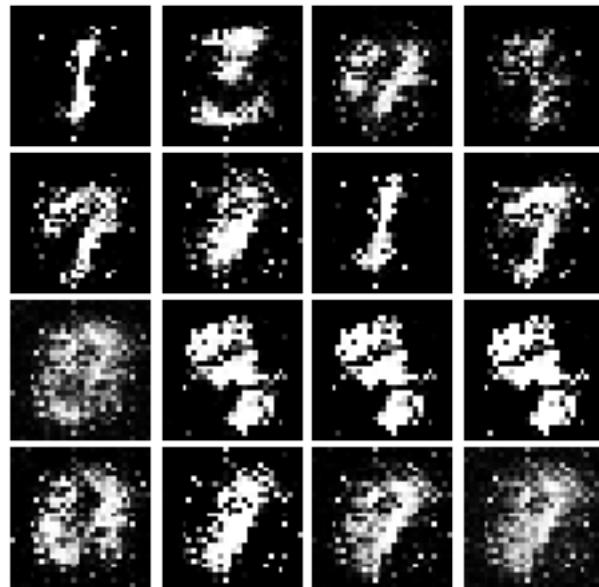
Iter: 250, D: 0.07226, G:0.5143



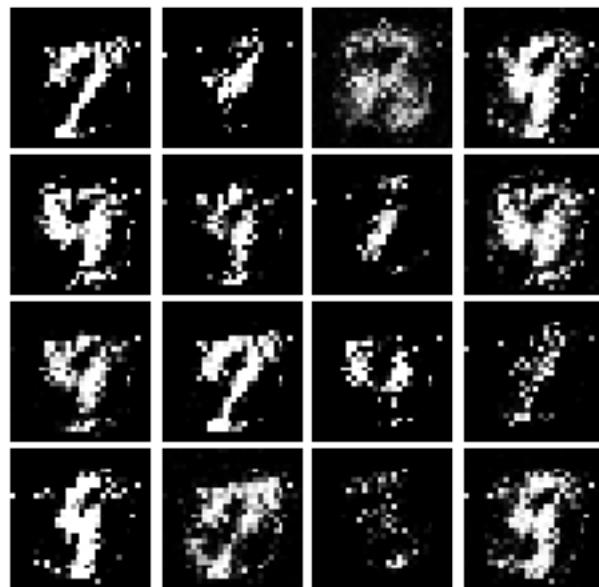
Iter: 500, D: 0.1363, G:0.328



Iter: 750, D: 0.1577, G:0.254



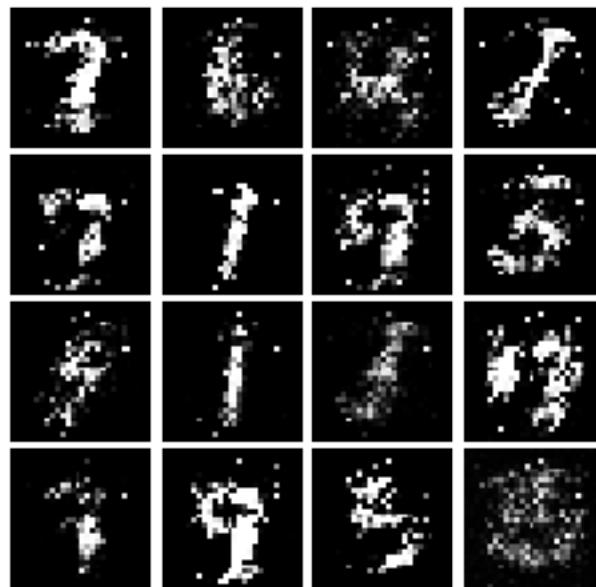
Iter: 1000, D: 0.1583, G:0.3071



Iter: 1250, D: 0.2053, G:0.3487



Iter: 1500, D: 0.1734, G: 0.2592



Iter: 1750, D: 0.1869, G: 0.2191



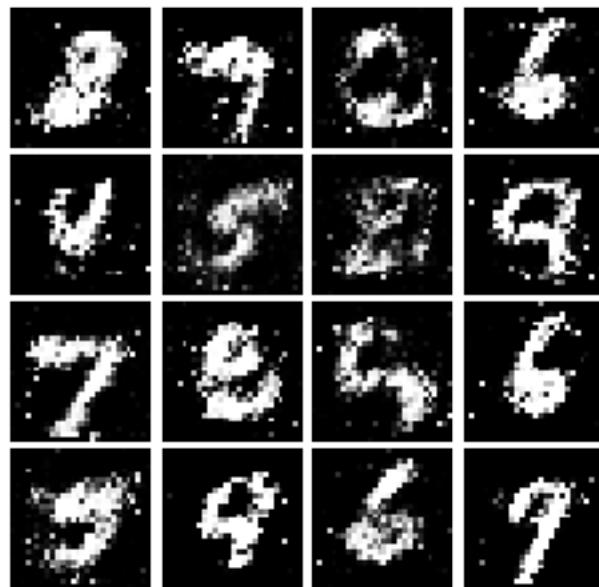
Iter: 2000, D: 0.2158, G: 0.1852



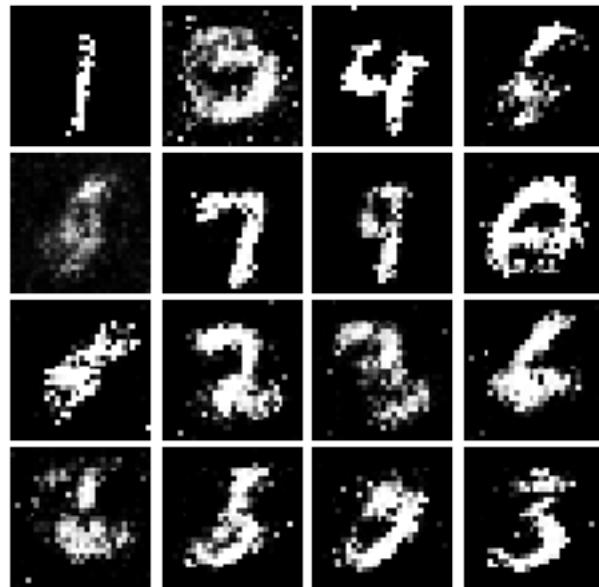
Iter: 2250, D: 0.2294, G: 0.1483



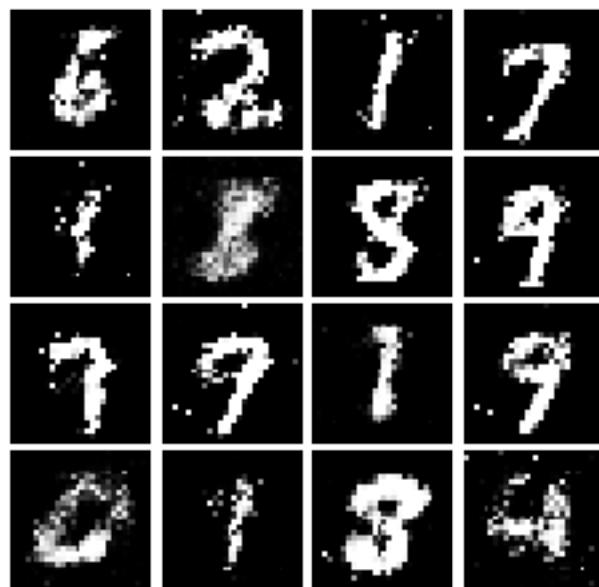
Iter: 2500, D: 0.2299, G: 0.2243



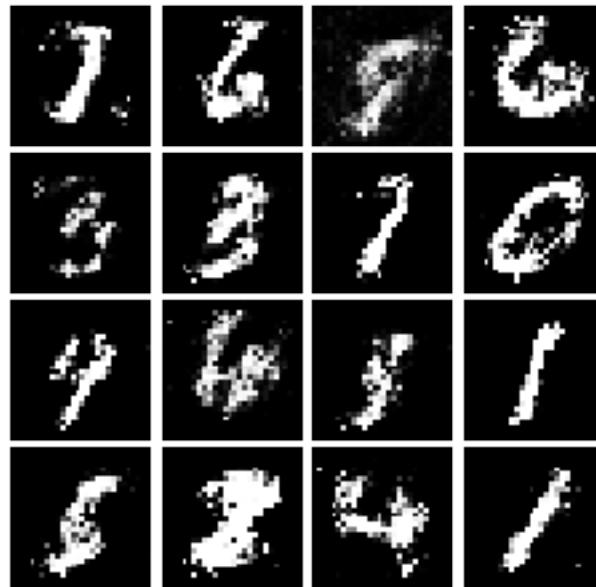
Iter: 2750, D: 0.2073, G: 0.1796



Iter: 3000, D: 0.236, G:0.151



Iter: 3250, D: 0.2285, G:0.1634



Iter: 3500, D: 0.2406, G: 0.1588



Iter: 3750, D: 0.2273, G: 0.1593



## 8 Deeply Convolutional GANs

In the first part of the notebook, we implemented an almost direct copy of the original GAN network from Ian Goodfellow. However, this