

RNN_Captioning

April 7, 2021

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.

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. 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:

```
In [1]: !pip install h5py
```

```
Requirement already satisfied: h5py in /anaconda/envs/msbd5601/lib/python3.7/site-packages (2.1.4)
Requirement already satisfied: six in /anaconda/envs/msbd5601/lib/python3.7/site-packages (from h5py)
Requirement already satisfied: numpy>=1.7 in /anaconda/envs/msbd5601/lib/python3.7/site-packages (from h5py)
```

```
In [2]: # As usual, a bit of setup
import time, os, json
import numpy as np
import matplotlib.pyplot as plt

from cs231n.gradient_check import eval_numerical_gradient, eval_numerical_gradient_array
from cs231n.rnn_layers import *
from cs231n.captioning_solver import CaptioningSolver
from cs231n.classifiers.rnn import CaptioningRNN
from cs231n.coco_utils import load_coco_data, sample_coco_minibatch, decode_captions
from cs231n.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))))

```

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 `cs231n/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 `cs231n/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 `cs231n/coco_utils.py`. Run the following cell to do so:

```

In [3]: # 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))

base dir  /home/msbd/msbd5601/assignment3/cs231n/datasets/coco_captioning
train_captions <class 'numpy.ndarray'> (400135, 17) int32
train_image_idxs <class 'numpy.ndarray'> (400135,) int32
val_captions <class 'numpy.ndarray'> (195954, 17) int32
val_image_idxs <class 'numpy.ndarray'> (195954,) int32
train_features <class 'numpy.ndarray'> (82783, 512) float32
val_features <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 `cs231n/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**.

```

In [4]: # 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 person sitting on a <UNK> with an umbrella <END>



<START> the person is taking a <UNK> from reading to <UNK> their cat <END>



<START> a bunch of young men riding bikes and skateboards <END>



3 Recurrent Neural Networks

As discussed in lecture, we will use recurrent neural network (RNN) language models for image captioning. The file `cs231n/rnn_layers.py` contains implementations of different layer types that are needed for recurrent neural networks, and the file `cs231n/classifiers/rnn.py` uses these layers to implement an image captioning model.

We will first implement different types of RNN layers in `cs231n/rnn_layers.py`.

4 Vanilla RNN: step forward

Open the file `cs231n/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.

In [5]: `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 `cs231n/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.

```

In [6]: from cs231n.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 `cs231n/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.

In [8]: `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 `cs231n/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.

```
In [15]: 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 `cs231n/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.

In [16]: `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: 1.0000000094736443e-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.

In [18]: `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(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 `cs231n/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.

```
In [19]: 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 `cs231n/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.

```
In [20]: # Sanity check for temporal softmax loss
        from cs231n.rnn_layers import temporal_softmax_loss

        N, T, V = 100, 1, 10

        def check_loss(N, T, V):
            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 within 2.2-2.4

        # 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, v

        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 `cs231n/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.

```
In [22]: 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.

```
In [23]: 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], verbose=False)
    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 `cs231n/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.

In [24]: `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,

```

```

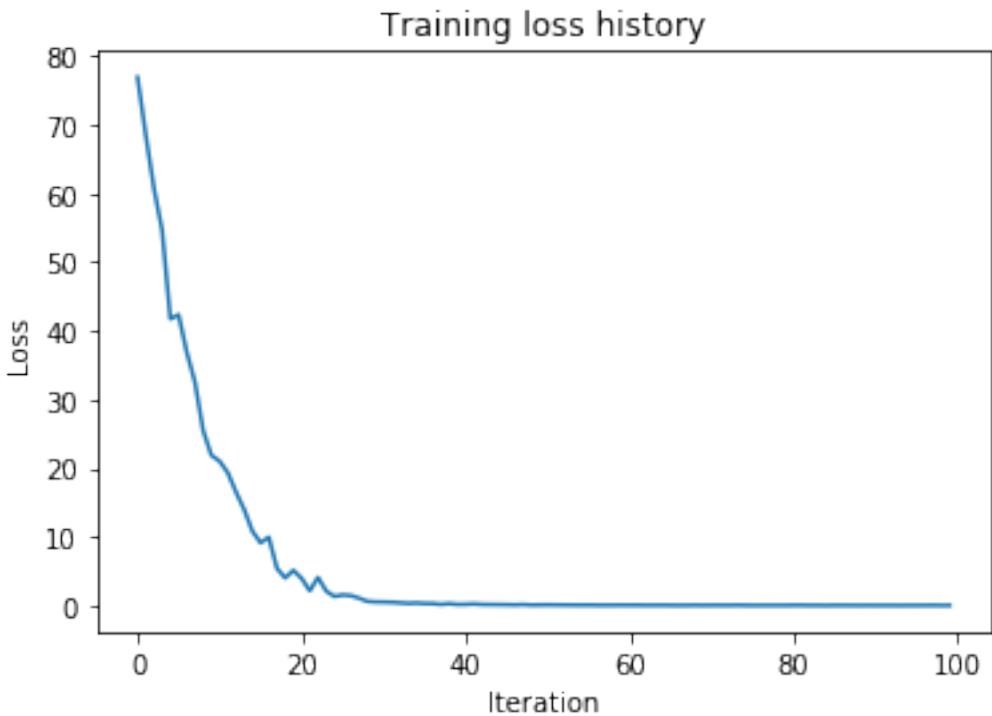
        )

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()

base dir  /home/msbd/msbd5601/assignment3/cs231n/datasets/coco_captioning
(Iteration 1 / 100) loss: 76.913486
(Iteration 11 / 100) loss: 21.063200
(Iteration 21 / 100) loss: 4.016248
(Iteration 31 / 100) loss: 0.567111
(Iteration 41 / 100) loss: 0.239442
(Iteration 51 / 100) loss: 0.162022
(Iteration 61 / 100) loss: 0.111543
(Iteration 71 / 100) loss: 0.097583
(Iteration 81 / 100) loss: 0.099097
(Iteration 91 / 100) loss: 0.073979

```



Print final training loss. You should see a final loss of less than 0.1.

```
In [25]: print('Final loss: ', small_rnn_solver.loss_history[-1])
```

```
Final loss:  0.08208879750046737
```

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 `cs231n/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.

```
In [28]: 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(gt_captions, sample_captions, urls):
    plt.imshow(image_from_url(url))
    plt.title('%s\n%s\nGT:%s' % (split, sample_caption, gt_caption))
    plt.axis('off')
    plt.show()
```

train

kids and a man walking down the sidewalk with suitcases <END>
GT:<START> kids and a man walking down the sidewalk with suitcases <END>



train

there is a male surfer coming out of the water <END>
GT:<START> there is a male surfer coming out of the water <END>



val

various <UNK> with <UNK> <END>
GT:<START> an elephant and baby elephant walk towards the water <END>



val
man truck woman on the <UNK> in a <UNK> <END>
GT:<START> a <UNK> <UNK> through a <UNK> <UNK> with benches <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.

Your Answer: * **Advantage:** Compared to word-level RNN, character-level RNN can save a lot of memory. There are only 26 letters in English, so character-level RNN only need to store these letters as vocabulary. While the vocabulary size of word-level RNN depends entirely on the dataset, if the dataset has one million different words, the vocabulary size would be up to one million. * **Disadvantage:** Character-level RNN is harder to learn the relationship between words and need larger parameter space. Seeing the example "A man is walking.", word-level RNN only extract 4 tokens, while character-level need 13 tokens, which means character-level RNN needs larger hidden layers. And it is also harder to remember the dependencies between words because of the more tokens.

In []:

LSTM_Captioning

April 7, 2021

1 Image Captioning with LSTMs

In the previous exercise you implemented a vanilla RNN and applied it to image captioning. In this notebook you will implement the LSTM update rule and use it for image captioning.

```
In [1]: # As usual, a bit of setup
    import time, os, json
    import numpy as np
    import matplotlib.pyplot as plt

    from cs231n.gradient_check import eval_numerical_gradient, eval_numerical_gradient_array
    from cs231n.rnn_layers import *
    from cs231n.captioning_solver import CaptioningSolver
    from cs231n.classifiers.rnn import CaptioningRNN
    from cs231n.coco_utils import load_coco_data, sample_coco_minibatch, decode_captions
    from cs231n.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))))



```

2 Load MS-COCO data

As in the previous notebook, we will use the Microsoft COCO dataset for captioning.

```
In [2]: # 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))

base dir /home/msbd/msbd5601/assignment3/cs231n/datasets/coco_captioning
train_captions <class 'numpy.ndarray'> (400135, 17) int32
train_image_idxs <class 'numpy.ndarray'> (400135,) int32
val_captions <class 'numpy.ndarray'> (195954, 17) int32
val_image_idxs <class 'numpy.ndarray'> (195954,) int32
train_features <class 'numpy.ndarray'> (82783, 512) float32
val_features <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

```

3 LSTM

If you read recent papers, you'll see that many people use a variant on the vanilla RNN called Long-Short Term Memory (LSTM) RNNs. Vanilla RNNs can be tough to train on long sequences due to vanishing and exploding gradients caused by repeated matrix multiplication. LSTMs solve this problem by replacing the simple update rule of the vanilla RNN with a gating mechanism as follows.

Similar to the vanilla RNN, at each timestep we receive an input $x_t \in \mathbb{R}^D$ and the previous hidden state $h_{t-1} \in \mathbb{R}^H$; the LSTM also maintains an H -dimensional *cell state*, so we also receive the previous cell state $c_{t-1} \in \mathbb{R}^H$. The learnable parameters of the LSTM are an *input-to-hidden* matrix $W_x \in \mathbb{R}^{4H \times D}$, a *hidden-to-hidden* matrix $W_h \in \mathbb{R}^{4H \times H}$ and a *bias vector* $b \in \mathbb{R}^{4H}$.

At each timestep we first compute an *activation vector* $a \in \mathbb{R}^{4H}$ as $a = W_x x_t + W_h h_{t-1} + b$. We then divide this into four vectors $a_i, a_f, a_o, a_g \in \mathbb{R}^H$ where a_i consists of the first H elements of a , a_f is the next H elements of a , etc. We then compute the *input gate* $g \in \mathbb{R}^H$, *forget gate* $f \in \mathbb{R}^H$, *output gate* $o \in \mathbb{R}^H$ and *block input* $g \in \mathbb{R}^H$ as

$$i = \sigma(a_i) \quad f = \sigma(a_f) \quad o = \sigma(a_o) \quad g = \tanh(a_g)$$

where σ is the sigmoid function and \tanh is the hyperbolic tangent, both applied elementwise. Finally we compute the next cell state c_t and next hidden state h_t as

$$c_t = f \odot c_{t-1} + i \odot g \quad h_t = o \odot \tanh(c_t)$$

where \odot is the elementwise product of vectors.

In the rest of the notebook we will implement the LSTM update rule and apply it to the image captioning task.

In the code, we assume that data is stored in batches so that $X_t \in \mathbb{R}^{N \times D}$, and will work with *transposed* versions of the parameters: $W_x \in \mathbb{R}^{D \times 4H}$, $W_h \in \mathbb{R}^{H \times 4H}$ so that activations $A \in \mathbb{R}^{N \times 4H}$ can be computed efficiently as $A = X_t W_x + H_{t-1} W_h$

4 LSTM: step forward

Implement the forward pass for a single timestep of an LSTM in the `lstm_step_forward` function in the file `cs231n/rnn_layers.py`. This should be similar to the `rnn_step_forward` function that you implemented above, but using the LSTM update rule instead.

Once you are done, run the following to perform a simple test of your implementation. You should see errors on the order of $e-8$ or less.

```
In [3]: N, D, H = 3, 4, 5
x = np.linspace(-0.4, 1.2, num=N*D).reshape(N, D)
prev_h = np.linspace(-0.3, 0.7, num=N*H).reshape(N, H)
prev_c = np.linspace(-0.4, 0.9, num=N*H).reshape(N, H)
Wx = np.linspace(-2.1, 1.3, num=4*D*H).reshape(D, 4 * H)
Wh = np.linspace(-0.7, 2.2, num=4*H*H).reshape(H, 4 * H)
b = np.linspace(0.3, 0.7, num=4*H)

next_h, next_c, cache = lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)

expected_next_h = np.asarray([
    [ 0.24635157,  0.28610883,  0.32240467,  0.35525807,  0.38474904],
    [ 0.49223563,  0.55611431,  0.61507696,  0.66844003,  0.7159181 ],
    [ 0.56735664,  0.66310127,  0.74419266,  0.80889665,  0.858299  ]])
expected_next_c = np.asarray([
    [ 0.32986176,  0.39145139,  0.451556,    0.51014116,  0.56717407],
    [ 0.66382255,  0.76674007,  0.87195994,  0.97902709,  1.08751345],
    [ 0.74192008,  0.90592151,  1.07717006,  1.25120233,  1.42395676]])

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

next_h error:  5.7054131967097955e-09
next_c error:  5.8143123088804145e-09
```

5 LSTM: step backward

Implement the backward pass for a single LSTM timestep in the function `lstm_step_backward` in the file `cs231n/rnn_layers.py`. Once you are done, run the following to perform numeric gradient checking on your implementation. You should see errors on the order of $e-7$ or less.

```
In [4]: np.random.seed(231)
```

```
N, D, H = 4, 5, 6
```

```

x = np.random.randn(N, D)
prev_h = np.random.randn(N, H)
prev_c = np.random.randn(N, H)
Wx = np.random.randn(D, 4 * H)
Wh = np.random.randn(H, 4 * H)
b = np.random.randn(4 * H)

next_h, next_c, cache = lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)

dnext_h = np.random.randn(*next_h.shape)
dnext_c = np.random.randn(*next_c.shape)

fx_h = lambda x: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[0]
fh_h = lambda h: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[0]
fc_h = lambda c: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[0]
fWx_h = lambda Wx: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[0]
fWh_h = lambda Wh: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[0]
fb_h = lambda b: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[0]

fx_c = lambda x: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[1]
fh_c = lambda h: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[1]
fc_c = lambda c: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[1]
fWx_c = lambda Wx: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[1]
fWh_c = lambda Wh: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[1]
fb_c = lambda b: lstm_step_forward(x, prev_h, prev_c, Wx, Wh, b)[1]

num_grad = eval_numerical_gradient_array

dx_num = num_grad(fx_h, x, dnext_h) + num_grad(fx_c, x, dnext_c)
dh_num = num_grad(fh_h, prev_h, dnext_h) + num_grad(fh_c, prev_h, dnext_c)
dc_num = num_grad(fc_h, prev_c, dnext_h) + num_grad(fc_c, prev_c, dnext_c)
dWx_num = num_grad(fWx_h, Wx, dnext_h) + num_grad(fWx_c, Wx, dnext_c)
dWh_num = num_grad(fWh_h, Wh, dnext_h) + num_grad(fWh_c, Wh, dnext_c)
db_num = num_grad(fb_h, b, dnext_h) + num_grad(fb_c, b, dnext_c)

dx, dh, dc, dWx, dWh, db = lstm_step_backward(dnext_h, dnext_c, cache)

print('dx error: ', rel_error(dx_num, dx))
print('dh error: ', rel_error(dh_num, dh))
print('dc error: ', rel_error(dc_num, dc))
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: 6.335032254429549e-10
dh error: 3.3963774090592634e-10
dc error: 1.5221723979041107e-10
dWx error: 2.1010960934639614e-09

```

```
dWh error:  9.712296180612259e-08
db error:  2.4915214652298706e-10
```

6 LSTM: forward

In the function `lstm_forward` in the file `cs231n/rnn_layers.py`, implement the `lstm_forward` function to run an LSTM forward on an entire timeseries of data.

When you are done, run the following to check your implementation. You should see an error on the order of $e-7$ or less.

```
In [5]: N, D, H, T = 2, 5, 4, 3
        x = np.linspace(-0.4, 0.6, num=N*T*D).reshape(N, T, D)
        h0 = np.linspace(-0.4, 0.8, num=N*H).reshape(N, H)
        Wx = np.linspace(-0.2, 0.9, num=4*D*H).reshape(D, 4 * H)
        Wh = np.linspace(-0.3, 0.6, num=4*H*H).reshape(H, 4 * H)
        b = np.linspace(0.2, 0.7, num=4*H)

        h, cache = lstm_forward(x, h0, Wx, Wh, b)

        expected_h = np.asarray([
            [[ 0.01764008,  0.01823233,  0.01882671,  0.0194232 ],
             [ 0.11287491,  0.12146228,  0.13018446,  0.13902939],
             [ 0.31358768,  0.33338627,  0.35304453,  0.37250975]],
            [[ 0.45767879,  0.4761092,   0.4936887,   0.51041945],
             [ 0.6704845,   0.69350089,  0.71486014,  0.7346449 ],
             [ 0.81733511,  0.83677871,  0.85403753,  0.86935314]]])

        print('h error: ', rel_error(expected_h, h))

h error:  8.610537452106624e-08
```

7 LSTM: backward

Implement the backward pass for an LSTM over an entire timeseries of data in the function `lstm_backward` in the file `cs231n/rnn_layers.py`. When you are done, run the following to perform numeric gradient checking on your implementation. You should see errors on the order of $e-8$ or less. (For dWh , it's fine if your error is on the order of $e-6$ or less).

```
In [6]: from cs231n.rnn_layers import lstm_forward, lstm_backward
        np.random.seed(231)

        N, D, T, H = 2, 3, 10, 6

        x = np.random.randn(N, T, D)
        h0 = np.random.randn(N, H)
```

```

Wx = np.random.randn(D, 4 * H)
Wh = np.random.randn(H, 4 * H)
b = np.random.randn(4 * H)

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

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

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

fx = lambda x: lstm_forward(x, h0, Wx, Wh, b)[0]
fh0 = lambda h0: lstm_forward(x, h0, Wx, Wh, b)[0]
fWx = lambda Wx: lstm_forward(x, h0, Wx, Wh, b)[0]
fWh = lambda Wh: lstm_forward(x, h0, Wx, Wh, b)[0]
fb = lambda b: lstm_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: 6.9939005453315376e-09
dh0 error: 1.5042752061355818e-09
dWx error: 3.2262956411424662e-09
dWh error: 2.6984653115906307e-06
db error: 8.236645826637549e-10

```

8 INLINE QUESTION

Recall that in an LSTM the input gate i , forget gate f , and output gate o are all outputs of a sigmoid function. Why don't we use the ReLU activation function instead of sigmoid to compute these values? Explain.

Your Answer: The LSTM uses the sigmoid function to simulate the role of a gate, i.e., 1 when needed and 0 when not needed, which requires that the range of the activation function must be limited to $[0, 1]$, and sigmoid satisfies this condition. When using Relu, the output will be $[0, +\infty]$, which will easily cause the problem of gradient exploding, and the feature of Relu to mitigate the gradient vanishing is not necessary in LSTM.

9 LSTM captioning model

Now that you have implemented an LSTM, update the implementation of the `loss` method of the `CaptioningRNN` class in the file `cs231n/classifiers/rnn.py` to handle the case where `self.cell_type` is `lstm`. This should require adding less than 10 lines of code.

Once you have done so, run the following to check your implementation. You should see a difference on the order of $e-10$ or less.

```
In [7]: 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='lstm',
                      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(-0.5, 1.7, num=N*D).reshape(N, D)
captions = (np.arange(N * T) % V).reshape(N, T)

loss, grads = model.loss(features, captions)
expected_loss = 9.82445935443

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

loss: 9.824459354432264
expected loss: 9.82445935443
difference: 2.2648549702353193e-12
```

10 Overfit LSTM captioning model

Run the following to overfit an LSTM captioning model on the same small dataset as we used for the RNN previously. You should see a final loss less than 0.5.

```
In [8]: np.random.seed(231)

small_data = load_coco_data(max_train=50)
```

```

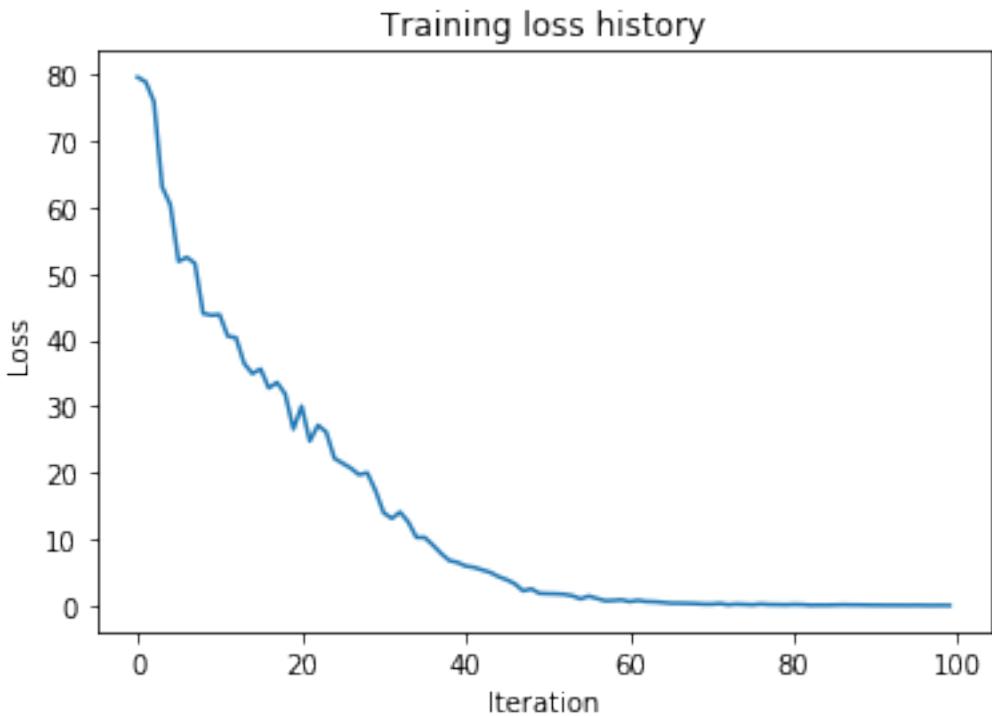
small_lstm_model = CaptioningRNN(
    cell_type='lstm',
    word_to_idx=data['word_to_idx'],
    input_dim=data['train_features'].shape[1],
    hidden_dim=512,
    wordvec_dim=256,
    dtype=np.float32,
)

small_lstm_solver = CaptioningSolver(small_lstm_model, small_data,
    update_rule='adam',
    num_epochs=50,
    batch_size=25,
    optim_config={
        'learning_rate': 5e-3,
    },
    lr_decay=0.995,
    verbose=True, print_every=10,
)
small_lstm_solver.train()

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

base dir /home/msbd/msbd5601/assignment3/cs231n/datasets/coco_captioning
(Iteration 1 / 100) loss: 79.551150
(Iteration 11 / 100) loss: 43.829102
(Iteration 21 / 100) loss: 30.062547
(Iteration 31 / 100) loss: 14.020211
(Iteration 41 / 100) loss: 6.007101
(Iteration 51 / 100) loss: 1.857024
(Iteration 61 / 100) loss: 0.662612
(Iteration 71 / 100) loss: 0.284369
(Iteration 81 / 100) loss: 0.243648
(Iteration 91 / 100) loss: 0.130096

```



Print final training loss. You should see a final loss of less than 0.5.

```
In [9]: print('Final loss: ', small_lstm_solver.loss_history[-1])
```

```
Final loss:  0.08172008973785243
```

11 LSTM test-time sampling

Modify the `sample` method of the `CaptioningRNN` class to handle the case where `self.cell_type` is `lstm`. This should take fewer than 10 lines of code.

When you are done run the following to sample from your overfit LSTM model on some training and validation set samples. As with the RNN, training results should be very good, and validation results probably won't make a lot of sense (because we're overfitting).

```
In [10]: 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_lstm_model.sample(features)
    sample_captions = decode_captions(sample_captions, data['idx_to_word'])

    for gt_caption, sample_caption, url in zip(gt_captions, sample_captions, urls):
```

```
plt.imshow(image_from_url(url))
plt.title('%s\n%s\nGT:%s' % (split, sample_caption, gt_caption))
plt.axis('off')
plt.show()
```

train

a man standing on the side of a road with bags of luggage <END>
GT:<START> a man standing on the side of a road with bags of luggage <END>



train

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



val

a person that is <UNK> of a <UNK> <END>

GT:<START> a sign that is on the front of a train station <END>



val

a cat sitting with a <UNK> <END>

GT:<START> a car is parked on a street at night <END>



In []:

Network Visualization-PyTorch

April 7, 2021

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.

```
In [1]: import torch
        import torchvision
        import numpy as np
        import random
        import matplotlib.pyplot as plt
        from PIL import Image
        from cs231n.image_utils import SQUEEZENET_MEAN, SQUEEZENET_STD

        %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

```

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 in `cs231n/net_visualization_pytorch`. You don't need to do anything here.

In [2]: `from cs231n.net_visualization_pytorch import preprocess, deprocess, rescale, blur_image`

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

In [3]: `# 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 continu`

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 `cs231n/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.

```
In [4]: from cs231n.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 a 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](#).

```
In [5]: # 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([[ 0.4636, -0.9071, -0.0325,  1.2165, -0.1103],
        [-0.0827,  0.9708,  0.5916,  0.8617,  0.8121],
        [-0.3395, -0.5872,  0.8550, -0.8999,  1.5639],
        [ 0.2683, -0.0613,  0.7454, -1.5803, -0.8741]])
tensor([1, 2, 1, 3])
tensor([-0.9071,  0.5916, -0.5872, -1.5803])
```

Implement `compute_saliency_maps` function inside `cs231n/net_visualization_pytorch.py`

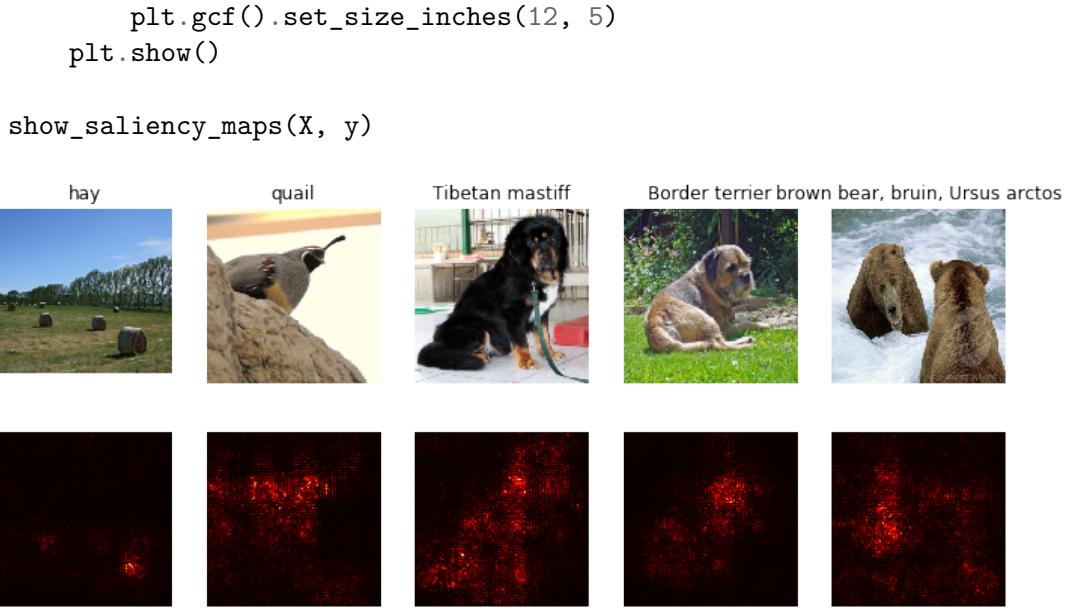
```
In [6]: # Load saliency maps computation function
from cs231n.net_visualization_pytorch import compute_saliency_maps
```

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

```
In [7]: 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')
```



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?

Your Answer: False * Size of saliency is (H, W) but size of input is $(3, H, W)$. Because we use the maximum of the absolute value of 3 channels, we lost information about other 2 channels, so we can not use saliency map to update the parameter. * We lost direction of the existed channel, so saliency map will always be positive, the resulting pixels would approximate to very large numbers (towards $+\infty$)

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

Implement `make_fooling_image` function inside `cs231n/net_visualization_pytorch.py`

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.

```
In [8]: from cs231n.net_visualization_pytorch import make_fooling_image
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)[1][0].item(), 'The model is not fooled!'

Iteration: 0
Needed iterations: 10
```

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

```
In [9]: 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.

[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

In `cs231n/net_visualization_pytorch.py` complete the implementation of the `image_visualization_update_step` used in the `create_class_visualization` function below. Once you have completed that implementation, run the following cells to generate an image of a Tarantula:

```
In [10]: from cs231n.net_visualization_pytorch import class_visualization_update_step, jitter,
def create_class_visualization(target_y, model, dtype, **kwargs):
    """
    Generate an image to maximize the score of target_y under a pretrained 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 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))
    class_visualization_update_step(img, model, target_y, l2_reg, learning_rate)
    # 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)

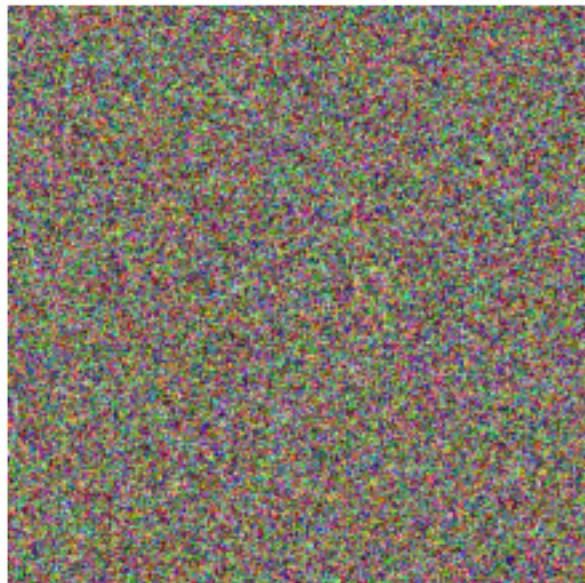
    # Periodically show the image
    if t == 0 or (t + 1) % show_every == 0 or t == num_iterations - 1:
        plt.imshow(deprocess(img.data.clone().cpu()))
        class_name = class_names[target_y]
        plt.title('%s\nIteration %d / %d' % (class_name, t + 1, num_iterations))
        plt.gcf().set_size_inches(4, 4)
        plt.axis('off')
        plt.show()
```

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

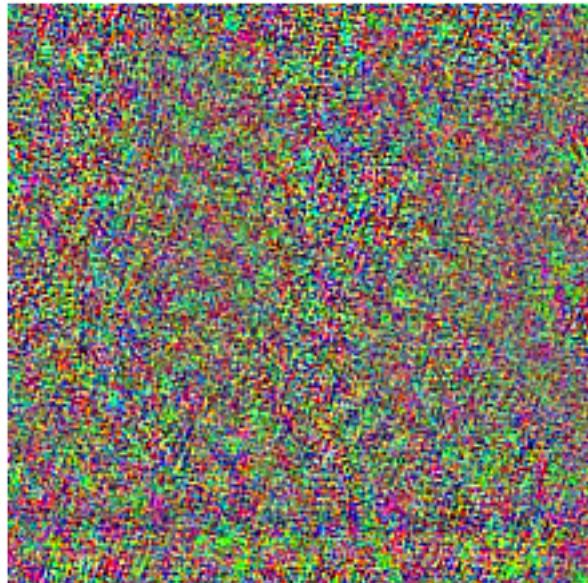
In [11]: dtype = torch.FloatTensor
        # dtype = torch.cuda.FloatTensor # Uncomment this to use GPU
        model.type(dtype)

        #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)
```