network architecture allows no real spatial reasoning. It is unable to reason about things like "sharp edges" in general because it lacks any convolutional layers. Thus, in this section, we will implement some of the ideas from [DCGAN](#), where we use convolutional networks

**Discriminator** We will use a discriminator inspired by the TensorFlow MNIST classification tutorial, which is able to get above 99% accuracy on the MNIST dataset fairly quickly.

- \* Reshape into image tensor (Use Unflatten!)
- \* Conv2D: 32 Filters, 5x5, Stride 1
- \* Leaky ReLU(alpha=0.01)
- \* Max Pool 2x2, Stride 2
- \* Conv2D: 64 Filters, 5x5, Stride 1
- \* Leaky ReLU(alpha=0.01)
- \* Max Pool 2x2, Stride 2
- \* Flatten
- \* Fully Connected with output size 4 x 4 x 64
- \* Leaky ReLU(alpha=0.01)
- \* Fully Connected with output size 1

```
[46]: def build_dc_classifier():
    """
    Build and return a PyTorch model for the DCGAN discriminator implementing
    the architecture above.
    """
    return nn.Sequential(
        Unflatten(batch_size, 1, 28, 28),
        nn.Conv2d(1, 32, 5, 1),
        nn.LeakyReLU(0.01),
        nn.MaxPool2d(2, 2),
```

```

        nn.Conv2d(32, 64, 5, 1),
        nn.LeakyReLU(0.01),
        nn.MaxPool2d(2, 2),
        Flatten(),
        nn.Linear(1024, 1024),
        nn.LeakyReLU(0.01),
        nn.Linear(1024, 1)
    )

data = next(enumerate(loader_train))[-1][0].type(dtype)
b = build_dc_classifier().type(dtype)
out = b(data)
print(out.size())

```

`torch.Size([128, 1])`

Check the number of parameters in your classifier as a sanity check:

```
[47]: def test_dc_classifier(true_count=1102721):
    model = build_dc_classifier()
    cur_count = count_params(model)
    if cur_count != true_count:
        print('Incorrect number of parameters in generator. Check your'
              'architecture.')
    else:
        print('Correct number of parameters in generator.')

test_dc_classifier()
```