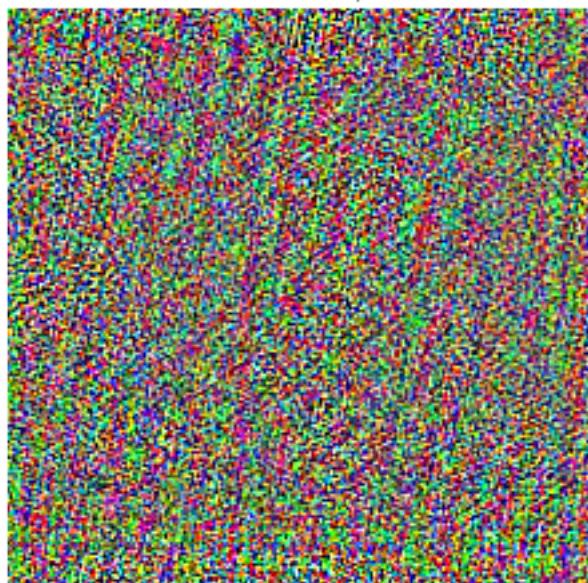
gorilla, Gorilla gorilla
Iteration 1 / 100



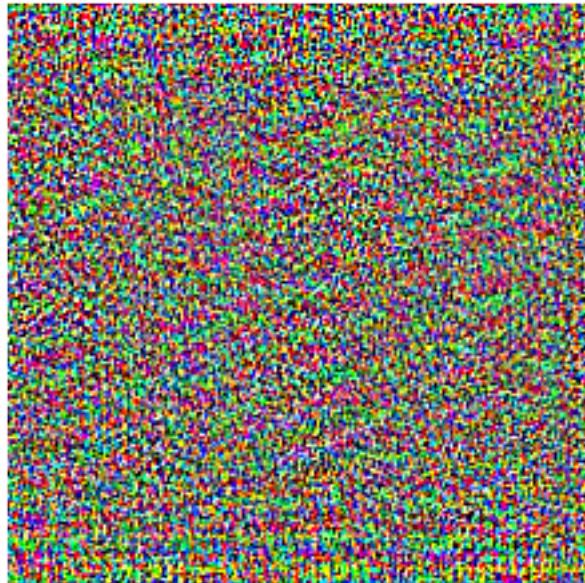
gorilla, Gorilla gorilla
Iteration 25 / 100



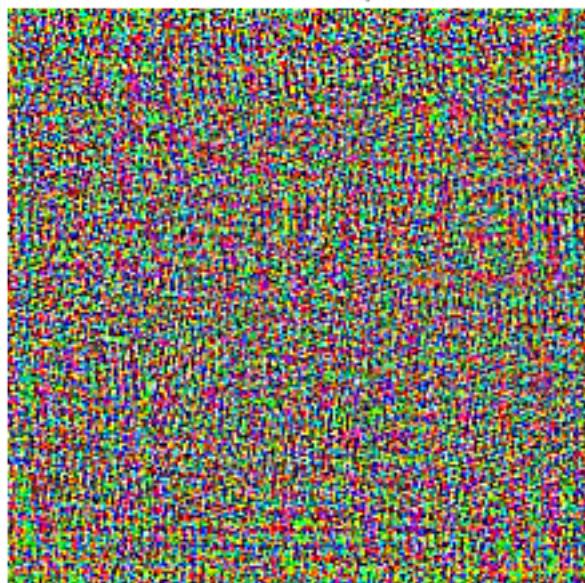
gorilla, Gorilla gorilla
Iteration 50 / 100



gorilla, Gorilla gorilla
Iteration 75 / 100



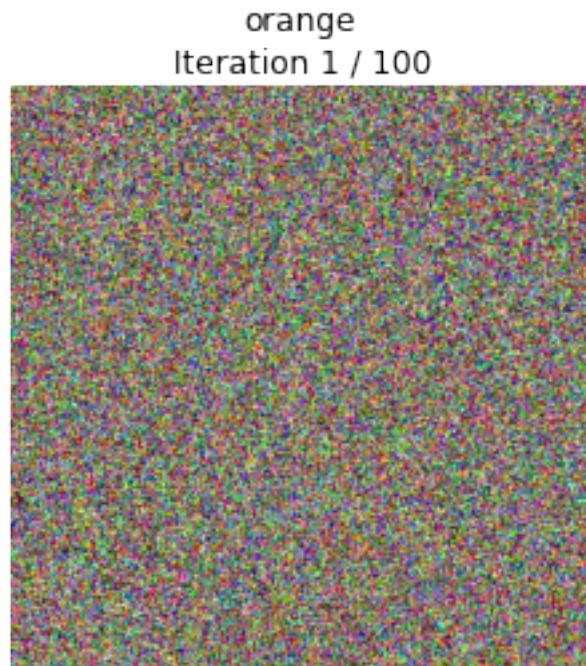
gorilla, Gorilla gorilla
Iteration 100 / 100



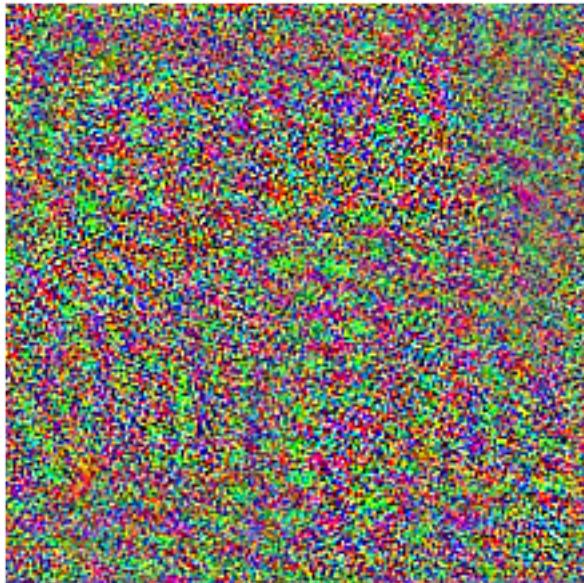
Try out your class visualization on other classes! You should also feel free to play with various hyperparameters to try and improve the quality of the generated image, but this is not required.

```
In [12]: # 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)
```

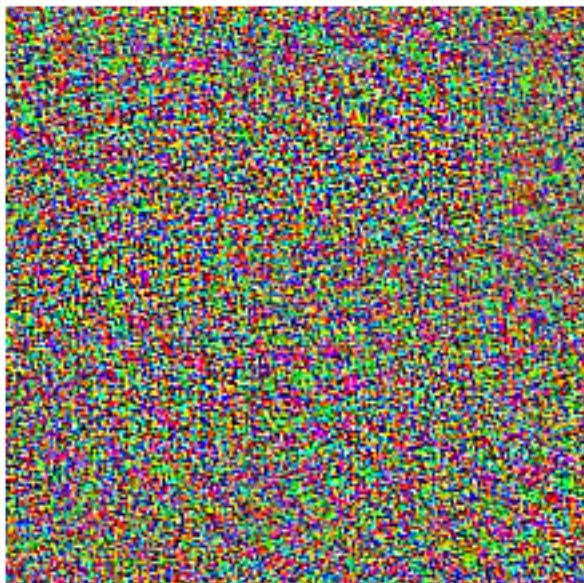
orange



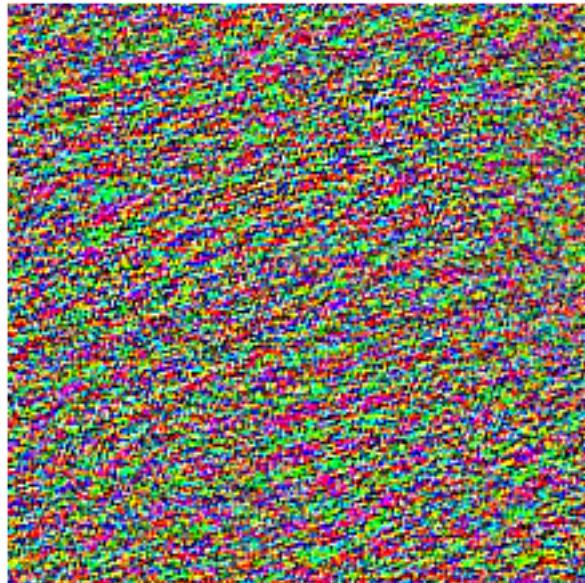
orange
Iteration 25 / 100



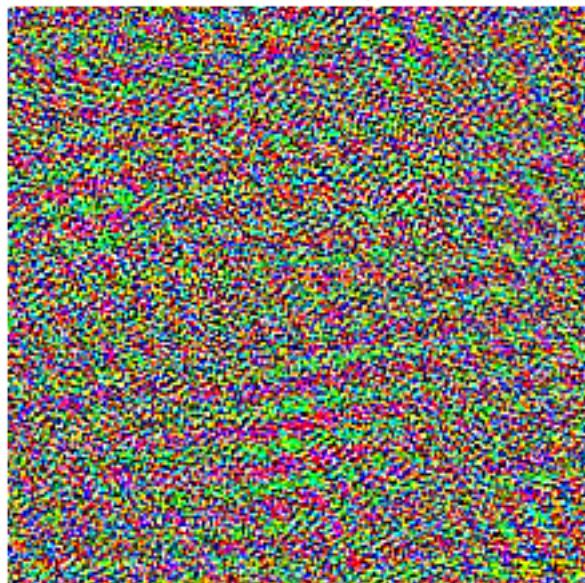
orange
Iteration 50 / 100



orange
Iteration 75 / 100



orange
Iteration 100 / 100



In []:

StyleTransfer-PyTorch

April 7, 2021

1 Style Transfer

In this notebook we will implement the style transfer technique from "["Image Style Transfer Using Convolutional Neural Networks" \(Gatys et al., CVPR 2015\)](#).

The general idea is to take two images, and produce a new image that reflects the content of one but the artistic "style" of the other. We will do this by first formulating a loss function that matches the content and style of each respective image in the feature space of a deep network, and then performing gradient descent on the pixels of the image itself.

The deep network we use as a feature extractor is [SqueezeNet](#), a small model that has been trained on ImageNet. You could use any network, but we chose SqueezeNet here for its small size and efficiency.

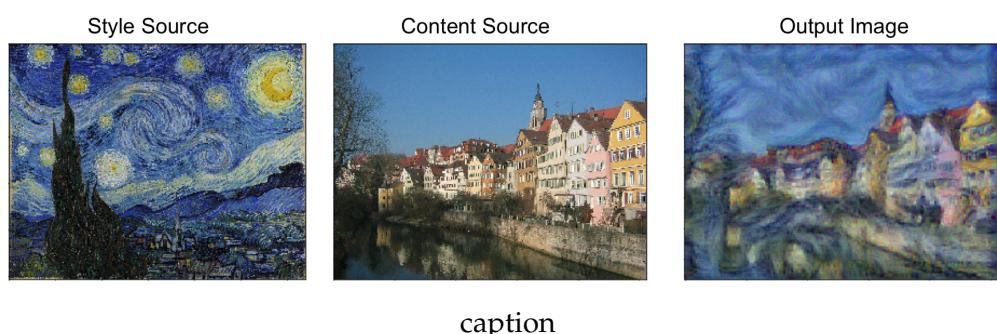
Here's an example of the images you'll be able to produce by the end of this notebook:

2 Part 0: Setup

```
In [1]: import torch
        import torch.nn as nn
        import torchvision
        import torchvision.transforms as T
        import PIL

        import numpy as np

        import matplotlib.pyplot as plt
```



```

from cs231n.image_utils import SQUEEZENET_MEAN, SQUEEZENET_STD
%matplotlib inline
# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2

```

We provide you with some helper functions to deal with images, since for this part of the assignment we're dealing with real JPEGs, not CIFAR-10 data.

In [2]: `from cs231n.style_transfer_pytorch import preprocess, deprocess, rescale, rel_error, f`

```

CHECKS_PATH = None

# Local
CHECKS_PATH = 'style-transfer-checks.npz'

# Colab
#CHECKS_PATH = '/content/drive/My Drive/{}{}'.format(FOLDERNAME, 'style-transfer-checks.npz')

assert CHECKS_PATH is not None, "[!] Choose path to style-transfer-checks.npz"

STYLES_FOLDER = CHECKS_PATH.replace('style-transfer-checks.npz', 'styles')

answers = dict(np.load(CHECKS_PATH))

```

Pytorch has two separate types for Tensors that contain floating-point numbers: one for operations on the CPU (`torch.FloatTensor`), and one using CUDA for operations on the GPU (`torch.cuda.FloatTensor`). We'll be using this type variable more later, so we need to set the tensor type to one of them.

In [11]: `# dtype = torch.FloatTensor`
`# Uncomment out the following line if you're on a machine with a GPU set up for PyTorch`
`dtype = torch.cuda.FloatTensor`

In [12]: `# Load the pre-trained SqueezeNet model.`
`cnn = torchvision.models.squeezeNet1_1(pretrained=True).features`
`cnn.type(dtype)`

`# We don't want to train the model any further, so we don't want PyTorch to waste computation`
`# computing gradients on parameters we're never going to update.`
`for param in cnn.parameters():`
 `param.requires_grad = False`

3 Part 1: Computing Loss

We're going to compute the three components of our loss function now. The loss function is a weighted sum of three terms: content loss + style loss + total variation loss. You'll fill in the functions that compute these weighted terms below.

3.1 Part 1A: Content loss

We can generate an image that reflects the content of one image and the style of another by incorporating both in our loss function. We want to penalize deviations from the content of the content image and deviations from the style of the style image. We can then use this hybrid loss function to perform gradient descent **not on the parameters** of the model, but instead **on the pixel values** of our original image.

Let's first write the content loss function. Content loss measures how much the feature map of the generated image differs from the feature map of the source image. We only care about the content representation of one layer of the network (say, layer ℓ), that has feature maps $A^\ell \in \mathbb{R}^{1 \times C_\ell \times H_\ell \times W_\ell}$. C_ℓ is the number of filters/channels in layer ℓ , H_ℓ and W_ℓ are the height and width. We will work with reshaped versions of these feature maps that combine all spatial positions into one dimension. Let $F^\ell \in \mathbb{R}^{C_\ell \times M_\ell}$ be the feature map for the current image and $P^\ell \in \mathbb{R}^{C_\ell \times M_\ell}$ be the feature map for the content source image where $M_\ell = H_\ell \times W_\ell$ is the number of elements in each feature map. Each row of F^ℓ or P^ℓ represents the vectorized activations of a particular filter, convolved over all positions of the image. Finally, let w_c be the weight of the content loss term in the loss function.

Then the content loss is given by:

$$L_c = w_c \times \sum_{i,j} (F_{ij}^\ell - P_{ij}^\ell)^2$$

Implement `content_loss` in `cs231n/style_transfer_pytorch.py`

Test your content loss. You should see errors less than 0.001.

In [16]: `from cs231n.style_transfer_pytorch import content_loss, extract_features, features_from`

```
def content_loss_test(correct):
    content_image = '%s/tubingen.jpg' % (STYLES_FOLDER)
    image_size = 192
    content_layer = 3
    content_weight = 6e-2

    c_feats, content_img_var = features_from_img(content_image, image_size, cnn)

    bad_img = torch.zeros(*content_img_var.data.size()).type(dtype)
    feats = extract_features(bad_img, cnn)

    student_output = content_loss(content_weight, c_feats[content_layer], feats[content_layer])
    error = rel_error(correct, student_output)
    print('Maximum error is {:.3f}'.format(error))

content_loss_test(answers['cl_out'])
```

Maximum error is 0.000

3.2 Part 1B: Style loss

Now we can tackle the style loss. For a given layer ℓ , the style loss is defined as follows:

First, compute the Gram matrix G which represents the correlations between the values in each channel of the feature map (i.e. the "responses" of the filter responsible for that channel), where F

is as above. The Gram matrix is an approximation of the covariance matrix -- it tells us how every channel's values (i.e. that filter's activations) correlate with every other channel's values. If we have C channels, matrix G will be of shape (C, C) to capture these correlations.

We want the activation statistics of our generated image to match the activation statistics of our style image, and matching the (approximate) covariance is one way to do that. There are a variety of ways you could do this, but the Gram matrix is nice because it's easy to compute and in practice shows good results.

Given a feature map F^ℓ of shape (C_ℓ, H_ℓ, W_ℓ) , we can flatten the height and width dimensions so they're just 1 dimension $M_\ell = H_\ell \times W_\ell$: the new shape of F^ℓ is (C_ℓ, M_ℓ) . Then, the Gram matrix has shape (C_ℓ, C_ℓ) where each element is given by the equation:

$$G_{ij}^\ell = \sum_k F_{ik}^\ell F_{jk}^\ell$$

Assuming G^ℓ is the Gram matrix from the feature map of the current image, A^ℓ is the Gram Matrix from the feature map of the source style image, and w_ℓ a scalar weight term, then the style loss for the layer ℓ is simply the weighted Euclidean distance between the two Gram matrices:

$$L_s^\ell = w_\ell \sum_{i,j} \left(G_{ij}^\ell - A_{ij}^\ell \right)^2$$

In practice we usually compute the style loss at a set of layers \mathcal{L} rather than just a single layer ℓ ; then the total style loss is the sum of style losses at each layer:

$$L_s = \sum_{\ell \in \mathcal{L}} L_s^\ell$$

Begin by implementing the Gram matrix computation function `gram_matrix` inside `cs231n\style_transfer_pytorch.py`:

Test your Gram matrix code. You should see errors less than 0.001.

```
In [17]: from cs231n.style_transfer_pytorch import gram_matrix
def gram_matrix_test(correct):
    style_image = '%s/starry_night.jpg' % (STYLES_FOLDER)
    style_size = 192
    feats, _ = features_from_img(style_image, style_size, cnn)
    student_output = gram_matrix(feats[5].clone()).cpu().data.numpy()
    error = rel_error(correct, student_output)
    print('Maximum error is {:.3f}'.format(error))

gram_matrix_test(answers['gm_out'])
```

Maximum error is 0.000

Next, put it together and implement the style loss function `style_loss` in `cs231n\style_transfer_pytorch.py`

Test your style loss implementation. The error should be less than 0.001.

```
In [18]: from cs231n.style_transfer_pytorch import style_loss
def style_loss_test(correct):
```

```

content_image = '%s/tubingen.jpg' % (STYLES_FOLDER)
style_image = '%s/starry_night.jpg' % (STYLES_FOLDER)
image_size = 192
style_size = 192
style_layers = [1, 4, 6, 7]
style_weights = [300000, 1000, 15, 3]

c_feats, _ = features_from_img(content_image, image_size, cnn)
feats, _ = features_from_img(style_image, style_size, cnn)
style_targets = []
for idx in style_layers:
    style_targets.append(gram_matrix(feats[idx].clone()))

student_output = style_loss(c_feats, style_layers, style_targets, style_weights)
error = rel_error(correct, student_output)
print('Error is {:.3f}'.format(error))

style_loss_test(answers['sl_out'])

```

Error is 0.000

3.3 Part 1C: Total-variation regularization

It turns out that it's helpful to also encourage smoothness in the image. We can do this by adding another term to our loss that penalizes wiggles or "total variation" in the pixel values.

You can compute the "total variation" as the sum of the squares of differences in the pixel values for all pairs of pixels that are next to each other (horizontally or vertically). Here we sum the total-variation regularization for each of the 3 input channels (RGB), and weight the total summed loss by the total variation weight, w_t :

$$L_{tv} = w_t \times \left(\sum_{c=1}^3 \sum_{i=1}^{H-1} \sum_{j=1}^W (x_{i+1,j,c} - x_{i,j,c})^2 + \sum_{c=1}^3 \sum_{i=1}^H \sum_{j=1}^{W-1} (x_{i,j+1,c} - x_{i,j,c})^2 \right)$$

In `cs231/style_transfer_pytorch.py`, fill in the definition for the TV loss term in `tv_loss`.

To receive full credit, your implementation should not have any loops.

Test your TV loss implementation. Error should be less than 0.0001.

```

In [19]: from cs231n.style_transfer_pytorch import tv_loss
        from inspect import getsourcelines
        import re

        def tv_loss_test(correct):
            content_image = '%s/tubingen.jpg' % (STYLES_FOLDER)
            image_size = 192
            tv_weight = 2e-2

            content_img = preprocess(PIL.Image.open(content_image), size=image_size).type(dtyp)

```

```

error = rel_error(correct, student_output)
print('Error is {:.4f}'.format(error))
lines, _ = getsourcelines(tv_loss)
used_loop = any(bool(re.search(r"for \S* in", line)) for line in lines)
if used_loop:
    print("WARNING!!!! - Your implementation of tv_loss contains a loop! To receive full credit, do not use loops in tv_loss.")

tv_loss_test(answers['tv_out'])

Error is 0.0000

```

4 Part 2: Style Transfer

Now we're ready to string it all together (you shouldn't have to modify this function):

```

In [20]: def style_transfer(content_image, style_image, image_size, style_size, content_layer,
                           style_layers, style_weights, tv_weight, init_random = False):
    """
    Run style transfer!
    """

    Inputs:
    - content_image: filename of content image
    - style_image: filename of style image
    - image_size: size of smallest image dimension (used for content loss and generating
    - style_size: size of smallest style image dimension
    - content_layer: layer to use for content loss
    - content_weight: weighting on content loss
    - style_layers: list of layers to use for style loss
    - style_weights: list of weights to use for each layer in style_layers
    - tv_weight: weight of total variation regularization term
    - init_random: initialize the starting image to uniform random noise
    """

    # Extract features for the content image
    content_img = preprocess(PIL.Image.open(content_image), size=image_size).type(dtype)
    feats = extract_features(content_img, cnn)
    content_target = feats[content_layer].clone()

    # Extract features for the style image
    style_img = preprocess(PIL.Image.open(style_image), size=style_size).type(dtype)
    feats = extract_features(style_img, cnn)
    style_targets = []
    for idx in style_layers:
        style_targets.append(gram_matrix(feats[idx].clone()))

    # Initialize output image to content image or noise
    if init_random:

```

```

        img = torch.Tensor(content_img.size()).uniform_(0, 1).type(dtype)
    else:
        img = content_img.clone().type(dtype)

    # We do want the gradient computed on our image!
    img.requires_grad_()

    # Set up optimization hyperparameters
    initial_lr = 3.0
    decayed_lr = 0.1
    decay_lr_at = 180

    # Note that we are optimizing the pixel values of the image by passing
    # in the img Torch tensor, whose requires_grad flag is set to True
    optimizer = torch.optim.Adam([img], lr=initial_lr)

    f, axarr = plt.subplots(1,2)
    axarr[0].axis('off')
    axarr[1].axis('off')
    axarr[0].set_title('Content Source Img.')
    axarr[1].set_title('Style Source Img.')
    axarr[0].imshow(deprocess(content_img.cpu()))
    axarr[1].imshow(deprocess(style_img.cpu()))
    plt.show()
    plt.figure()

    for t in range(200):
        if t < 190:
            img.data.clamp_(-1.5, 1.5)
            optimizer.zero_grad()

        feats = extract_features(img, cnn)

        # Compute loss
        c_loss = content_loss(content_weight, feats[content_layer], content_target)
        s_loss = style_loss(feats, style_layers, style_targets, style_weights)
        t_loss = tv_loss(img, tv_weight)
        loss = c_loss + s_loss + t_loss

        loss.backward()

        # Perform gradient descents on our image values
        if t == decay_lr_at:
            optimizer = torch.optim.Adam([img], lr=decayed_lr)
            optimizer.step()

        if t % 100 == 0:
            print('Iteration {}'.format(t))

```

```

    plt.axis('off')
    plt.imshow(deprocess(img.data.cpu()))
    plt.show()
    print('Iteration {}'.format(t))
    plt.axis('off')
    plt.imshow(deprocess(img.data.cpu()))
    plt.show()

```

4.1 Generate some pretty pictures!

Try out `style_transfer` on the three different parameter sets below. Make sure to run all three cells. Feel free to add your own, but make sure to include the results of style transfer on the third parameter set (starry night) in your submitted notebook.

- The `content_image` is the filename of content image.
- The `style_image` is the filename of style image.
- The `image_size` is the size of smallest image dimension of the content image (used for content loss and generated image).
- The `style_size` is the size of smallest style image dimension.
- The `content_layer` specifies which layer to use for content loss.
- The `content_weight` gives weighting on content loss in the overall loss function. Increasing the value of this parameter will make the final image look more realistic (closer to the original content).
- `style_layers` specifies a list of which layers to use for style loss.
- `style_weights` specifies a list of weights to use for each layer in `style_layers` (each of which will contribute a term to the overall style loss). We generally use higher weights for the earlier style layers because they describe more local/smaller scale features, which are more important to texture than features over larger receptive fields. In general, increasing these weights will make the resulting image look less like the original content and more distorted towards the appearance of the style image.
- `tv_weight` specifies the weighting of total variation regularization in the overall loss function. Increasing this value makes the resulting image look smoother and less jagged, at the cost of lower fidelity to style and content.

Below the next three cells of code (in which you shouldn't change the hyperparameters), feel free to copy and paste the parameters to play around them and see how the resulting image changes.

```
In [21]: # Composition VII + Tubingen
params1 = {
    'content_image' : '%s/tubingen.jpg' % (STYLES_FOLDER),
    'style_image' : '%s/composition_vii.jpg' % (STYLES_FOLDER),
    'image_size' : 192,
    'style_size' : 512,
    'content_layer' : 3,
    'content_weight' : 5e-2,
    'style_layers' : (1, 4, 6, 7),
    'style_weights' : (20000, 500, 12, 1),
    'tv_weight' : 5e-2
```

```
    }  
  
    style_transfer(**params1)
```

Content Source Img.



Style Source Img.



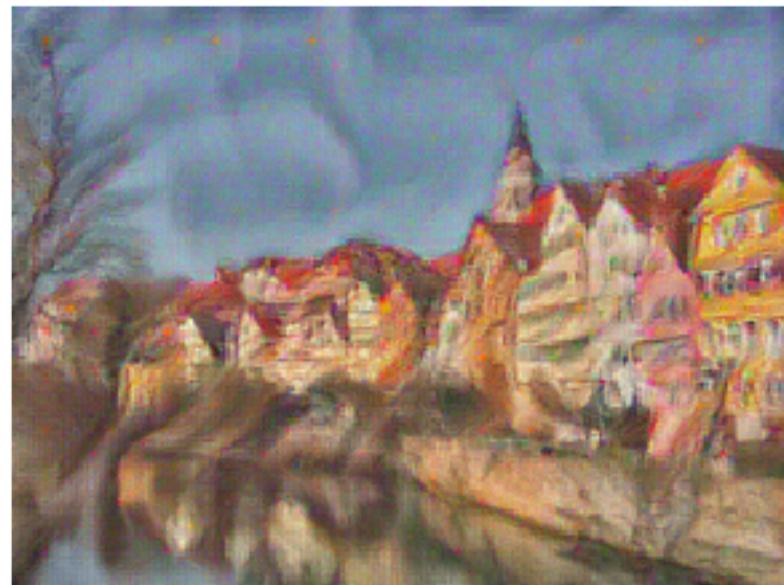
Iteration 0



Iteration 100



Iteration 199

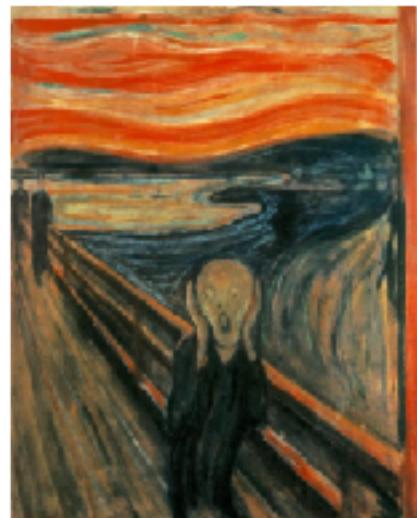


```
In [22]: # Scream + Tubingen
params2 = {
    'content_image': '%s/tubingen.jpg' % (STYLES_FOLDER),
    'style_image': '%s/the_scream.jpg' % (STYLES_FOLDER),
    'image_size': 192,
    'style_size': 224,
    'content_layer': 3,
    'content_weight': 3e-2,
    'style_layers': [1, 4, 6, 7],
    'style_weights': [200000, 800, 12, 1],
    'tv_weight': 2e-2
}
style_transfer(**params2)
```

Content Source Img.



Style Source Img.



Iteration 0



Iteration 100



Iteration 199



```
In [23]: # Starry Night + Tubingen
params3 = {
    'content_image' : '%s/tubingen.jpg' % (STYLES_FOLDER),
    'style_image' : '%s/starry_night.jpg' % (STYLES_FOLDER),
    'image_size' : 192,
    'style_size' : 192,
    'content_layer' : 3,
    'content_weight' : 6e-2,
    'style_layers' : [1, 4, 6, 7],
    'style_weights' : [300000, 1000, 15, 3],
    'tv_weight' : 2e-2
}
style_transfer(**params3)
```

Content Source Img.