`Correct number of parameters in generator.`

**Generator** For the generator, we will copy the architecture exactly from the [InfoGAN paper](#). See Appendix C.1 MNIST. See the documentation for `tf.nn.conv2d_transpose`. We are always "training" in GAN mode. \* Fully connected with output size 1024 \* ReLU \* BatchNorm \* Fully connected with output size  $7 \times 7 \times 128$  \* ReLU \* BatchNorm \* Reshape into Image Tensor of shape  $7, 7, 128$  \* Conv2D<sup>T</sup> (Transpose): 64 filters of 4x4, stride 2, 'same' padding \* ReLU \* BatchNorm \* Conv2D<sup>T</sup> (Transpose): 1 filter of 4x4, stride 2, 'same' padding \* TanH \* Should have a 28x28x1 image, reshape back into 784 vector

```
[53]: def build_dc_generator(noise_dim=NOISE_DIM):
    """
    Build and return a PyTorch model implementing the DCGAN generator using
    the architecture described above.
    """
    return nn.Sequential(
        nn.Linear(noise_dim, 1024),
        nn.ReLU(),
        nn.BatchNorm1d(1024),
```

```

        nn.Linear(1024, 6272),
        nn.ReLU(),
        nn.BatchNorm1d(6272),
        Unflatten(batch_size, 128, 7, 7),
        nn.ConvTranspose2d(128, 64, 4, 2, 1),
        nn.ReLU(),
        nn.BatchNorm2d(64),
        nn.ConvTranspose2d(64, 1, 4, 2, 1),
        nn.Tanh(),
        Flatten()
    )

test_g_gan = build_dc_generator().type(dtype)
test_g_gan.apply(initialize_weights)

fake_seed = torch.randn(batch_size, NOISE_DIM).type(dtype)
fake_images = test_g_gan.forward(fake_seed)
fake_images.size()

```

[53]: torch.Size([128, 784])

Check the number of parameters in your generator as a sanity check:

[54]:

```

def test_dc_generator(true_count=6580801):
    model = build_dc_generator(4)
    cur_count = count_params(model)
    if cur_count != true_count:
        print('Incorrect number of parameters in generator. Check your'
              'architecture.')
    else:
        print('Correct number of parameters in generator.')

```

test\_dc\_generator()

Correct number of parameters in generator.

[55]:

```

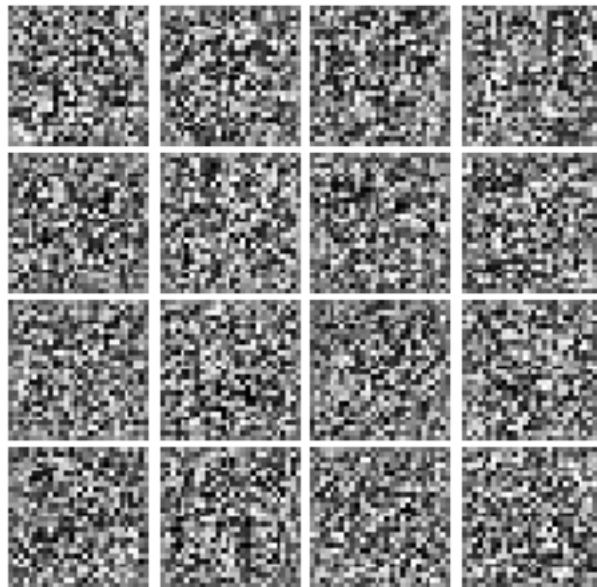
D_DC = build_dc_classifier().type(dtype)
D_DC.apply(initialize_weights)
G_DC = build_dc_generator().type(dtype)
G_DC.apply(initialize_weights)

D_DC_solver = get_optimizer(D_DC)
G_DC_solver = get_optimizer(G_DC)

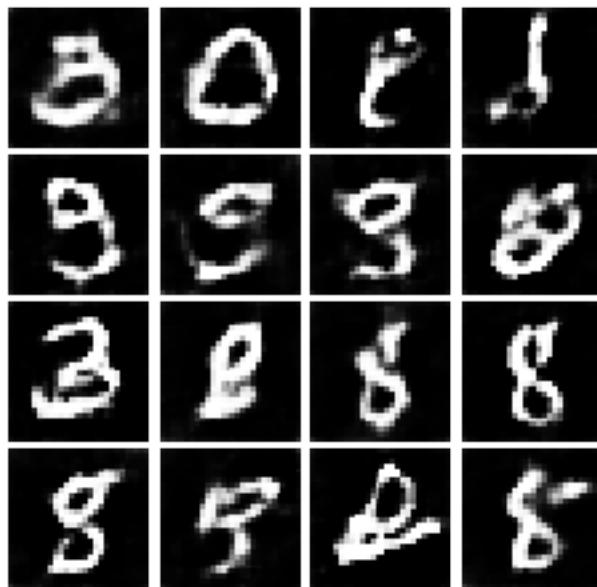
run_a_gan(D_DC, G_DC, D_DC_solver, G_DC_solver, discriminator_loss,
          generator_loss, num_epochs=5)