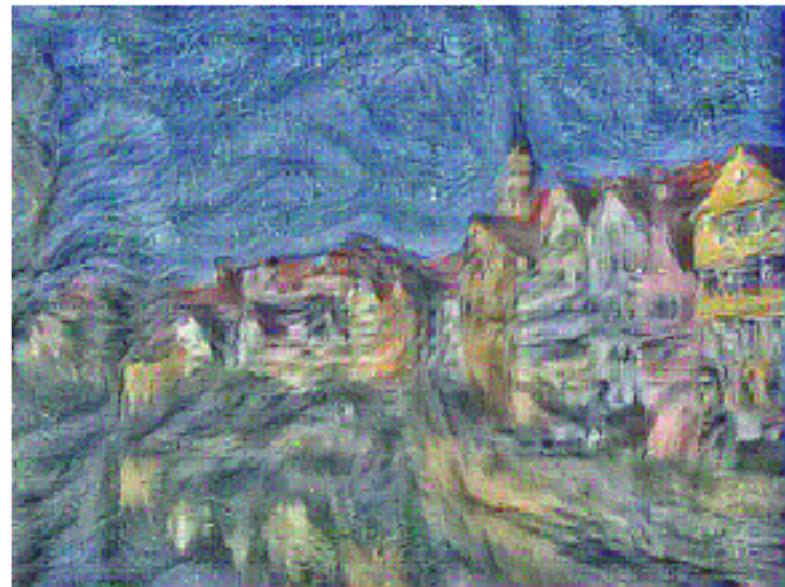
Style Source Img.



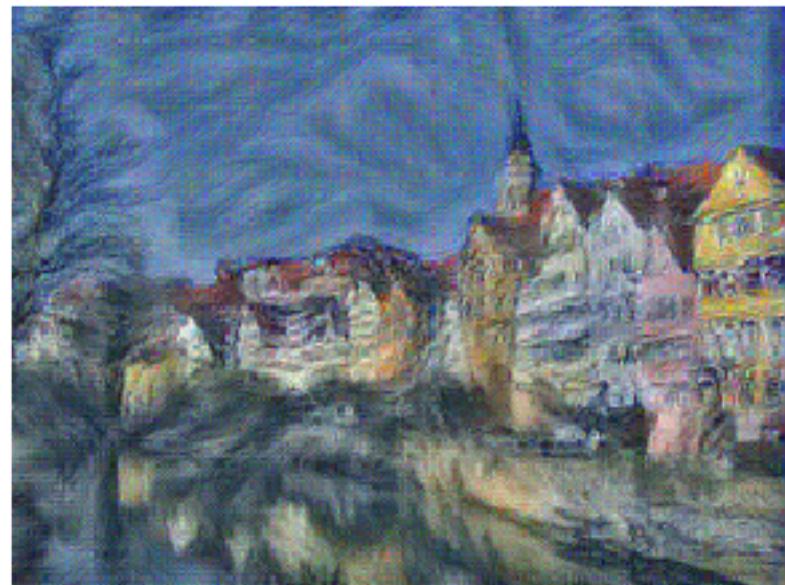
Iteration 0



Iteration 100



Iteration 199



5 Part 3: Feature Inversion

The code you've written can do another cool thing. In an attempt to understand the types of features that convolutional networks learn to recognize, a recent paper "[Understanding Deep Image Representations by Inverting Them](#)" attempts to reconstruct an image from its feature representation. We can easily implement this idea using image gradients from the pretrained network, which is exactly what we did above (but with two different feature representations).

Now, if you set the style weights to all be 0 and initialize the starting image to random noise instead of the content source image, you'll reconstruct an image from the feature representation of the content source image. You're starting with total noise, but you should end up with something that looks quite a bit like your original image.

(Similarly, you could do "texture synthesis" from scratch if you set the content weight to 0 and initialize the starting image to random noise, but we won't ask you to do that here.)

Run the following cell to try out feature inversion.

[1] Aravindh Mahendran, Andrea Vedaldi, "Understanding Deep Image Representations by Inverting Them", CVPR 2015

```
In [24]: # Feature Inversion -- Starry Night + Tubingen
params_inv = {
    'content_image' : '%s/tubingen.jpg' % (STYLES_FOLDER),
    'style_image' : '%s/starry_night.jpg' % (STYLES_FOLDER),
    'image_size' : 192,
    'style_size' : 192,
    'content_layer' : 3,
    'content_weight' : 6e-2,
    'style_layers' : [1, 4, 6, 7],
    'style_weights' : [0, 0, 0, 0], # we discard any contributions from style to the
    'tv_weight' : 2e-2,
    'init_random': True # we want to initialize our image to be random
}
style_transfer(**params_inv)
```

Content Source Img.



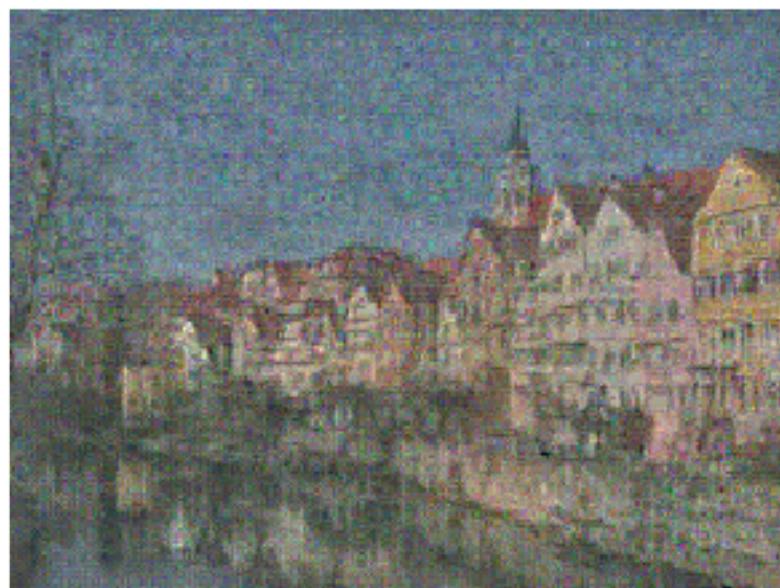
Style Source Img.



Iteration 0



Iteration 100



Iteration 199



In []:

Generative_Adversarial_Networks_PyTorch

April 7, 2021

1 Generative Adversarial Networks (GANs)

So far in CS231N, 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.

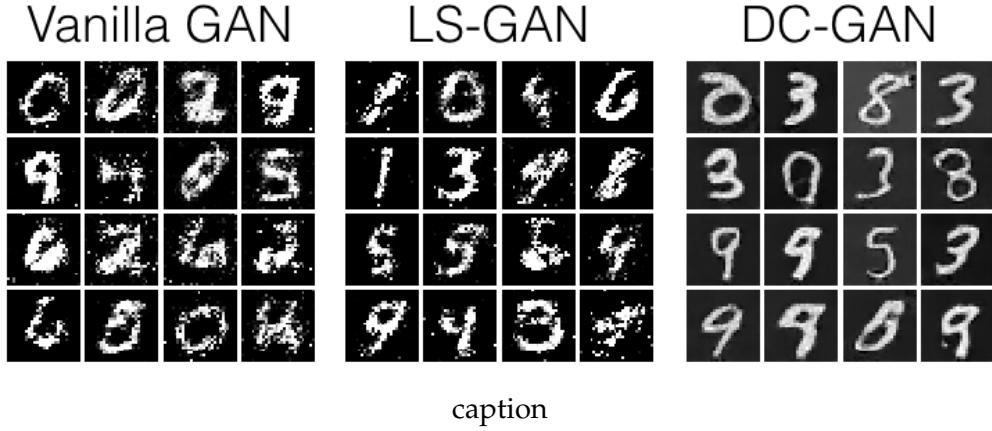
We can think of this back and forth process of the generator (G) trying to fool the discriminator (D), and the discriminator trying to correctly classify real vs. fake as a minimax game:

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

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

```
In [1]: 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'

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

def show_images(images):
    images = np.reshape(images, [images.shape[0], -1]) # images reshape to (batch_size, C*H*W)
    sqrt_n = int(np.ceil(np.sqrt(images.shape[0])))
    sqrt_im = int(np.ceil(np.sqrt(images.shape[1])))

    fig = plt.figure(figsize=(sqrt_n, sqrt_n))
    gs = gridspec.GridSpec(sqrt_n, sqrt_n)
    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([sqrt_im, sqrt_im]))
    return

```

In [2]: `from cs231n.gan_pytorch import preprocess_img, deprocess_img, rel_error, count_params,`
`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`.

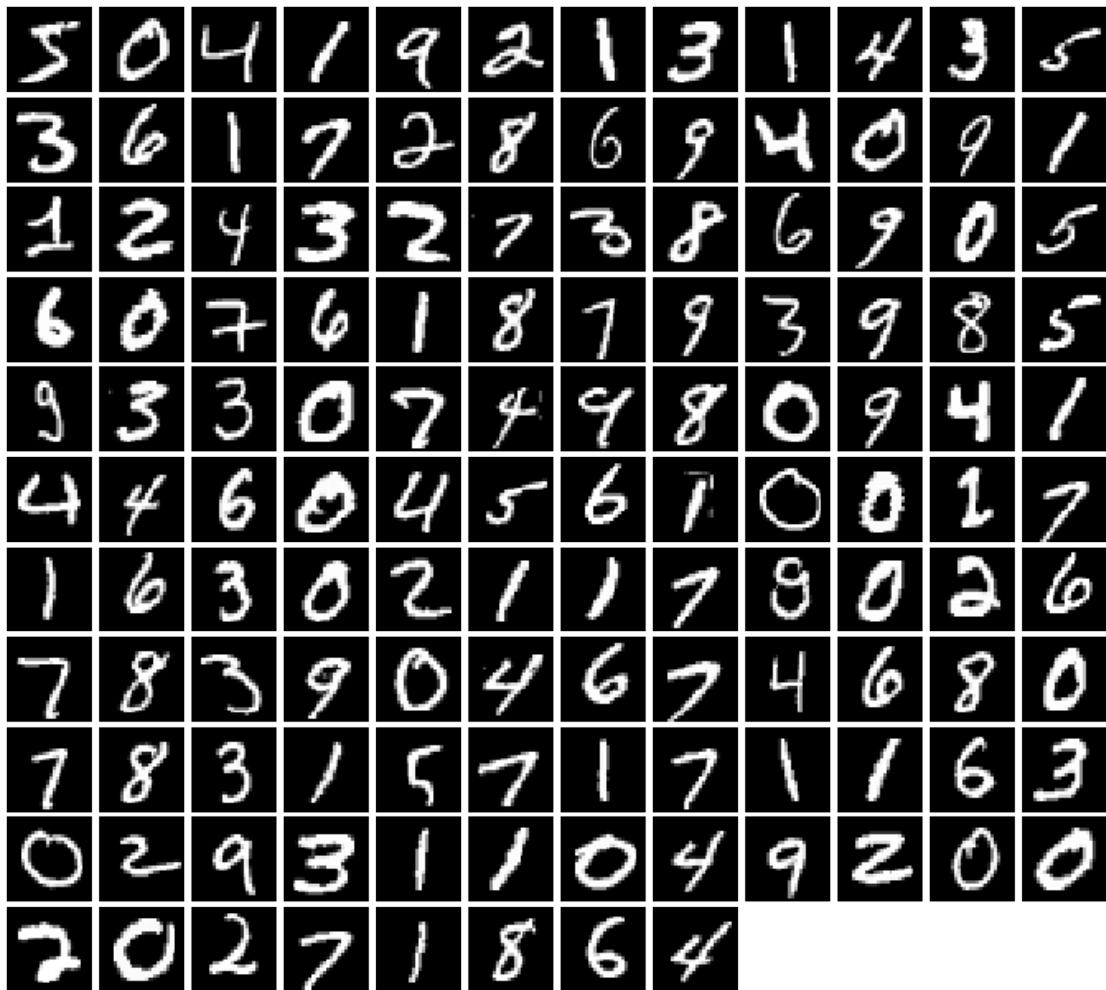
```
In [3]: NUM_TRAIN = 50000
        NUM_VAL = 5000

        NOISE_DIM = 96
        batch_size = 128

        mnist_train = dset.MNIST('./cs231n/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('./cs231n/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].

Implement `sample_noise` in `cs231n/gan_pytorch.py`.

Hint: use `torch.rand`.

Make sure noise is the correct shape and type:

In [4]: `from cs231n.gan_pytorch import sample_noise`

```
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.

In [5]: `from cs231n.gan_pytorch import Flatten, Unflatten, initialize_weights`

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. **If you are a Colab user, it is recommended to change colab runtime to GPU.**

In [6]: `# dtype = torch.FloatTensor`
`dtype = torch.cuda.FloatTensor`

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 α ; 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.

Implement discriminator in `cs231n/gan_pytorch.py`

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

In [7]: `from cs231n.gan_pytorch import discriminator`

```

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])

Implement generator in cs231n/gan_pytorch.py

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

```
In [8]: from cs231n.gan_pytorch import generator
```

```

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.

Implement `bce_loss`, `discriminator_loss`, `generator_loss` in `cs231n/gan_pytorch.py`

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

```
In [9]: from cs231n.gan_pytorch import bce_loss, discriminator_loss, generator_loss

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

```
In [10]: 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, $\text{beta1}=0.5$, $\text{beta2}=0.999$. You'll use this to construct optimizers for the generators and discriminators for the rest of the notebook.