```

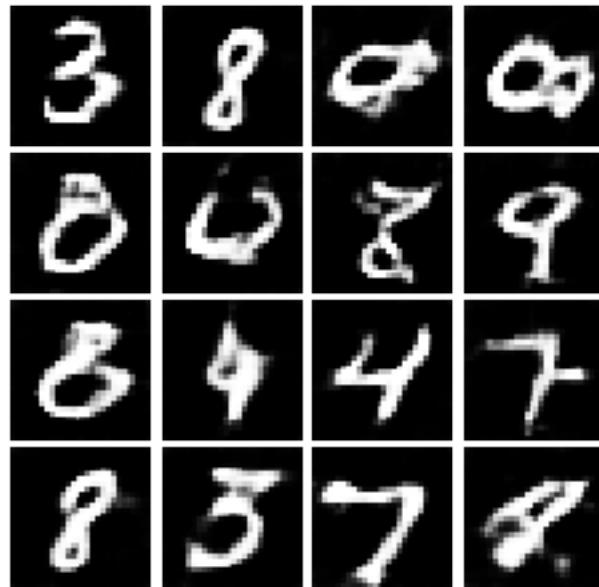
Iter: 0, D: 1.37, G:0.3721



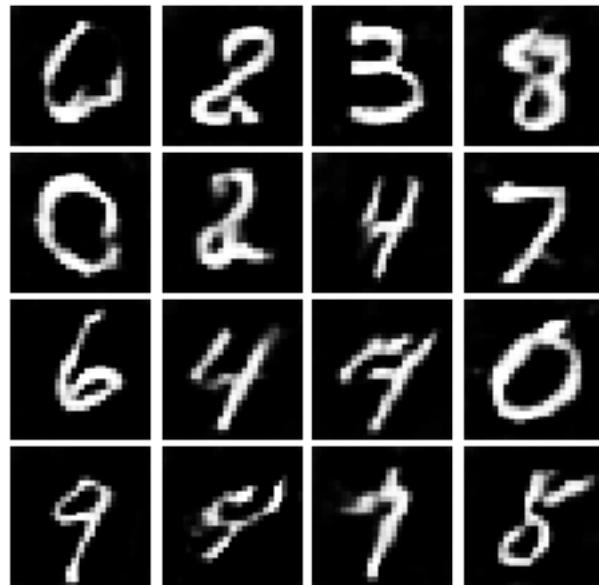
Iter: 250, D: 1.246, G:0.8127



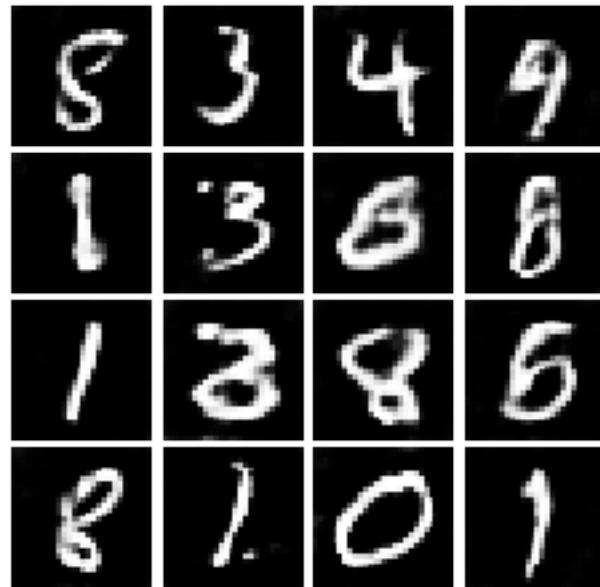
Iter: 500, D: 1.176, G:1.04



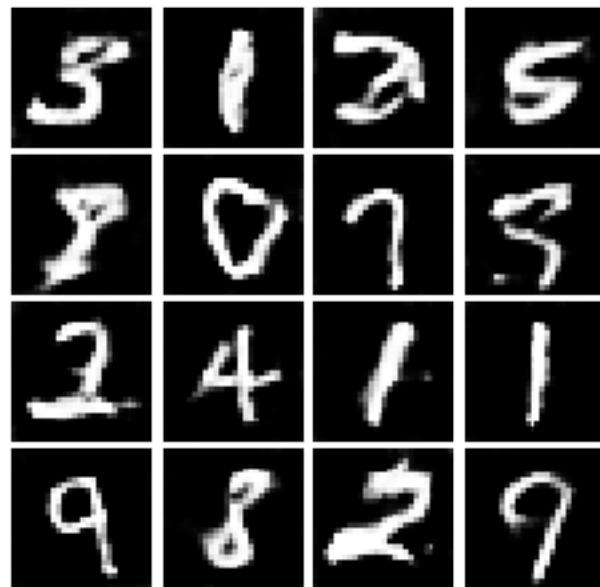
Iter: 750, D: 1.168, G:0.8971



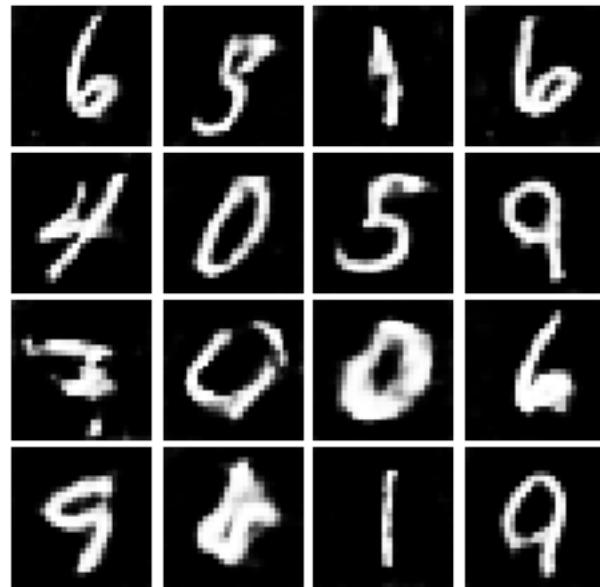
Iter: 1000, D: 1.249, G:1.087



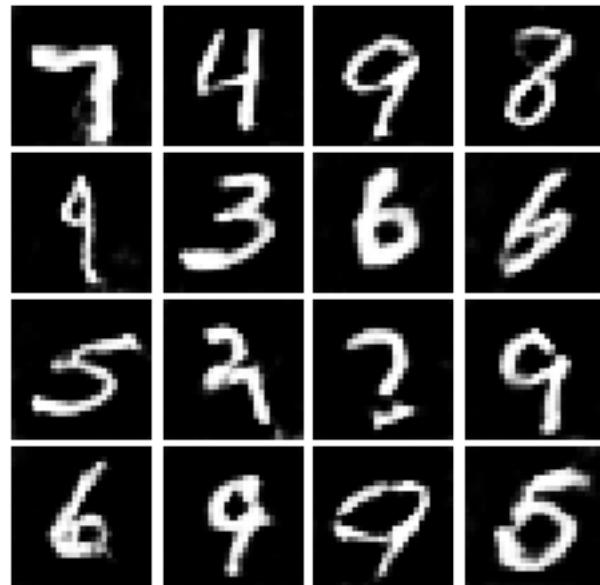
Iter: 1250, D: 1.3, G:0.9757



Iter: 1500, D: 1.169, G:0.9851



Iter: 1750, D: 1.123, G:0.8439



## 8.1 INLINE QUESTION 1

If the generator loss decreases during training while the discriminator loss stays at a constant high value from the start, is this a good sign? Why or why not? A qualitative answer is sufficient

Your answer: This is not a good sign. It is because the Generator is very easily able to generate fake images that the discriminator is easily fooled by (i.e. discriminates fake images as real images).