Implement `get_optimizer` in `cs231n/gan_pytorch.py`

6 Training a GAN!

We provide you the main training loop... you won't need to change `run_a_gan` in `cs231n/gan_pytorch.py`, but we encourage you to read through and understand it.

```
In [11]: from cs231n.gan_pytorch import get_optimizer, run_a_gan

In [12]: # Make the discriminator
          D = discriminator().type(dtype)

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

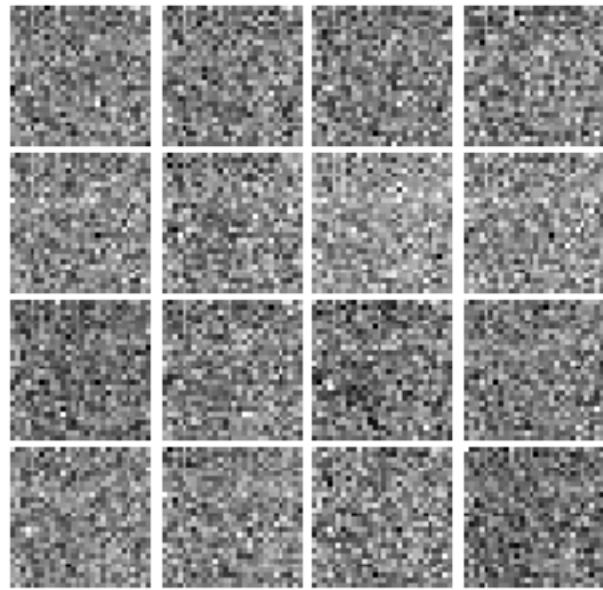
          # Use the function you wrote earlier to get optimizers for the Discriminator and the Generator
          D_solver = get_optimizer(D)
          G_solver = get_optimizer(G)
          # Run it!
          images = run_a_gan(D, G, D_solver, G_solver, discriminator_loss, generator_loss, loader)

Iter: 0, D: 1.328, G:0.7202
Iter: 250, D: 1.245, G:1.103
Iter: 500, D: 1.175, G:1.068
Iter: 750, D: 1.36, G:0.7679
Iter: 1000, D: 1.216, G:1.032
Iter: 1250, D: 1.174, G:0.997
Iter: 1500, D: 1.306, G:0.8871
Iter: 1750, D: 1.25, G:1.137
Iter: 2000, D: 1.282, G:0.8806
Iter: 2250, D: 1.389, G:1.246
Iter: 2500, D: 1.243, G:0.974
Iter: 2750, D: 1.374, G:0.909
Iter: 3000, D: 1.27, G:0.7621
Iter: 3250, D: 1.31, G:0.8235
Iter: 3500, D: 1.318, G:0.771
Iter: 3750, D: 1.326, G:0.892
```

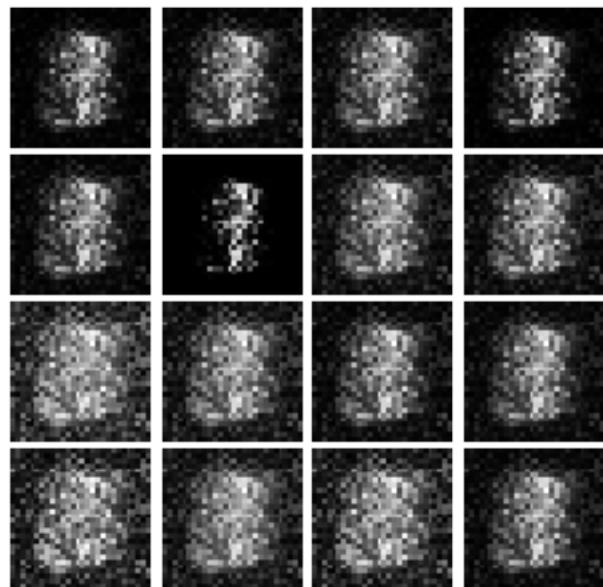
Run the cell below to show the generated images.

```
In [13]: numIter = 0
          for img in images:
              print("Iter: {}".format(numIter))
              show_images(img)
              plt.show()
              numIter += 250
              print()
```

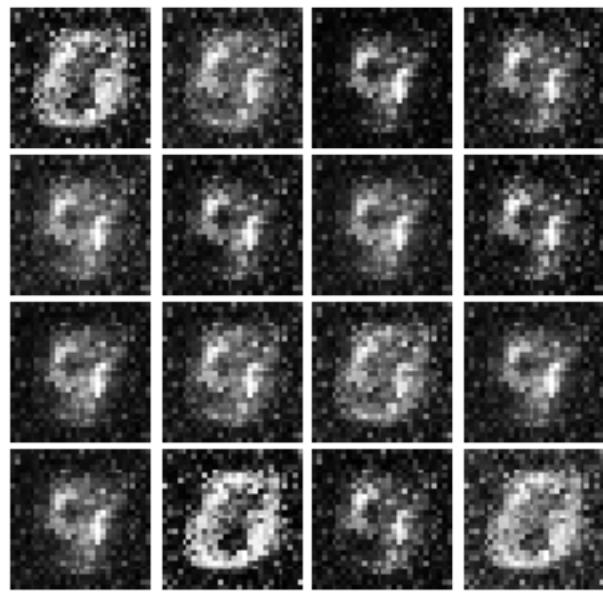
```
Iter: 0
```



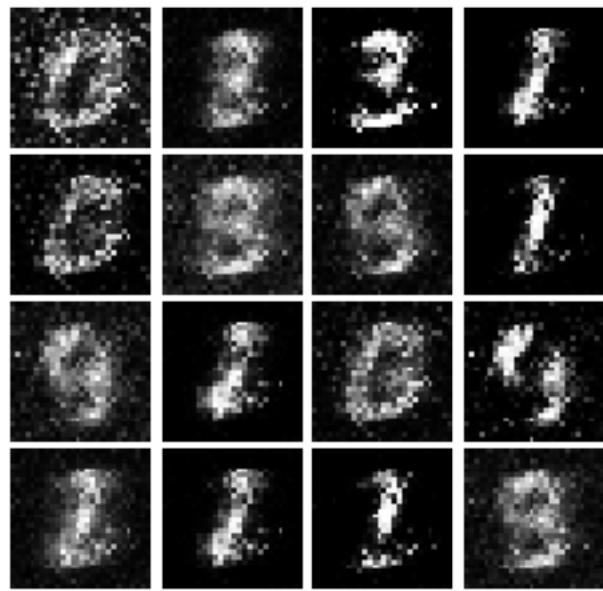
Iter: 250



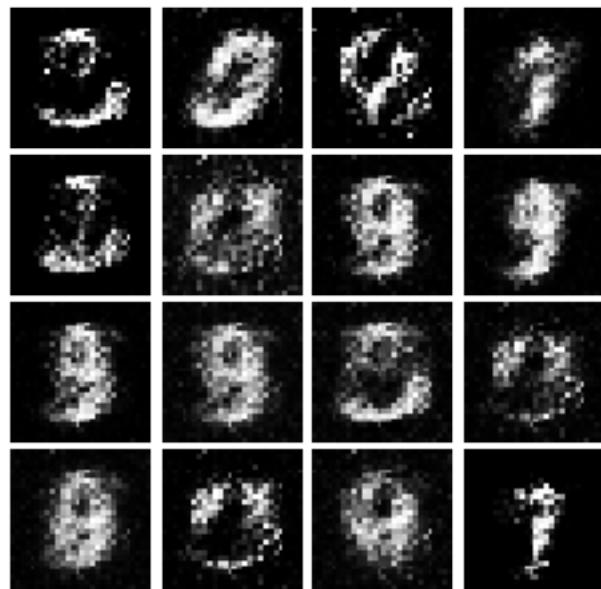
Iter: 500



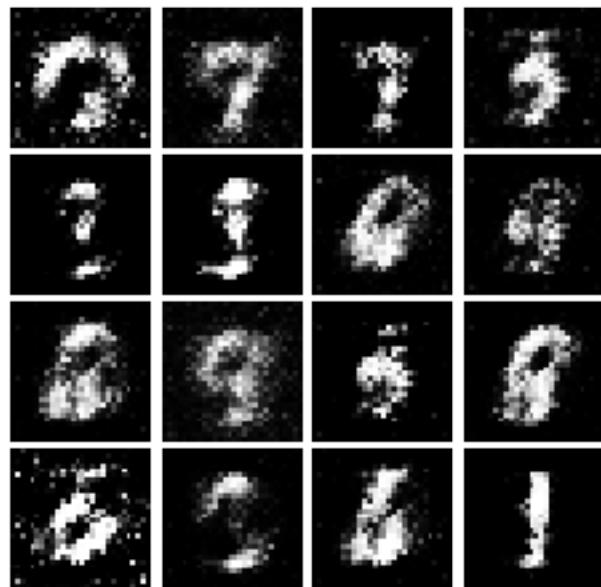
Iter: 750



Iter: 1000



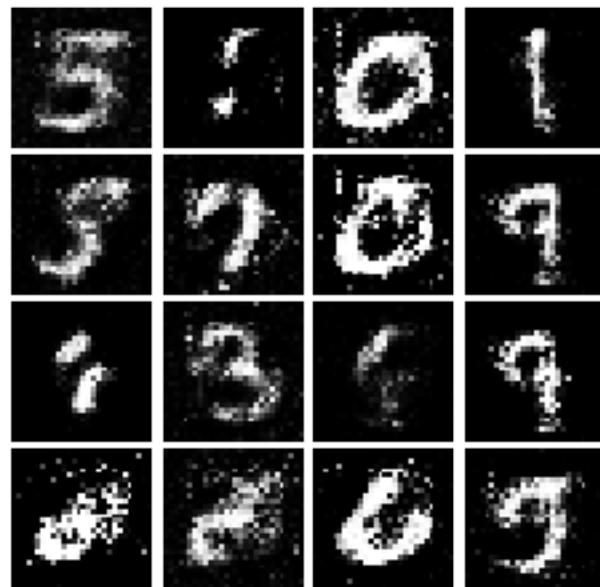
Iter: 1250



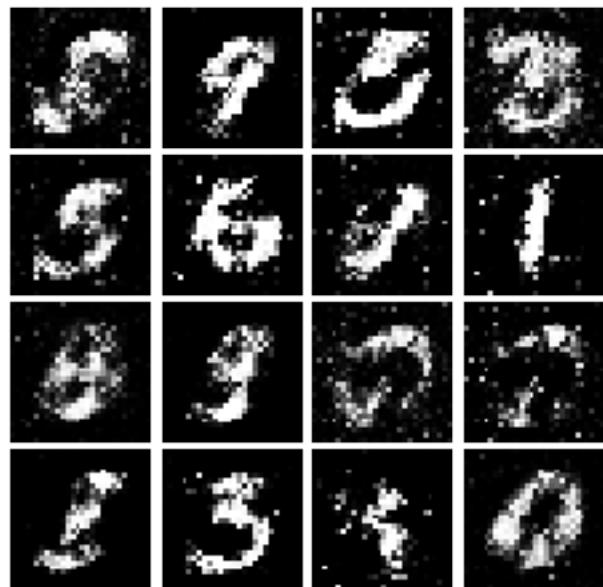
Iter: 1500



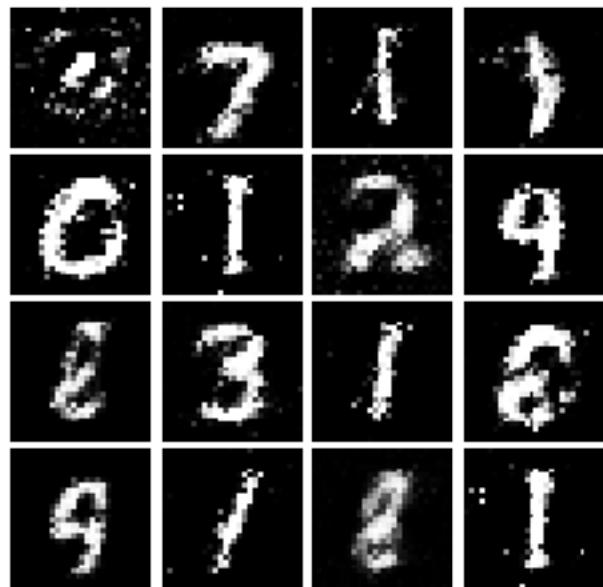
Iter: 1750



Iter: 2000



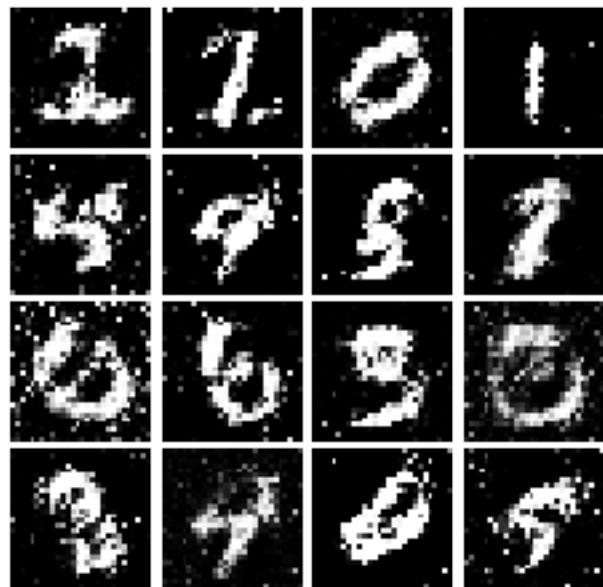
Iter: 2250



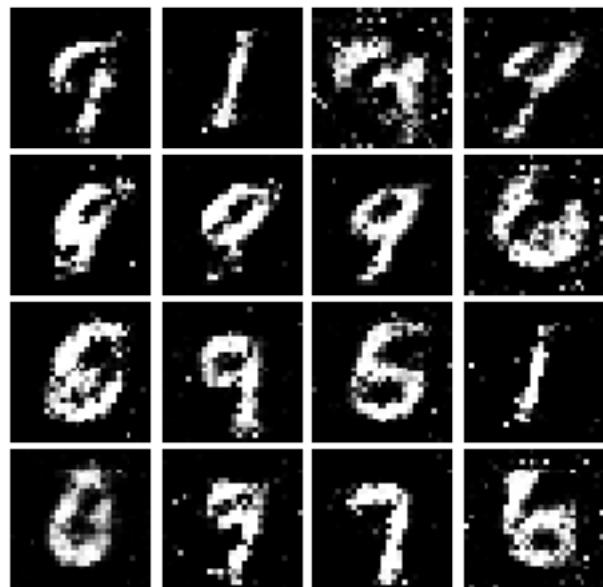
Iter: 2500



Iter: 2750



Iter: 3000



Iter: 3250



Iter: 3500



Iter: 3750



Please tag the cell below on Gradescope while submitting.

```
In [14]: print("Vanilla GAN Fianl image:")
show_images(images[-1])
plt.show()
```

Vanilla GAN Fianl image:



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`).

Implement `ls_discriminator_loss`, `ls_generator_loss` in `cs231n/gan_pytorch.py`

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

```
In [15]: from cs231n.gan_pytorch import ls_discriminator_loss, ls_generator_loss

    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: 1.64377e-08
Maximum error in g_loss: 2.7837e-09
```

Run the following cell to train your model!

```
In [16]: D_LS = discriminator().type(dtype)
        G_LS = generator().type(dtype)

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

        images = run_a_gan(D_LS, G_LS, D_LS_solver, G_LS_solver, ls_discriminator_loss, ls_gen...
```

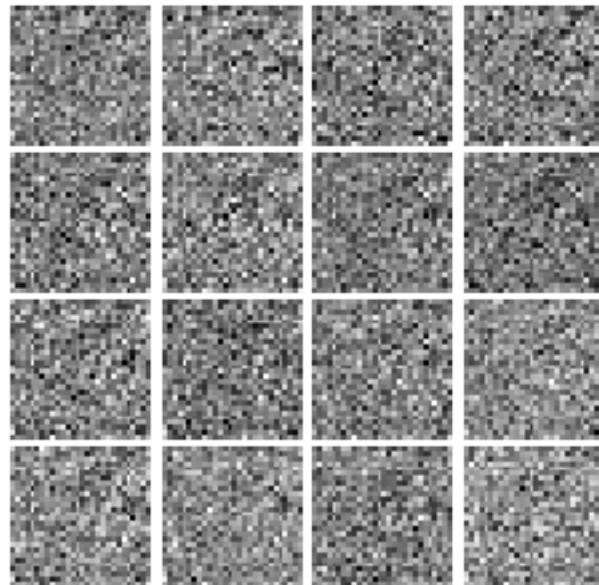
Iter: 0, D: 0.5689, G:0.51
Iter: 250, D: 0.7447, G:0.596
Iter: 500, D: 0.1766, G:0.6847
Iter: 750, D: 0.2041, G:0.2435
Iter: 1000, D: 0.1592, G:0.3196
Iter: 1250, D: 0.1197, G:0.3032
Iter: 1500, D: 0.1171, G:0.3392
Iter: 1750, D: 0.1624, G:0.3267
Iter: 2000, D: 0.2187, G:0.2212
Iter: 2250, D: 0.2123, G:0.1661
Iter: 2500, D: 0.2189, G:0.1746
Iter: 2750, D: 0.2117, G:0.1856
Iter: 3000, D: 0.2557, G:0.1662
Iter: 3250, D: 0.2192, G:0.1812
Iter: 3500, D: 0.2266, G:0.1981
Iter: 3750, D: 0.2333, G:0.1745

Run the cell below to show generated images.

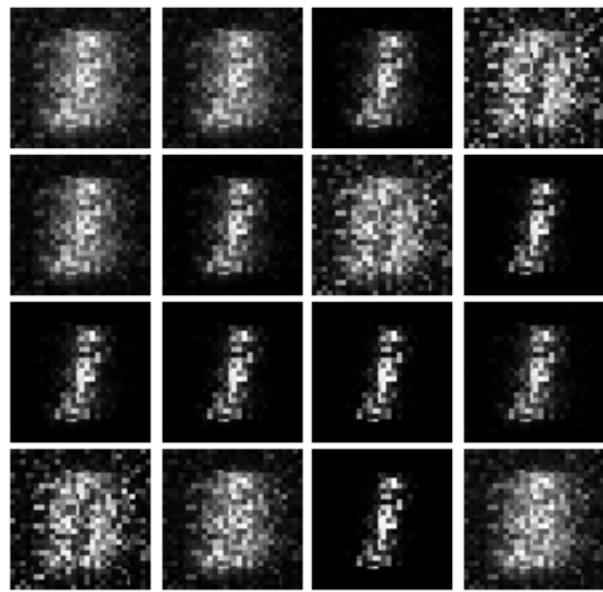
```
In [17]: numIter = 0
        for img in images:
```

```
print("Iter: {}".format(numIter))
show_images(img)
plt.show()
numIter += 250
print()
```

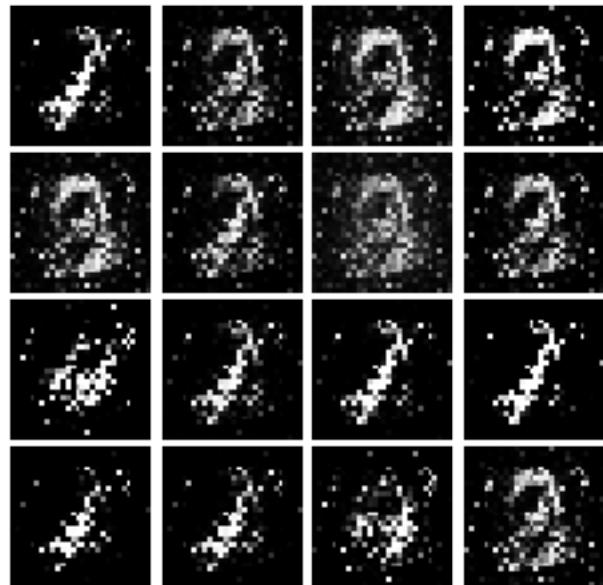
Iter: 0



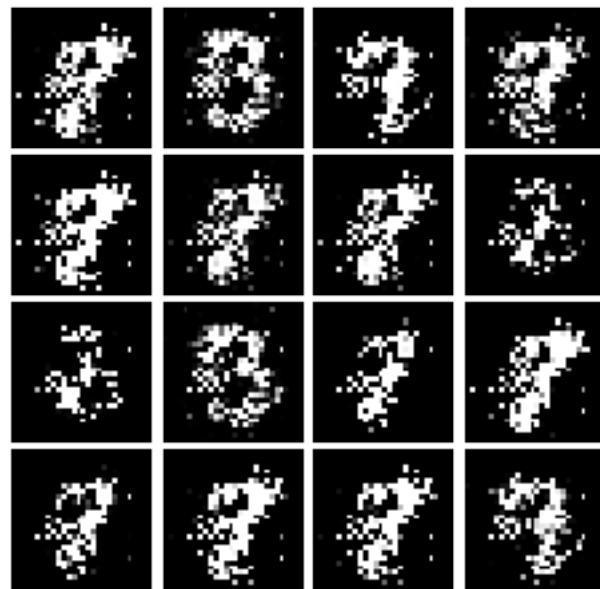
Iter: 250



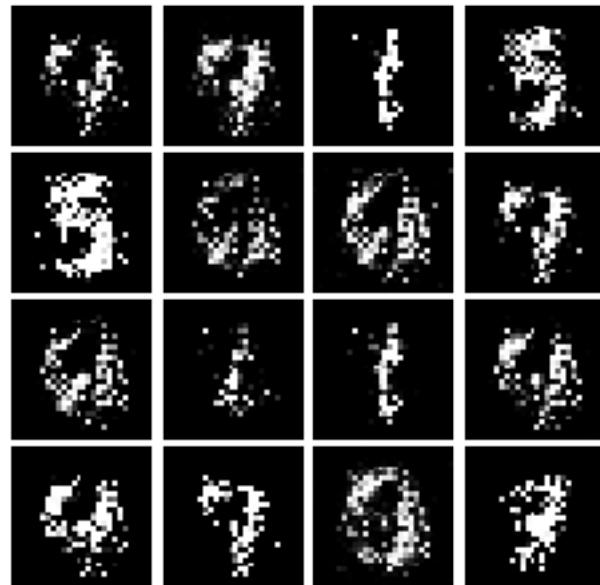
Iter: 500



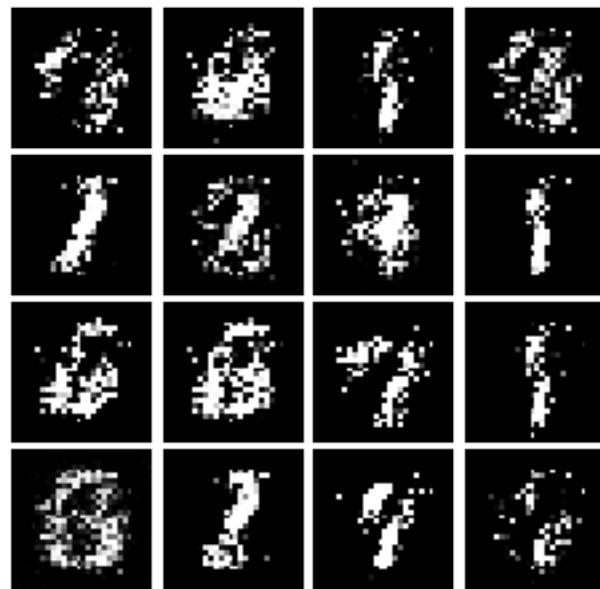
Iter: 750



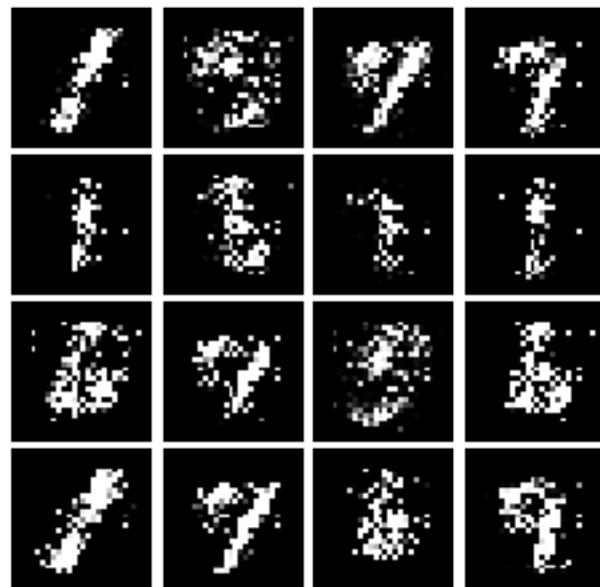
Iter: 1000



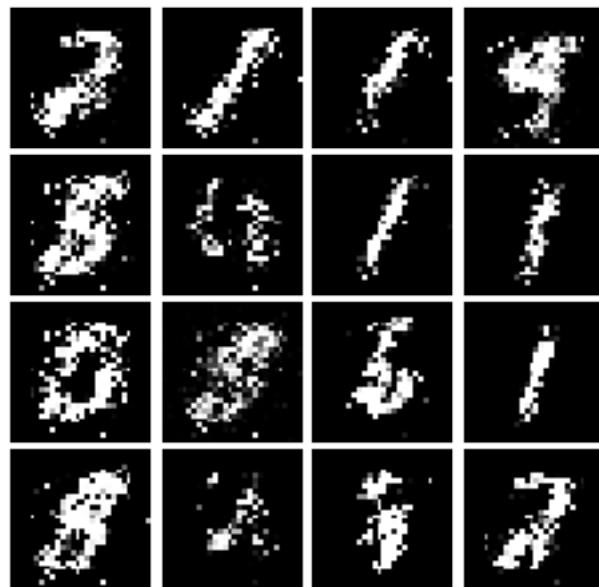
Iter: 1250



Iter: 1500



Iter: 1750



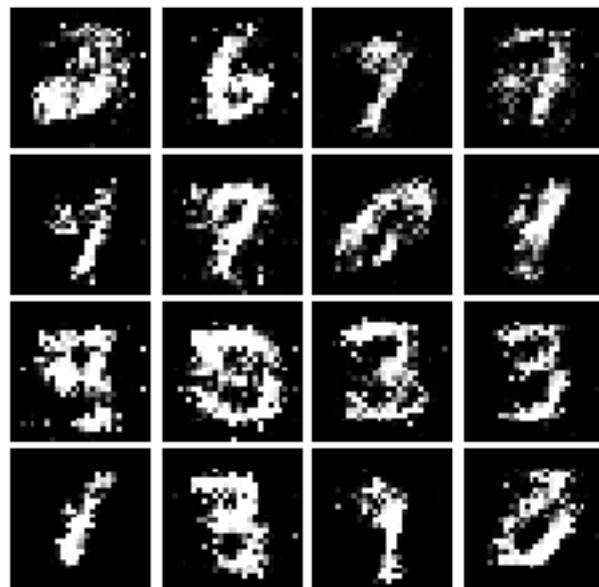
Iter: 2000



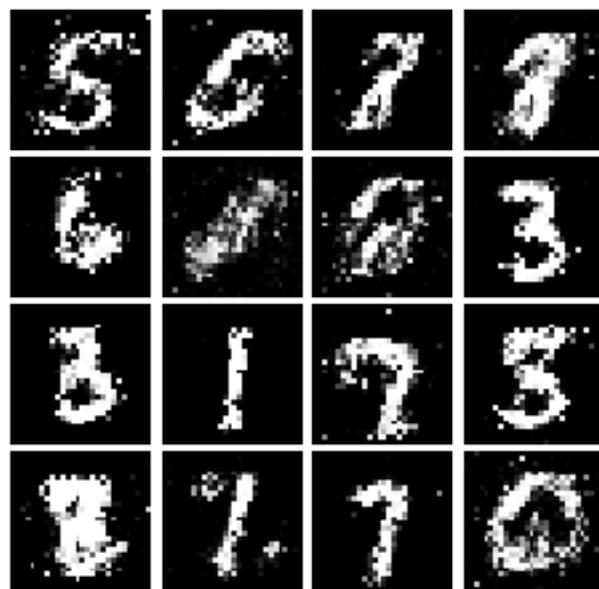
Iter: 2250



Iter: 2500



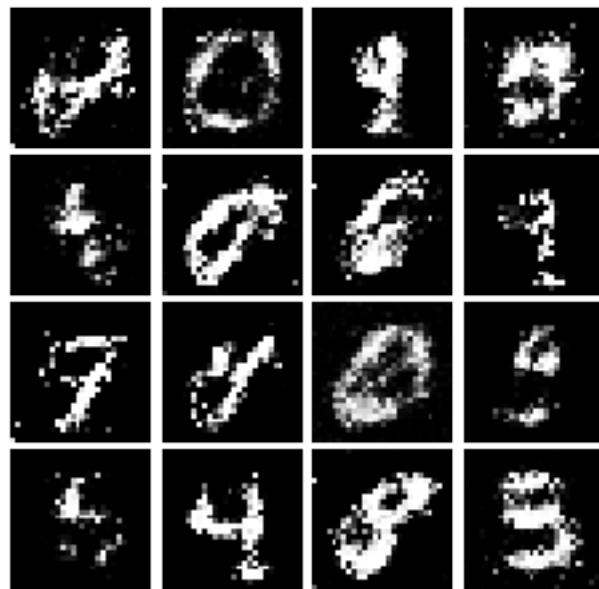
Iter: 2750



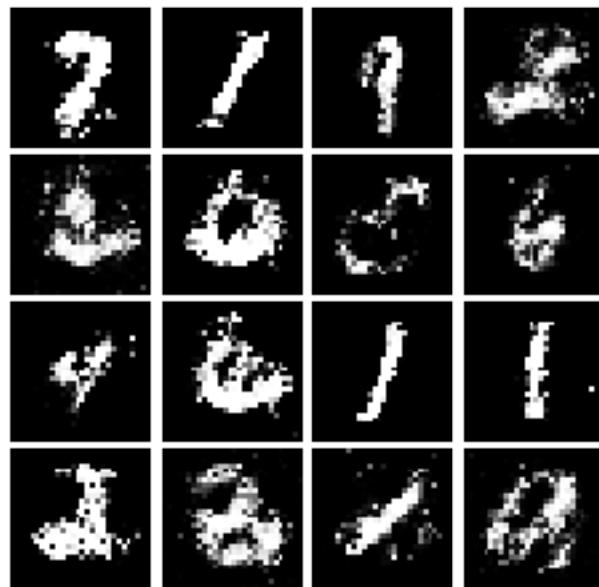
Iter: 3000



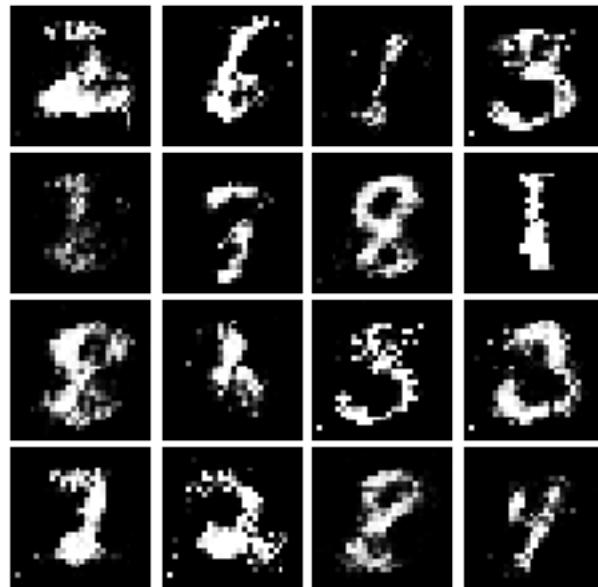
Iter: 3250



Iter: 3500



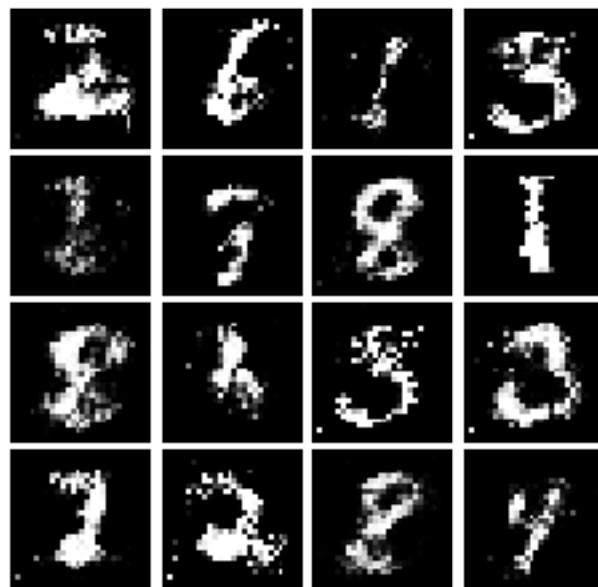
Iter: 3750



Please tag the cell below on Gradescope while submitting.

```
In [18]: print("LSGAN Fianl image:")
    show_images(images[-1])
    plt.show()
```

LSGAN Fianl image:



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

Implement `build_dc_classifier` in `cs231n/gan_pytorch.py`

```
In [19]: from cs231n.gan_pytorch import build_dc_classifier
```

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

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

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

```
In [20]: def test_dc_classifier(true_count=1102721):
    model = build_dc_classifier(batch_size)
    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 x 7 x 128
- * ReLU
- * BatchNorm
- * Reshape into Image Tensor of shape

7, 7, 128 * Conv2D^T (Transpose): 64 filters of 4x4, stride 2, 'same' padding (use padding=1) * ReLU
* BatchNorm * Conv2D^T (Transpose): 1 filter of 4x4, stride 2, 'same' padding (use padding=1) *
TanH * Should have a 28x28x1 image, reshape back into 784 vector

Implement build_dc_generator in cs231n/gan_pytorch.py

In [21]: `from cs231n.gan_pytorch import build_dc_generator`

```
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()
```

Out [21]: `torch.Size([128, 784])`

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

In [22]: `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.

In [23]: `D_DC = build_dc_classifier(batch_size).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)
```

```
images = run_a_gan(D_DC, G_DC, D_DC_solver, G_DC_solver, discriminator_loss, generator_loss)
```

Iter: 0, D: 1.448, G:1.464

Iter: 250, D: 1.312, G:0.8799

Iter: 500, D: 1.277, G:1.001

Iter: 750, D: 1.173, G:0.8774

Iter: 1000, D: 1.301, G:0.8848

Iter: 1250, D: 1.187, G:1.081

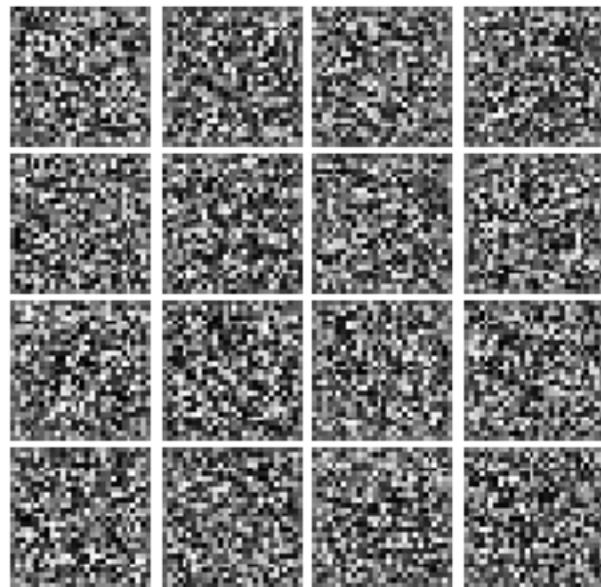
Iter: 1500, D: 1.128, G:1.176

Iter: 1750, D: 1.039, G:1.051

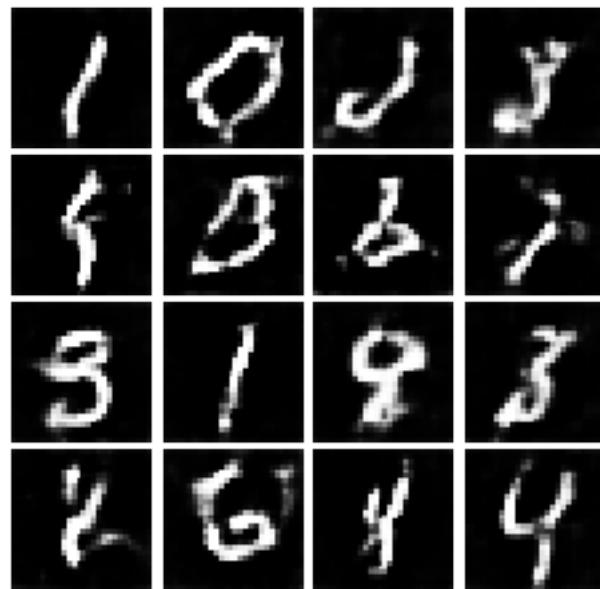
Run the cell below to show generated images.

```
In [24]: numIter = 0
      for img in images:
          print("Iter: {}".format(numIter))
          show_images(img)
          plt.show()
          numIter += 250
          print()
```

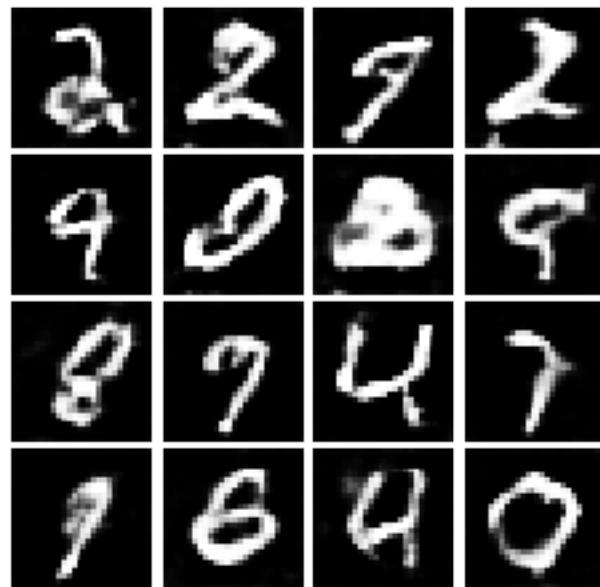
Iter: 0



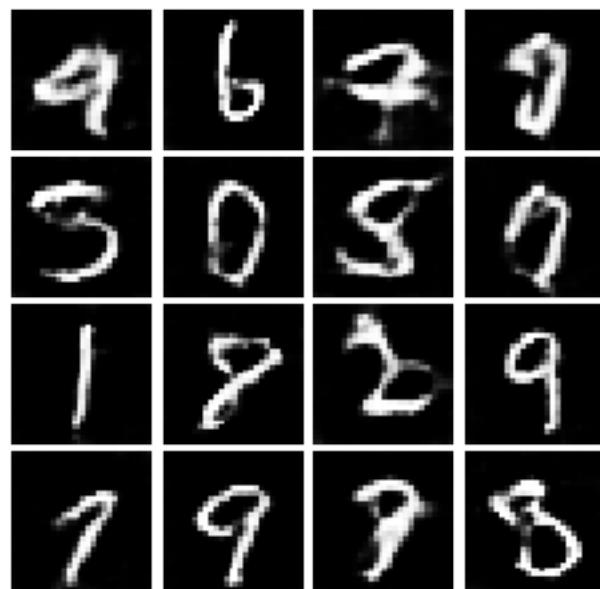
Iter: 250



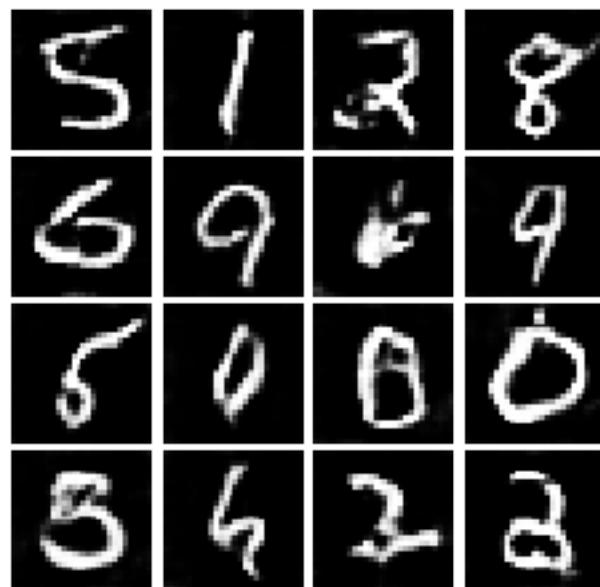
Iter: 250



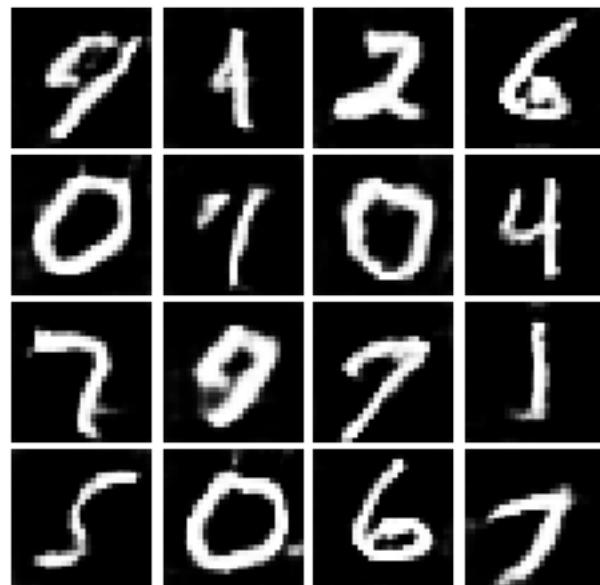
Iter: 500



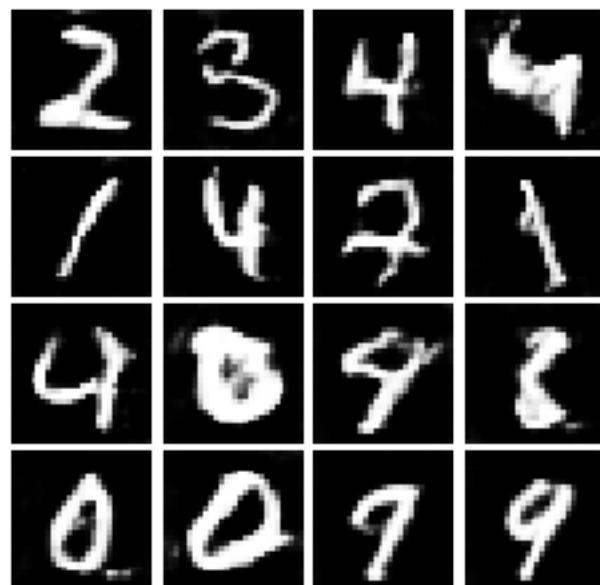
Iter: 1000



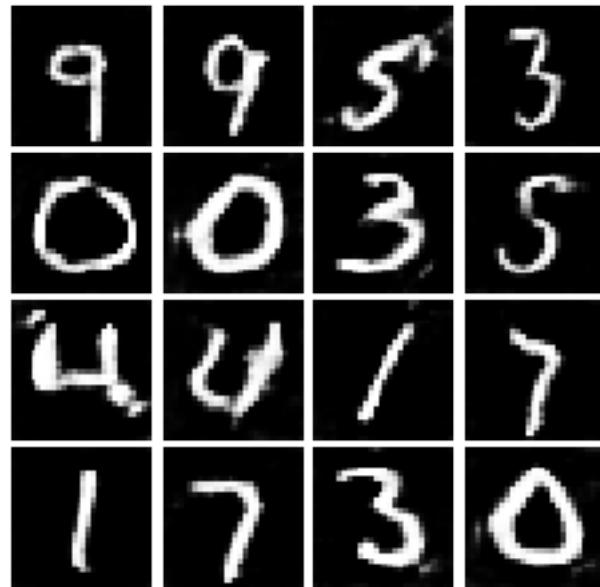
Iter: 1250



Iter: 1500



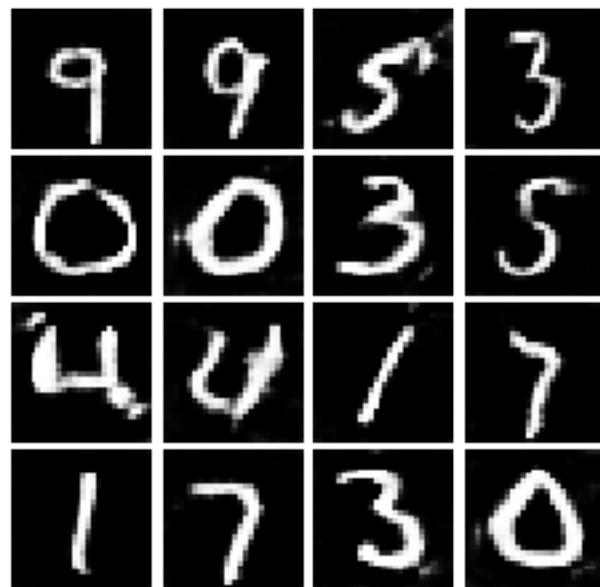
Iter: 1750



Please tag the cell below on Gradescope while submitting.

```
In [25]: print("DCGAN Fianl image:")
    show_images(images[-1])
    plt.show()
```

DCGAN Fianl image:



8.1 INLINE QUESTION 1

We will look at an example to see why alternating minimization of the same objective (like in a GAN) can be tricky business.

Consider $f(x, y) = xy$. What does $\min_x \max_y f(x, y)$ evaluate to? (Hint: minmax tries to minimize the maximum value achievable.)

Now try to evaluate this function numerically for 6 steps, starting at the point $(1, 1)$, by using alternating gradient (first updating y , then updating x using that updated y) with step size 1. **Here step size is the learning_rate, and steps will be learning_rate * gradient.** You'll find that writing out the update step in terms of $x_t, y_t, x_{t+1}, y_{t+1}$ will be useful.

Breifly explain what $\min_x \max_y f(x, y)$ evaluates to and record the six pairs of explicit values for (x_t, y_t) in the table below.

8.1.1 Your answer:

we have: $y_{t+1} = y_t + \alpha \times \frac{df}{dy_t} = y_t + x_t$ $x_{t+1} = x_t - \alpha \times \frac{df}{dx_t} = x_t - y_{t+1} = -y_t$ So we can update x, y as follows, which is a loop:

y_0	y_1	y_2	y_3	y_4	y_5	y_6
1	2	1	-1	-2	-1	1
x_0	x_1	x_2	x_3	x_4	x_5	x_6
1	-1	-2	-1	1	2	1

8.2 INLINE QUESTION 2

Using this method, will we ever reach the optimal value? Why or why not?

8.2.1 Your answer:

NO. We will not reach the optimal value. As we can see in question 1, x, y are trapped in a circle of every 6 step, the value of x, y just oscillating, and never converge.

8.3 INLINE QUESTION 3

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.

8.3.1 Your answer:

This is not a good sign. * A high value of the discriminator's loss implies that the discriminator hasn't learnt a good distribution of real dataset, which leads to pass fake images as real ones. This may be because the network architecture is too simple and the poor discriminator will also fail to provide a meaningful signal to the generator. * Poor discriminator likely pass images as real ones, which gives the generator the illusion that the generated image is realistic, which also leads to a drop in the generator loss.

In []: