

# RNN\_Captioning

February 6, 2022

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

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

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

# This downloads the COCO dataset to your Drive
# if it doesn't already exist.
%cd /content/drive/My\ Drive/$FOLDERNAME/cs231n/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME
```

```
Mounted at /content/drive
/content/drive/My Drive/cs231n/assignments/assignment3/cs231n/datasets
--2022-02-04 14:33:42-- http://cs231n.stanford.edu/coco_captioning.zip
Resolving cs231n.stanford.edu (cs231n.stanford.edu)... 171.64.68.10
Connecting to cs231n.stanford.edu (cs231n.stanford.edu)|171.64.68.10|:80...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 1035210391 (987M) [application/zip]
Saving to: coco_captioning.zip

coco_captioning.zip 100%[=====] 987.25M 36.3MB/s    in 31s

2022-02-04 14:34:13 (31.9 MB/s) - coco_captioning.zip saved
[1035210391/1035210391]
```

```

Archive: coco_captioning.zip
  creating: coco_captioning/
  inflating: coco_captioning/coco2014_captions.h5
  inflating: coco_captioning/coco2014_vocab.json
  inflating: coco_captioning/train2014_images.txt
  inflating: coco_captioning/train2014_urls.txt
  inflating: coco_captioning/train2014_vgg16_fc7.h5
  inflating: coco_captioning/train2014_vgg16_fc7_pca.h5
  inflating: coco_captioning/val2014_images.txt
  inflating: coco_captioning/val2014_urls.txt
  inflating: coco_captioning/val2014_vgg16_fc7.h5
  inflating: coco_captioning/val2014_vgg16_fc7_pca.h5
--2022-02-04 14:34:53-- http://cs231n.stanford.edu/imagenet_val_25.npz
Resolving cs231n.stanford.edu (cs231n.stanford.edu)... 171.64.68.10
Connecting to cs231n.stanford.edu (cs231n.stanford.edu)|171.64.68.10|:80...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 3940548 (3.8M) [text/plain]
Saving to: imagenet_val_25.npz

imagenet_val_25.npz 100%[=====] 3.76M 1.63MB/s in 2.3s

2022-02-04 14:34:55 (1.63 MB/s) - imagenet_val_25.npz saved [3940548/3940548]

/content/drive/My Drive/cs231n/assignments/assignment3

```

## 1 Image Captioning with RNNs

In this exercise, you will implement vanilla Recurrent Neural Networks and use them to train a model that can generate novel captions for images.

```
[2]: # Setup cell.
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'

%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 COCO Dataset

For this exercise, we will use the 2014 release of the [COCO dataset](#), a 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.

**Image features.** 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 pre-trained on ImageNet, and these features are stored in the files `train2014_vgg16_fc7.h5` and `val2014_vgg16_fc7.h5`. To cut down on processing time and memory requirements, we have reduced the dimensionality of the features from 4096 to 512 using Principal Component Analysis (PCA), and these features are stored in the files `train2014_vgg16_fc7_pca.h5` and `val2014_vgg16_fc7_pca.h5`. The raw images take up nearly 20GB of space so we have not included them in the download. Since all images are taken from Flickr, we have stored the URLs of the training and validation images in the files `train2014_urls.txt` and `val2014_urls.txt`. This allows you to download images on-the-fly for visualization.

**Captions.** 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.

**Tokens.** There are a couple special tokens that we add to the vocabulary, and we have taken care of all implementation details around special tokens for you. 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.

You can load all of the 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:

```
[3]: # Load COCO data from disk into a dictionary.
# We'll work with dimensionality-reduced features for the remainder of this ↴ assignment,
# but you can also experiment with the original features on your own 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 /content/drive/My
Drive/cs231n/assignments/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 Inspect 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.

```
[4]: # Sample a minibatch and show the images and captions.
# If you get an error, the URL just no longer exists, so don't worry!
# You can re-sample as many times as you want.

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 man flying a kite with lots of other people flying kites <END>



<START> a field with no grass is well made and a train moving <UNK> it <END>



<START> the room has two beds a <UNK> and a <UNK> with three <UNK> on it <END>



### 3 Recurrent Neural Network

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`.

**NOTE:** The Long-Short Term Memory (LSTM) RNN is a common variant of the vanilla RNN. `LSTM_Captioning.ipynb` is optional extra credit, so don't worry about references to LSTM in `cs231n/classifiers/rnn.py` and `cs231n/rnn_layers.py` for now.

### 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.

```
[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.

```
[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.

[7]: 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.

```
[8]: 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.

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

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

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

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

```
out error: 1.000000094736443e-08
```

## 9 Word Embedding: Backward

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

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

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

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

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

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

```
dW error: 3.2774595693100364e-12
```

## 10 Temporal Affine Layer

At every timestep we use an affine function to transform the RNN hidden vector at that timestep into scores for each word in the vocabulary. Because this is very similar to the affine layer that you implemented in assignment 2, we have provided this function for you in the `temporal_affine_forward` and `temporal_affine_backward` functions in the file `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.

```
[11]: 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.

```
[12]: # 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, verbose=False)

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

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

## 12 RNN for Image Captioning

Now that you have implemented the necessary layers, you can combine them to build an image captioning model. Open the file `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 vanilla 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.

```
[13]: 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.

```
[14]: 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, h=1e-6)
    e = rel_error(param_grad_num, grads[param_name])
    print('%s relative error: %e' % (param_name, e))

```

```

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

```

## 13 Overfit RNN Captioning Model on 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.

[15]: `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()

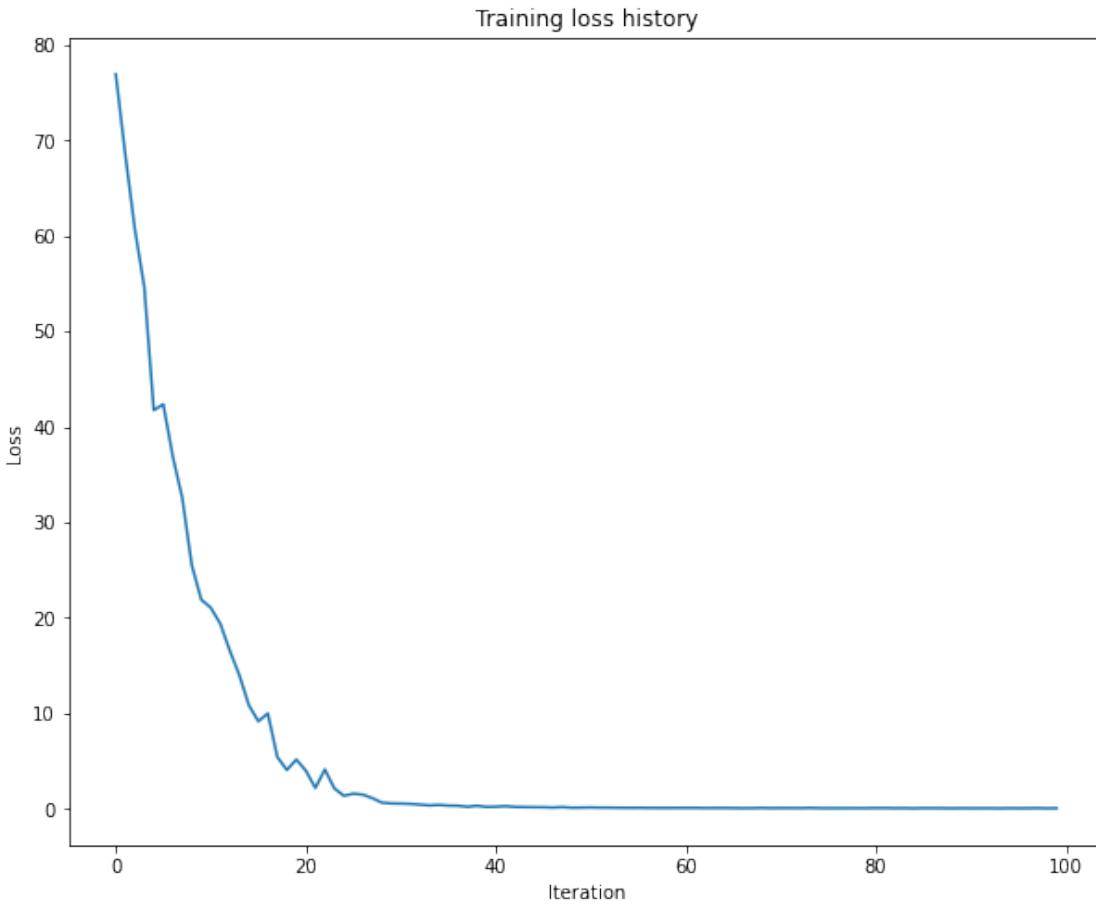
# Plot the training losses.
plt.plot(small_rnn_solver.loss_history)
plt.xlabel('Iteration')
plt.ylabel('Loss')
plt.title('Training loss history')
plt.show()

```

```

base dir /content/drive/My
Drive/cs231n/assignments/assignment3/cs231n/datasets/coco_captioning
(Iteration 1 / 100) loss: 76.913487
(Iteration 11 / 100) loss: 21.062729
(Iteration 21 / 100) loss: 4.016257
(Iteration 31 / 100) loss: 0.567182
(Iteration 41 / 100) loss: 0.239398
(Iteration 51 / 100) loss: 0.161974
(Iteration 61 / 100) loss: 0.111529
(Iteration 71 / 100) loss: 0.097580
(Iteration 81 / 100) loss: 0.099065
(Iteration 91 / 100) loss: 0.073967

```



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

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

Final loss: 0.08207126626105228

## 14 RNN Sampling at Test Time

Unlike classification models, image captioning models behave very differently at training time vs. 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, however, probably won't make sense.

```
[17]: # If you get an error, the URL just no longer exists, so don't worry!
# You can re-sample as many times as you want.
for split in ['train', 'val']:
```

```

minibatch = sample_coco_minibatch(small_data, split=split, batch_size=2)
gt_captions, features, urls = minibatch
gt_captions = decodeCaptions(gt_captions, data['idx_to_word'])

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

for gt_caption, sample_caption, url in zip(gt_captions, sample_captions, urls):
    img = imageFromUrl(url)
    # Skip missing URLs.
    if img is None: continue
    plt.imshow(img)
    plt.title('%s\n%s\nGT:%s' % (split, sample_caption, gt_caption))
    plt.axis('off')
    plt.show()

```

train

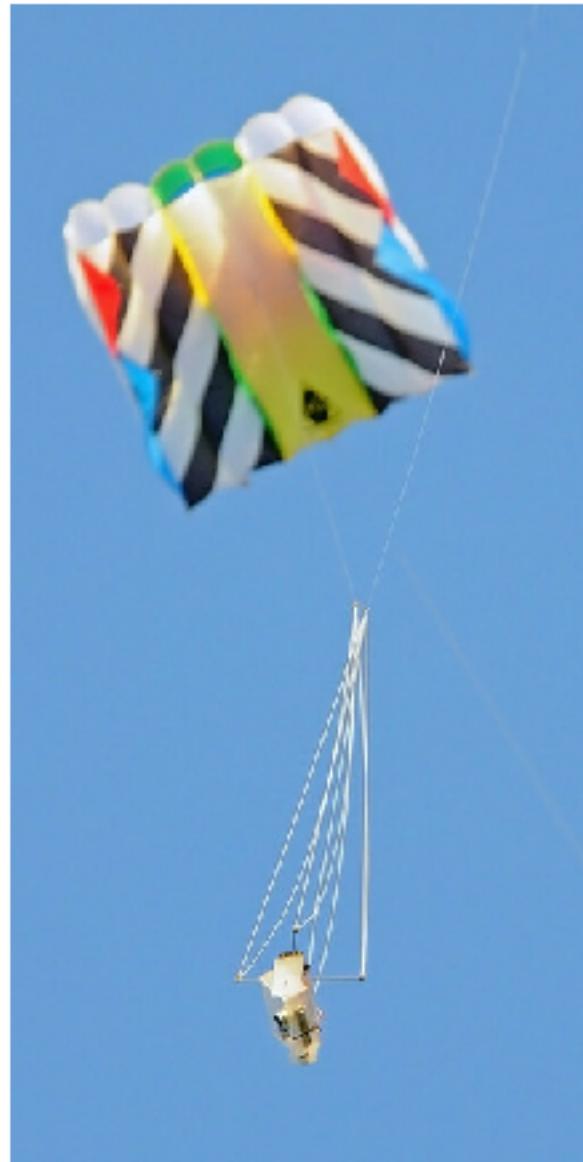
a boy sitting with <UNK> on with a donut in his hand <END>  
GT:<START> a boy sitting with <UNK> on with a donut in his hand <END>



train

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

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



val

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



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



## 15 Inline Question 1

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

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

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

**Your Answer:** The number of parameters of the character-level models are smaller. But the character-level models may produce meaningless outputs.

[17]:

# Transformer\_Captioning

February 6, 2022

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

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

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

# This downloads the COCO dataset to your Drive
# if it doesn't already exist.
%cd /content/drive/My\ Drive/$FOLDERNAME/cs231n/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME
```

Mounted at /content/drive  
/content/drive/My Drive/cs231n/assignments/assignment3/cs231n/datasets  
/content/drive/My Drive/cs231n/assignments/assignment3

## 1 Image Captioning with Transformers

You have now implemented a vanilla RNN and for the task of image captioning. In this notebook you will implement key pieces of a transformer decoder to accomplish the same task.

**NOTE:** This notebook will be primarily written in PyTorch rather than NumPy, unlike the RNN notebook.

```
[2]: # Setup cell.
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.transformer_layers import *
from cs231n.captioning_solver_transformer import CaptioningSolverTransformer
from cs231n.classifiers.transformer import CaptioningTransformer
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'

%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 COCO Dataset

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

```
[3]: # Load COCO data from disk into a dictionary.
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  /content/drive/My
Drive/cs231n/assignments/assignment3/cs231n/datasets/coco_captioning
trainCaptions <class 'numpy.ndarray'> (400135, 17) int32
trainImageIdxs <class 'numpy.ndarray'> (400135,) int32
valCaptions <class 'numpy.ndarray'> (195954, 17) int32
valImageIdxs <class 'numpy.ndarray'> (195954,) int32
trainFeatures <class 'numpy.ndarray'> (82783, 512) float32
valFeatures <class 'numpy.ndarray'> (40504, 512) float32
idx_to_word <class 'list'> 1004
word_to_idx <class 'dict'> 1004
trainUrls <class 'numpy.ndarray'> (82783,) <U63
valUrls <class 'numpy.ndarray'> (40504,) <U63

```

## 3 Transformer

As you have seen, RNNs are incredibly powerful but often slow to train. Further, RNNs struggle to encode long-range dependencies (though LSTMs are one way of mitigating the issue). In 2017, Vaswani et al introduced the Transformer in their paper "[Attention Is All You Need](#)" to a) introduce parallelism and b) allow models to learn long-range dependencies. The paper not only led to famous models like BERT and GPT in the natural language processing community, but also an explosion of interest across fields, including vision. While here we introduce the model in the context of image captioning, the idea of attention itself is much more general.

## 4 Transformer: Multi-Headed Attention

### 4.0.1 Dot-Product Attention

Recall that attention can be viewed as an operation on a query  $q \in \mathbb{R}^d$ , a set of value vectors  $\{v_1, \dots, v_n\}, v_i \in \mathbb{R}^d$ , and a set of key vectors  $\{k_1, \dots, k_n\}, k_i \in \mathbb{R}^d$ , specified as

$$c = \sum_{i=1}^n v_i \alpha_i \alpha_i^\top = \frac{\exp(k_i^\top q)}{\sum_{j=1}^n \exp(k_j^\top q)} \quad (1)$$

(2)

where  $\alpha_i$  are frequently called the "attention weights", and the output  $c \in \mathbb{R}^d$  is a correspondingly weighted average over the value vectors.

### 4.0.2 Self-Attention

In Transformers, we perform self-attention, which means that the values, keys and query are derived from the input  $X \in \mathbb{R}^{\ell \times d}$ , where  $\ell$  is our sequence length. Specifically, we learn parameter matrices  $V, K, Q \in \mathbb{R}^{d \times d}$  to map our input  $X$  as follows:

$$v_i = Vx_i \quad i \in \{1, \dots, \ell\} \quad (3)$$

$$k_i = Kx_i \quad i \in \{1, \dots, \ell\} \quad (4)$$

$$q_i = Qx_i \quad i \in \{1, \dots, \ell\} \quad (5)$$

### 4.0.3 Multi-Headed Scaled Dot-Product Attention

In the case of multi-headed attention, we learn a parameter matrix for each head, which gives the model more expressivity to attend to different parts of the input. Let  $h$  be number of heads, and  $Y_i$  be the attention output of head  $i$ . Thus we learn individual matrices  $Q_i, K_i$  and  $V_i$ . To keep our overall computation the same as the single-headed case, we choose  $Q_i \in \mathbb{R}^{d \times d/h}, K_i \in \mathbb{R}^{d \times d/h}$  and  $V_i \in \mathbb{R}^{d \times d/h}$ . Adding in a scaling term  $\frac{1}{\sqrt{d/h}}$  to our simple dot-product attention above, we have

$$Y_i = \text{softmax}\left(\frac{(XQ_i)(XK_i)^\top}{\sqrt{d/h}}\right)(XV_i) \quad (6)$$

where  $Y_i \in \mathbb{R}^{\ell \times d/h}$ , where  $\ell$  is our sequence length.

In our implementation, we then apply dropout here (though in practice it could be used at any step):

$$Y_i = \text{dropout}(Y_i)$$

Finally, then the output of the self-attention is a linear transformation of the concatenation of the heads:

$$Y = [Y_1; \dots; Y_h]A \quad (7)$$

were  $A \in \mathbb{R}^{d \times d}$  and  $[Y_1; \dots; Y_h] \in \mathbb{R}^{\ell \times d}$ .

Implement multi-headed scaled dot-product attention in the `MultiHeadAttention` class in the file `cs231n/transformer_layers.py`. The code below will check your implementation. The relative error should be less than `1e-3`.

[4]: `torch.manual_seed(231)`

```
# Choose dimensions such that they are all unique for easier debugging:
# Specifically, the following values correspond to N=1, H=2, T=3, E//H=4, and
# E=8.
batch_size = 1
sequence_length = 3
embed_dim = 8
attn = MultiHeadAttention(embed_dim, num_heads=2)

# Self-attention.
data = torch.randn(batch_size, sequence_length, embed_dim)
self_attn_output = attn(query=data, key=data, value=data)

# Masked self-attention.
mask = torch.randn(sequence_length, sequence_length) < 0.5
masked_self_attn_output = attn(query=data, key=data, value=data, attn_mask=mask)

# Attention using two inputs.
other_data = torch.randn(batch_size, sequence_length, embed_dim)
attn_output = attn(query=data, key=other_data, value=other_data)

expected_self_attn_output = np.asarray([
    [-0.2494,  0.1396,  0.4323, -0.2411, -0.1547,  0.2329, -0.1936,
     -0.1444],
    [-0.1997,  0.1746,  0.7377, -0.3549, -0.2657,  0.2693, -0.2541,
     -0.2476],
    [-0.0625,  0.1503,  0.7572, -0.3974, -0.1681,  0.2168, -0.2478,
     -0.3038]]))

expected_masked_self_attn_output = np.asarray([
    [-0.1347,  0.1934,  0.8628, -0.4903, -0.2614,  0.2798, -0.2586,
     -0.3019],
    [-0.1013,  0.3111,  0.5783, -0.3248, -0.3842,  0.1482, -0.3628,
```

```

        -0.1496],
       [-0.2071,  0.1669,  0.7097, -0.3152, -0.3136,  0.2520, -0.2774,
       -0.2208]]])

expected_attn_output = np.asarray([[[
[-0.1980,  0.4083,  0.1968, -0.3477,  0.0321,  0.4258, -0.8972,
-0.2744],
[-0.1603,  0.4155,  0.2295, -0.3485, -0.0341,  0.3929, -0.8248,
-0.2767],
[-0.0908,  0.4113,  0.3017, -0.3539, -0.1020,  0.3784, -0.7189,
-0.2912]]]

print('self_attn_output error: ', rel_error(expected_self_attn_output, self_attn_output.detach().numpy()))
print('masked_self_attn_output error: ', rel_error(expected_masked_self_attn_output, masked_self_attn_output.detach().numpy()))
print('attn_output error: ', rel_error(expected_attn_output, attn_output.detach().numpy()))

```

```

self_attn_output error:  0.0003775124598178026
masked_self_attn_output error:  0.0001526367643724865
attn_output error:  0.0003527921483788199

```

## 5 Positional Encoding

While transformers are able to easily attend to any part of their input, the attention mechanism has no concept of token order. However, for many tasks (especially natural language processing), relative token order is very important. To recover this, the authors add a positional encoding to the embeddings of individual word tokens.

Let us define a matrix  $P \in \mathbb{R}^{l \times d}$ , where  $P_{ij} =$

$$\begin{cases} \sin\left(i \cdot 10000^{-\frac{j}{d}}\right) & \text{if } j \text{ is even} \\ \cos\left(i \cdot 10000^{-\frac{(j-1)}{d}}\right) & \text{otherwise} \end{cases}$$

Rather than directly passing an input  $X \in \mathbb{R}^{l \times d}$  to our network, we instead pass  $X + P$ .

Implement this layer in `PositionalEncoding` in `cs231n/transformer_layers.py`. Once you are done, run the following to perform a simple test of your implementation. You should see errors on the order of  $e-3$  or less.

```

[5]: torch.manual_seed(231)

batch_size = 1
sequence_length = 2
embed_dim = 6
data = torch.randn(batch_size, sequence_length, embed_dim)

```

```

pos_encoder = PositionalEncoding(embed_dim)
output = pos_encoder(data)

expected_pe_output = np.asarray([[[[-1.2340,  1.1127,  1.6978, -0.0865, -0.0000,
    ↪ 1.2728],
    [ 0.9028, -0.4781,  0.5535,  0.8133,  1.2644,
    ↪ 1.7034]]]])
print('pe_output error: ', rel_error(expected_pe_output, output.detach().
    ↪numpy()))

```

pe\_output error: 0.00010421011374914356

## 6 Inline Question 1

Several key design decisions were made in designing the scaled dot product attention we introduced above. Explain why the following choices were beneficial: 1. Using multiple attention heads as opposed to one. 2. Dividing by  $\sqrt{d/h}$  before applying the softmax function. Recall that  $d$  is the feature dimension and  $h$  is the number of heads. 3. Adding a linear transformation to the output of the attention operation. What would happen if we were to stack attention operations directly?

Only one or two sentences per choice is necessary, but be sure to be specific in addressing what would have happened without each given implementation detail, why such a situation would be suboptimal, and how the proposed implementation improves the situation.

**Your Answer:** 1. Avoid overfitting. 2. Avoid gradient vanishing. 3. Without the linear transformation, the output would be the linear combination of input.

## 7 Overfit Transformer Captioning Model on Small Data

Run the following to overfit the Transformer-based captioning model on the same small dataset as we used for the RNN previously.

```
[6]: torch.manual_seed(231)
np.random.seed(231)

data = load_coco_data(max_train=50)

transformer = CaptioningTransformer(
    word_to_idx=data['word_to_idx'],
    input_dim=data['train_features'].shape[1],
    wordvec_dim=256,
    num_heads=2,
    num_layers=2,
    max_length=30
)
```

```

transformer_solver = CaptioningSolverTransformer(transformer, data,
    ↪idx_to_word=data['idx_to_word'],
    num_epochs=100,
    batch_size=25,
    learning_rate=0.001,
    verbose=True, print_every=10,
)
transformer_solver.train()

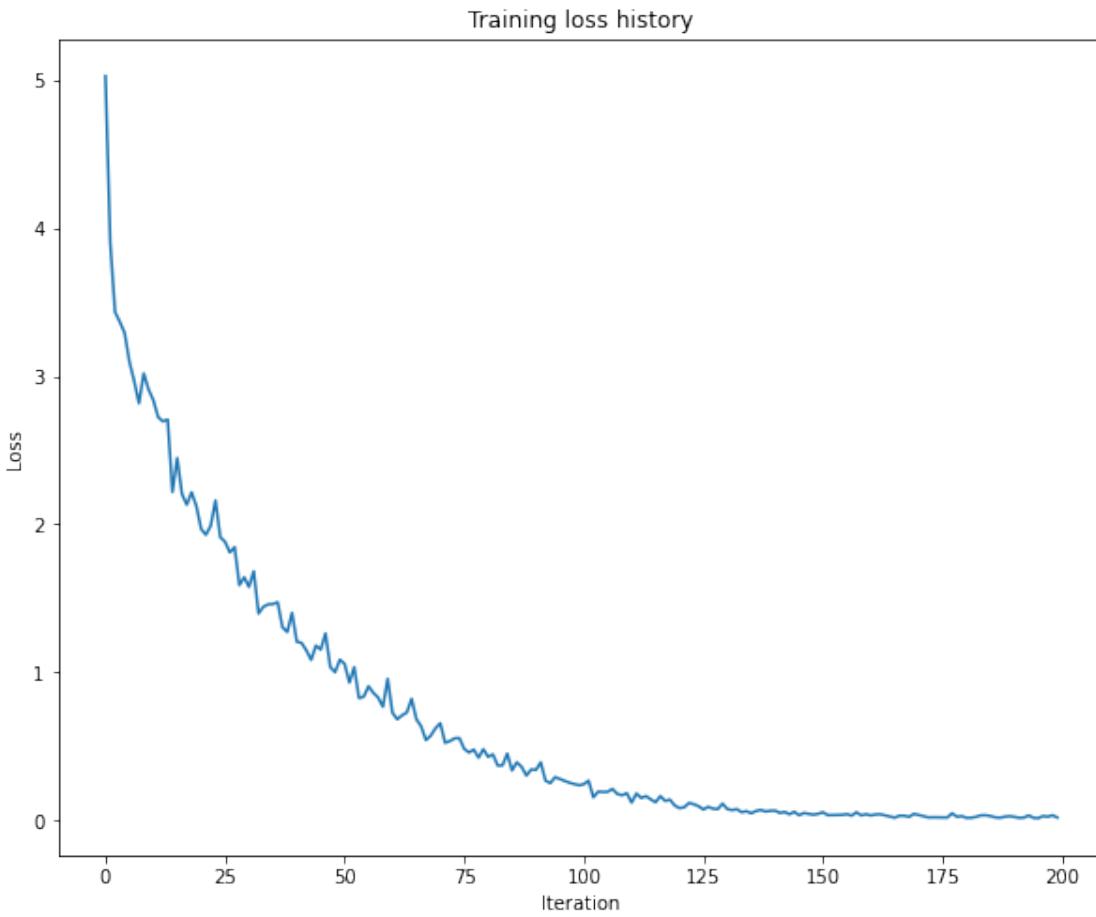
# Plot the training losses.
plt.plot(transformer_solver.loss_history)
plt.xlabel('Iteration')
plt.ylabel('Loss')
plt.title('Training loss history')
plt.show()

```

```

base dir /content/drive/My
Drive/cs231n/assignments/assignment3/cs231n/datasets/coco_captioning
(Iteration 1 / 200) loss: 5.023862
(Iteration 11 / 200) loss: 2.838898
(Iteration 21 / 200) loss: 1.968956
(Iteration 31 / 200) loss: 1.578086
(Iteration 41 / 200) loss: 1.208408
(Iteration 51 / 200) loss: 1.058599
(Iteration 61 / 200) loss: 0.729436
(Iteration 71 / 200) loss: 0.658647
(Iteration 81 / 200) loss: 0.431792
(Iteration 91 / 200) loss: 0.343782
(Iteration 101 / 200) loss: 0.247111
(Iteration 111 / 200) loss: 0.123138
(Iteration 121 / 200) loss: 0.085727
(Iteration 131 / 200) loss: 0.079453
(Iteration 141 / 200) loss: 0.068579
(Iteration 151 / 200) loss: 0.057809
(Iteration 161 / 200) loss: 0.036667
(Iteration 171 / 200) loss: 0.039795
(Iteration 181 / 200) loss: 0.020310
(Iteration 191 / 200) loss: 0.026851

```



Print final training loss. You should see a final loss of less than 0.03.

```
[7]: print('Final loss: ', transformer_solver.loss_history[-1])
```

Final loss: 0.022183504

## 8 Transformer Sampling at Test Time

The sampling code has been written for you. You can simply run the following to compare with the previous results with the RNN. As before the training results should be much better than the validation set results, given how little data we trained on.

```
[8]: # If you get an error, the URL just no longer exists, so don't worry!
# You can re-sample as many times as you want.
for split in ['train', 'val']:
    minibatch = sample_coco_minibatch(data, split=split, batch_size=2)
    gt_captions, features, urls = minibatch
    gt_captions = decode_captions(gt_captions, data['idx_to_word'])

    sample_captions = transformer.sample(features, max_length=30)
```

```

sample_captions = decode_captions(sample_captions, data['idx_to_word'])

for gt_caption, sample_caption, url in zip(gt_captions, sample_captions, urls):
    img = image_from_url(url)
    # Skip missing URLs.
    if img is None: continue
    plt.imshow(img)
    plt.title('%s\n%s\nGT:%s' % (split, sample_caption, gt_caption))
    plt.axis('off')
    plt.show()

```

URL Error: Gone [http://farm1.staticflickr.com/202/487987371\\_489a65d670\\_z.jpg](http://farm1.staticflickr.com/202/487987371_489a65d670_z.jpg)

train  
 a <UNK> decorated living room with a big tv in it <END>  
 GT:<START> a <UNK> decorated living room with a big tv in it <END>



val

a large dog with a stuffed and red front <END>

GT:<START> a bedroom with a bed desk and <UNK> <UNK> <END>



val

a man is jumping in a office his food while <END>  
GT:<START> a group of people <UNK> outside by a wall <END>



[8]:

# Network\_Visualization

February 6, 2022

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

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

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

# This downloads the COCO dataset to your Drive
# if it doesn't already exist.
%cd /content/drive/My\ Drive/$FOLDERNAME/cs231n/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME
```

```
Mounted at /content/drive
/content/drive/My\ Drive/cs231n/assignments/assignment3/cs231n/datasets
/content/drive/My\ Drive/cs231n/assignments/assignment3
```

## 1 Network Visualization

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 CNN 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 we will 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.

We will explore three techniques for image generation.

**Saliency Maps.** We can use saliency maps to tell which part of the image influenced the classification decision made by the network.

**Fooling Images.** We can perturb an input image so that it appears the same to humans but will be misclassified by the pretrained network.

**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.

```
[2]: # Setup cell.  
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  
from cs231n.net_visualization_pytorch import *  
  
%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'  
  
%load_ext autoreload  
%autoreload 2
```

## 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

```
[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
```

Downloading: "https://download.pytorch.org/models/squeezezenet1\_1-b8a52dc0.pth" to  
/root/.cache/torch/hub/checkpoints/squeezezenet1\_1-b8a52dc0.pth

0% | 0.00/4.73M [00:00<?, ?B/s]

## 2.1 Loading ImageNet Validation Images

We have provided a few example images from the validation set of the ImageNet ILSVRC 2012 Classification dataset. 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.

```
[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 an numpy array of shape  $(N, C)$  and  $y$  is a numpy array of shape  $(N,)$  containing integers  $0 \leq y[i] < C$ , then  $s[\text{np.arange}(N), y]$  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](#).

```
[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([[-0.7979, -0.5358,  0.4669,  0.0496, -0.5996],
       [-1.3072,  0.1847,  0.2190, -0.4592,  0.4198],
       [-1.5751, -0.7156,  0.5860, -0.2349, -0.4309],
       [ 1.6461,  1.3617,  0.0345,  0.9657, -1.5401]])
tensor([1, 2, 1, 3])
tensor([-0.5358,  0.2190, -0.7156,  0.9657])
```

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

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:

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

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

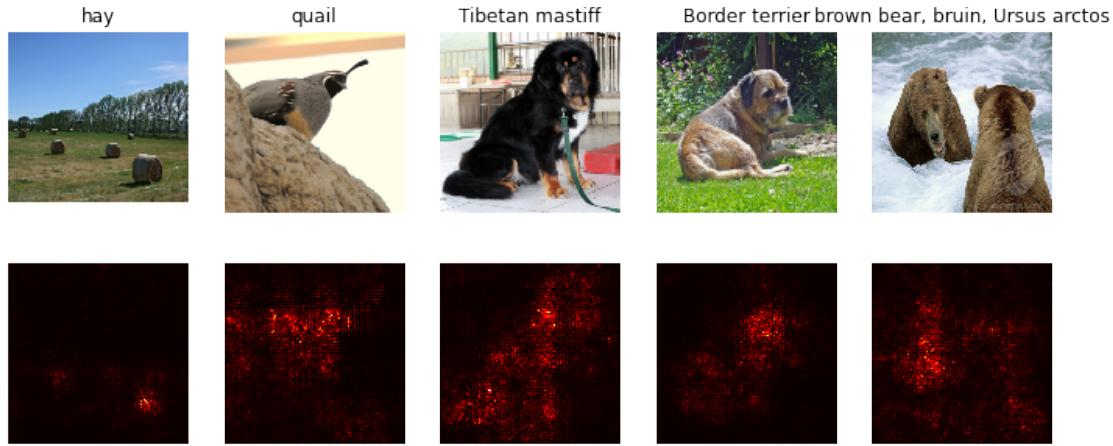
        # Convert the saliency map from Torch Tensor to numpy array and show images
        # and saliency maps together.
        saliency = saliency.numpy()
        N = X.shape[0]
        for i in range(N):
            plt.subplot(2, N, i + 1)
```

```

plt.imshow(X[i])
plt.axis('off')
plt.title(class_names[y[i]])
plt.subplot(2, N, N + i + 1)
plt.imshow(saliency[i], cmap=plt.cm.hot)
plt.axis('off')
plt.gcf().set_size_inches(12, 5)
plt.show()

show_saliency_maps(X, y)

```



## 4 Inline Question 1

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. There are only positive values in the saliency map.

## 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.

```
[7]: 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!'
```

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

```
[8]: 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:

```
[9]: 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 requiresgradient.

```
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  
 $ox, oy = \text{random.randint}(0, max\_jitter), \text{random.randint}(0, max\_jitter)$   
 $\text{img}.\text{data}.\text{copy\_}(jitter(\text{img}.\text{data}, ox, oy))$   
 $\text{class\_visualization\_update\_step}(\text{img}, \text{model}, \text{target\_y}, l2\_reg,$

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))
```

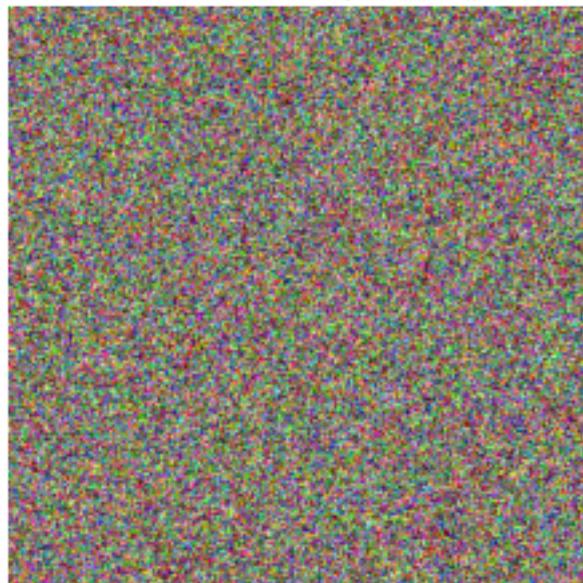
```
plt.gcf().set_size_inches(4, 4)
plt.axis('off')
plt.show()

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

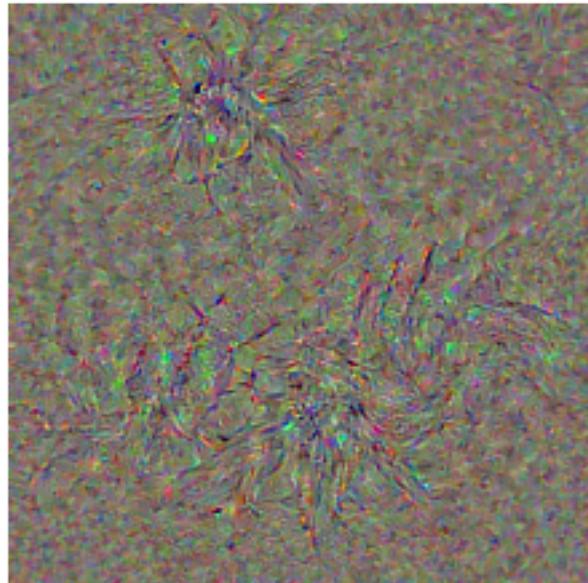
```
[10]: dtype = torch.FloatTensor
model.type(dtype)

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

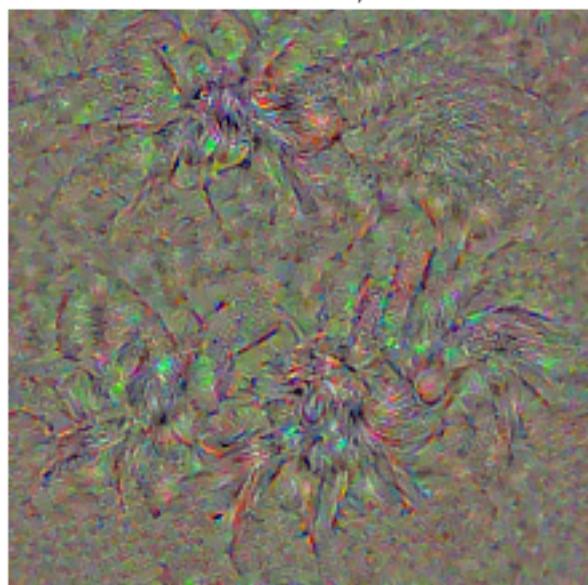
tarantula  
Iteration 1 / 100



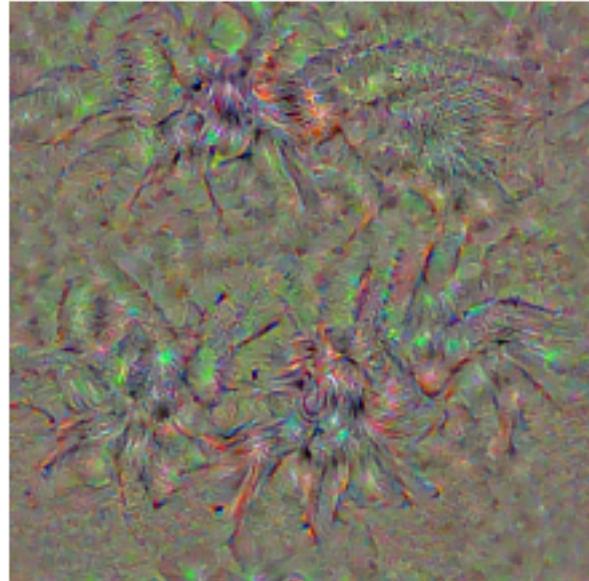
tarantula  
Iteration 25 / 100



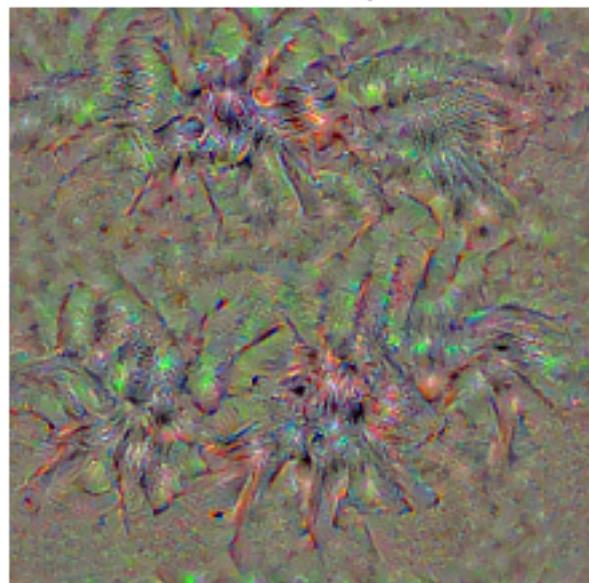
tarantula  
Iteration 50 / 100



tarantula  
Iteration 75 / 100



tarantula  
Iteration 100 / 100

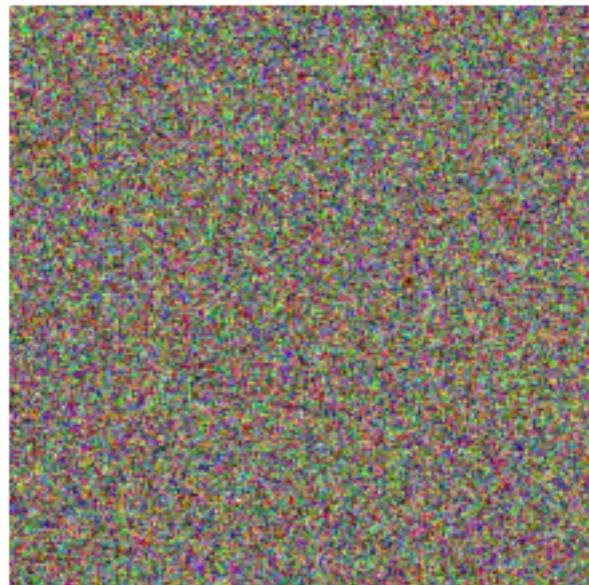


Try out your class visualization on other classes! You should also feel free to play with various hyperparameters to try and improve the quality of the generated image, but this is not required.

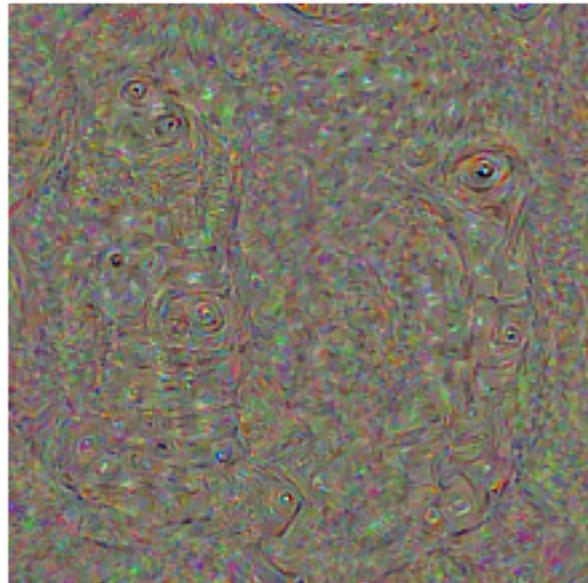
```
[11]: # target_y = 78 # Tick
# target_y = 187 # Yorkshire Terrier
# target_y = 683 # Oboe
# target_y = 366 # Gorilla
# target_y = 604 # Hourglass
target_y = np.random.randint(1000)
print(class_names[target_y])
X = create_class_visualization(target_y, model, dtype)
```

breastplate, aegis, egis

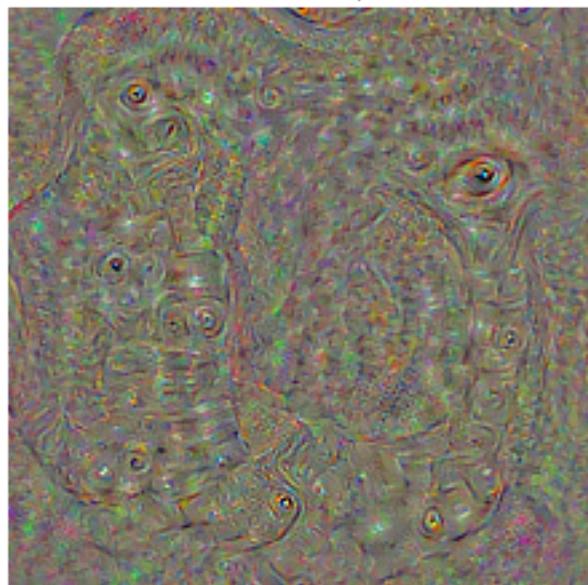
breastplate, aegis, egis  
Iteration 1 / 100



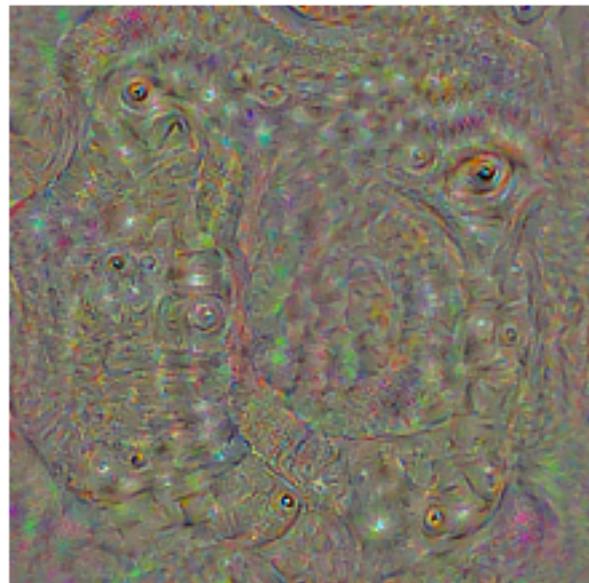
breastplate, aegis, egis  
Iteration 25 / 100



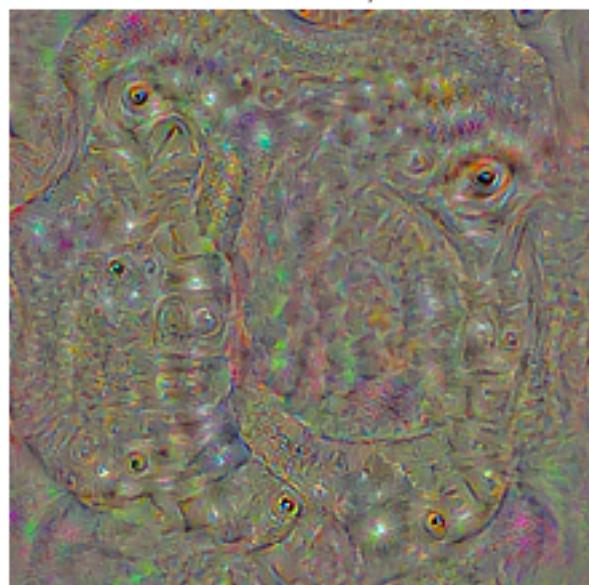
breastplate, aegis, egis  
Iteration 50 / 100



breastplate, aegis, egis  
Iteration 75 / 100



breastplate, aegis, egis  
Iteration 100 / 100



[11]:

# Generative\_Adversarial\_Networks

February 6, 2022

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

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

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

# This downloads the COCO dataset to your Drive
# if it doesn't already exist.
%cd /content/drive/My\ Drive/$FOLDERNAME/cs231n/datasets/
!bash get_datasets.sh
%cd /content/drive/My\ Drive/$FOLDERNAME
```

```
Mounted at /content/drive
/content/drive/My Drive/cs231n/assignments/assignment3/cs231n/datasets
/content/drive/My Drive/cs231n/assignments/assignment3
```

## 0.1 Using GPU

Go to Runtime > Change runtime type and set Hardware accelerator to GPU. This will reset Colab. **Rerun the top cell to mount your Drive again.**

## 1 Generative Adversarial Networks (GANs)

So far in CS 231N, 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:

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

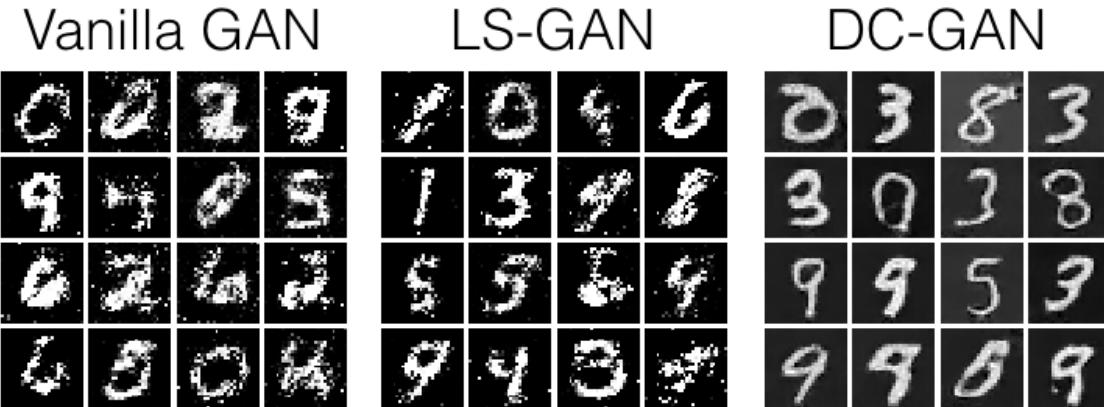
2. Update the discriminator ( $D$ ), to maximize the probability of the discriminator making the correct choice on real and generated data:

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

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:

```
[2]: # Run this cell to see sample outputs.
from IPython.display import Image
Image('images/gan_outputs_pytorch.png')
```

[2]:



```
[3]: # Setup cell.
import numpy as np
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 matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
from cs231n.gan_pytorch import preprocess_img, deprocess_img, rel_error, □
→count_params, ChunkSampler

%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'

%load_ext autoreload
%autoreload 2

def show_images(images):
    images = np.reshape(images, [images.shape[0], -1]) # Images reshape to □
→(batch_size, D).
    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([sqrtimg,sqrtimg]))
return

answers = dict(np.load('gan-checks.npz'))
dtype = torch.cuda.FloatTensor if torch.cuda.is_available() else torch.
→FloatTensor

```

## 1.1 Dataset

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

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

```
[4]: 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)

```

Downloading http://yann.lecun.com/exdb/mnist/train-images-idx3-ubyte.gz  
 Downloading http://yann.lecun.com/exdb/mnist/train-images-idx3-ubyte.gz to  
 ./cs231n/datasets/MNIST\_data/MNIST/raw/train-images-idx3-ubyte.gz  
 0%| | 0/9912422 [00:00<?, ?it/s]

Extracting ./cs231n/datasets/MNIST\_data/MNIST/raw/train-images-idx3-ubyte.gz to  
 ./cs231n/datasets/MNIST\_data/MNIST/raw

Downloading http://yann.lecun.com/exdb/mnist/train-labels-idx1-ubyte.gz  
 Downloading http://yann.lecun.com/exdb/mnist/train-labels-idx1-ubyte.gz to  
 ./cs231n/datasets/MNIST\_data/MNIST/raw/train-labels-idx1-ubyte.gz  
 0%| | 0/28881 [00:00<?, ?it/s]

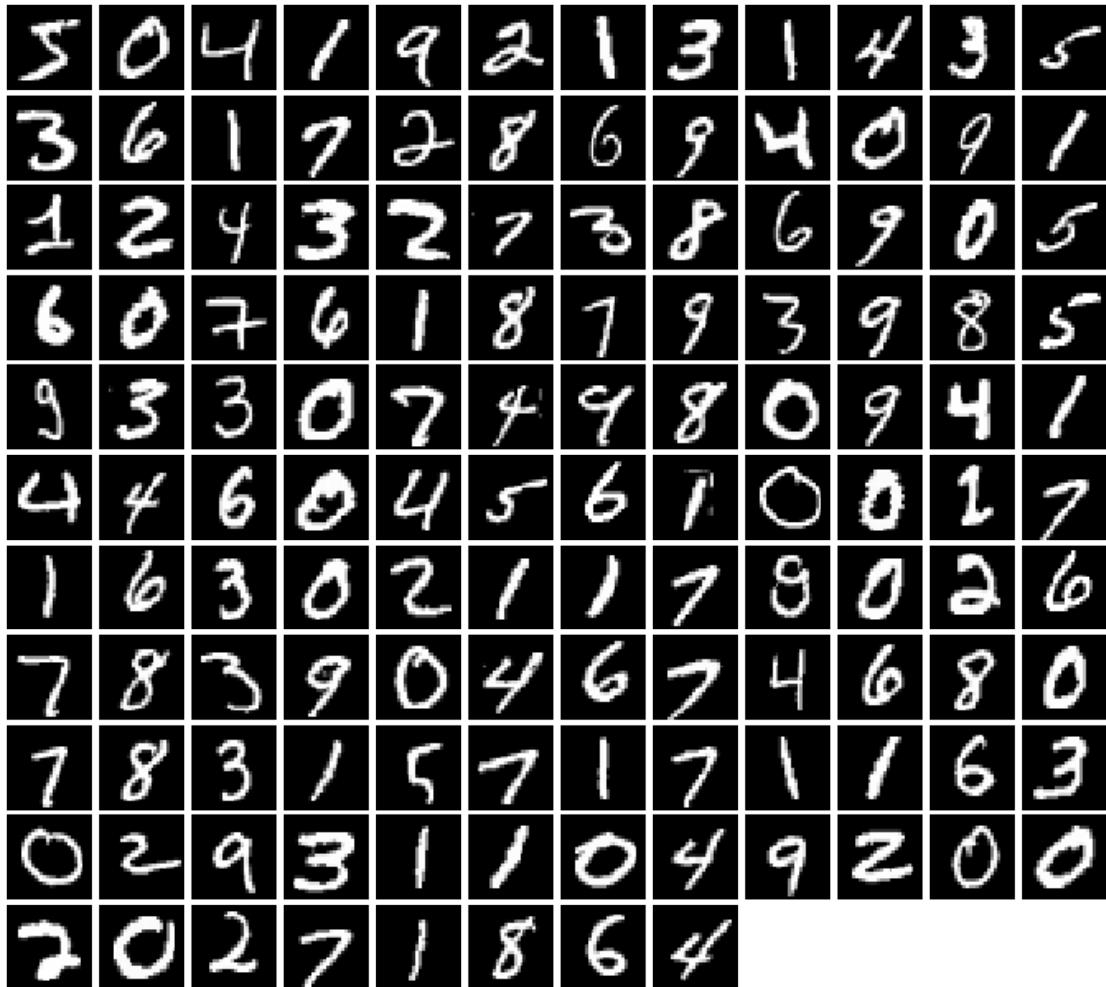
Extracting ./cs231n/datasets/MNIST\_data/MNIST/raw/train-labels-idx1-ubyte.gz to  
 ./cs231n/datasets/MNIST\_data/MNIST/raw

Downloading http://yann.lecun.com/exdb/mnist/t10k-images-idx3-ubyte.gz  
 Downloading http://yann.lecun.com/exdb/mnist/t10k-images-idx3-ubyte.gz to  
 ./cs231n/datasets/MNIST\_data/MNIST/raw/t10k-images-idx3-ubyte.gz  
 0%| | 0/1648877 [00:00<?, ?it/s]

Extracting ./cs231n/datasets/MNIST\_data/MNIST/raw/t10k-images-idx3-ubyte.gz to  
 ./cs231n/datasets/MNIST\_data/MNIST/raw

Downloading http://yann.lecun.com/exdb/mnist/t10k-labels-idx1-ubyte.gz  
 Downloading http://yann.lecun.com/exdb/mnist/t10k-labels-idx1-ubyte.gz to  
 ./cs231n/datasets/MNIST\_data/MNIST/raw/t10k-labels-idx1-ubyte.gz  
 0%| | 0/4542 [00:00<?, ?it/s]

```
Extracting ./cs231n/datasets/MNIST_data/MNIST/raw/t10k-labels-idx1-ubyte.gz to
./cs231n/datasets/MNIST_data/MNIST/raw
```



## 1.2 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:

```
[5]: 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!

### 1.3 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.

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

## 2 Discriminator

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

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

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

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

Implement discriminator in `cs231n/gan_pytorch.py`

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

[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:

```
[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.

```
[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

```
[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, beta1=0.5, 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 it for your own understanding.

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

# 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_train
)
```

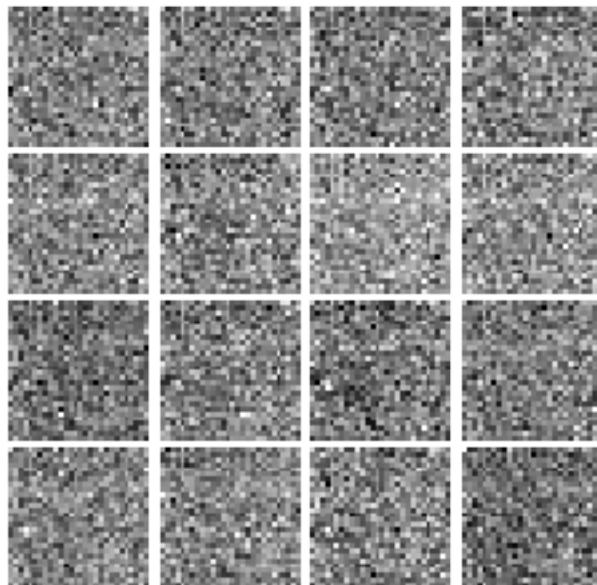
Iter: 0, D: 1.328, G:0.7202  
 Iter: 250, D: 1.092, G:2.223  
 Iter: 500, D: 1.164, G:1.465  
 Iter: 750, D: 1.554, G:0.6274  
 Iter: 1000, D: 1.194, G:0.8218  
 Iter: 1250, D: 1.098, G:1.088  
 Iter: 1500, D: 1.278, G:1.225  
 Iter: 1750, D: 1.244, G:0.976  
 Iter: 2000, D: 1.368, G:1.263  
 Iter: 2250, D: 1.258, G:0.8057  
 Iter: 2500, D: 1.34, G:0.9299  
 Iter: 2750, D: 1.393, G:0.8777  
 Iter: 3000, D: 1.268, G:1.014  
 Iter: 3250, D: 1.346, G:0.8724  
 Iter: 3500, D: 1.298, G:0.7736  
 Iter: 3750, D: 1.365, G:0.7962

Run the cell below to show the generated images.

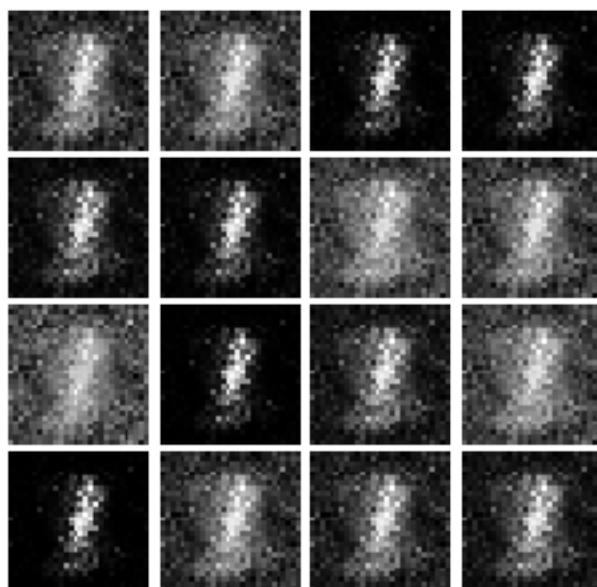
```
[12]: numIter = 0
for img in images:
    print("Iter: {}".format(numIter))
    show_images(img)
    plt.show()
```

```
numIter += 250  
print()
```

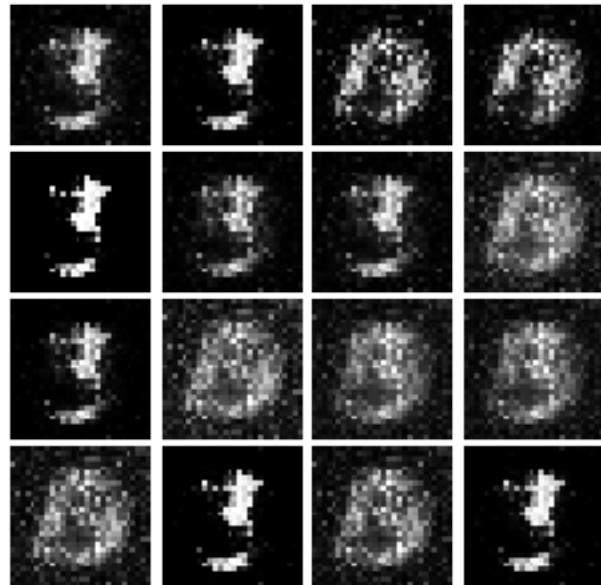
Iter: 0



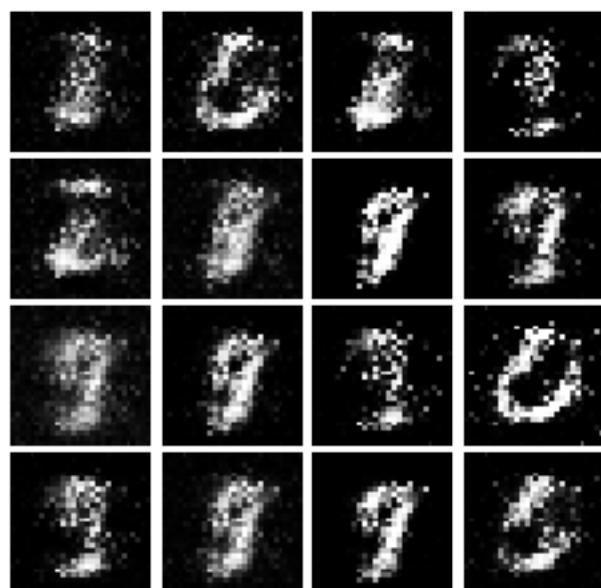
Iter: 250



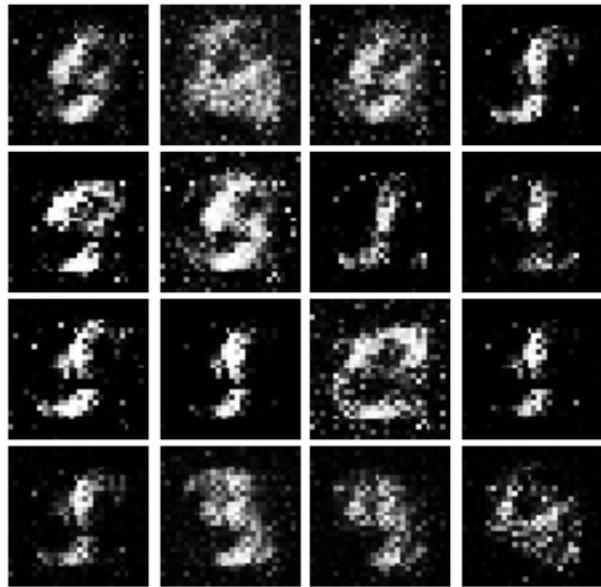
Iter: 500



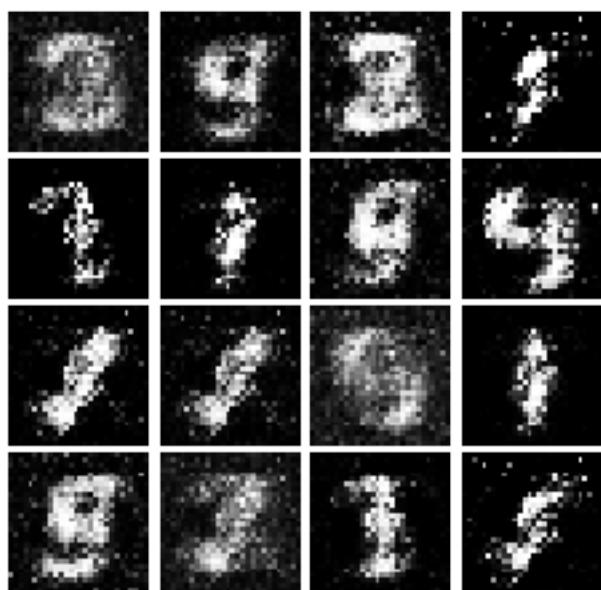
Iter: 750



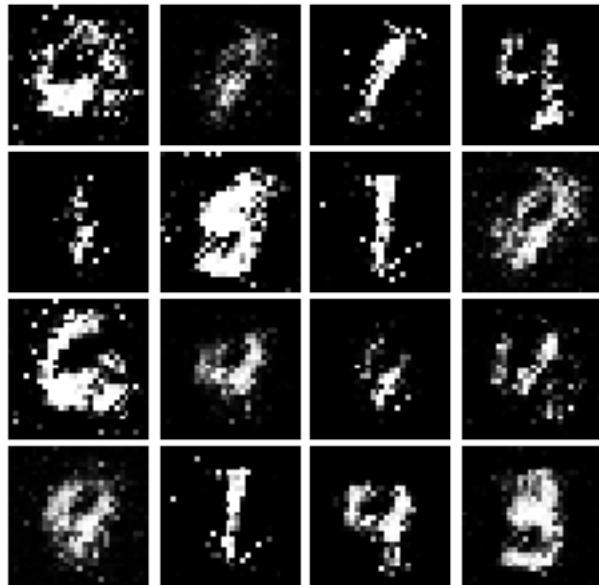
Iter: 1000



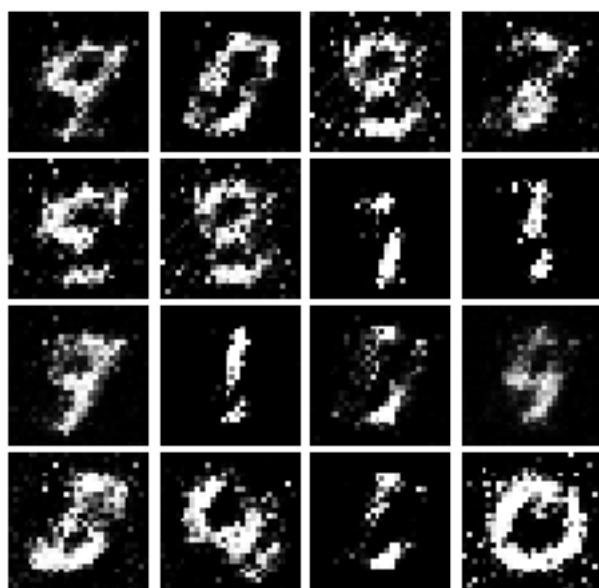
Iter: 1250



Iter: 1500



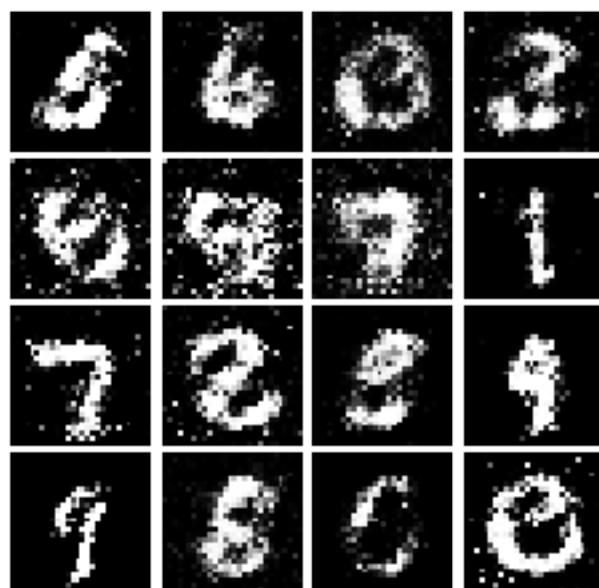
Iter: 1750



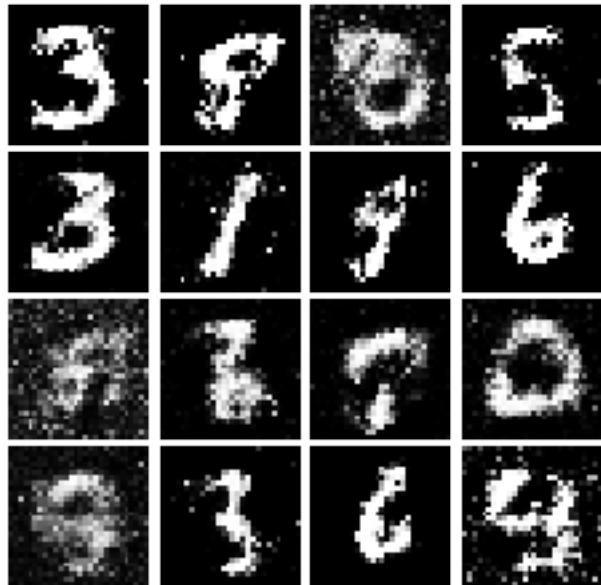
Iter: 2000



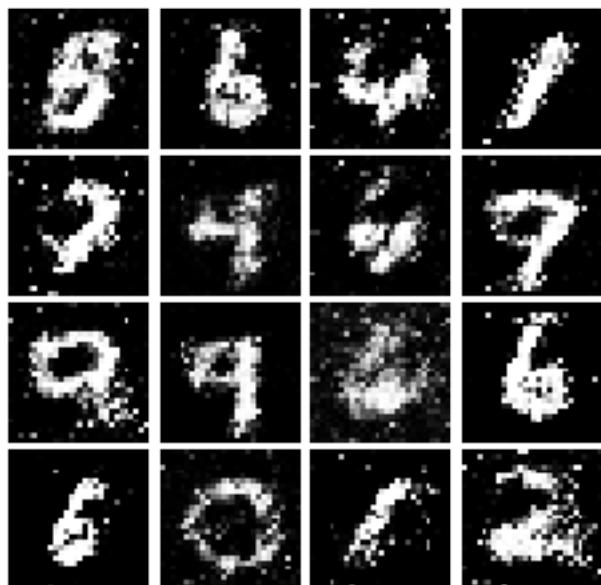
Iter: 2250



Iter: 2500



Iter: 2750



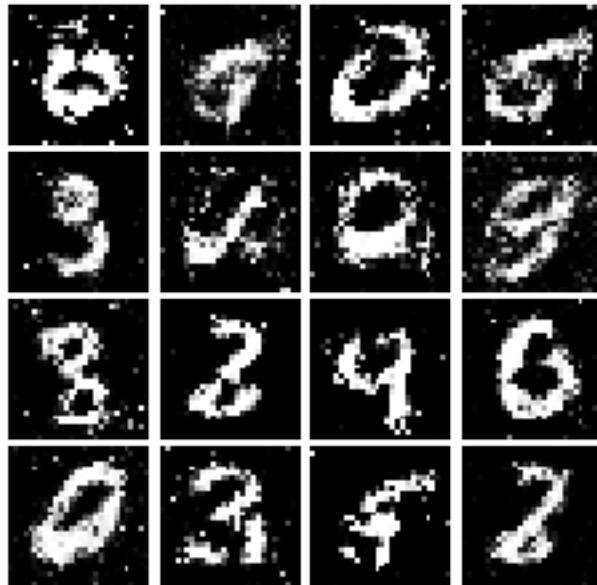
Iter: 3000



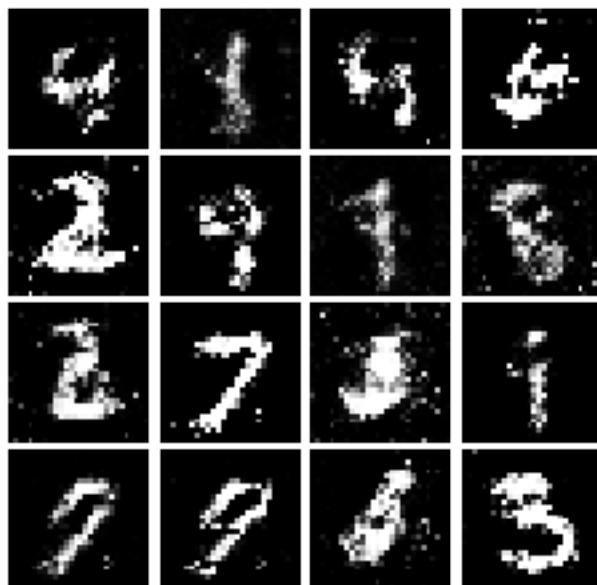
Iter: 3250



Iter: 3500



Iter: 3750

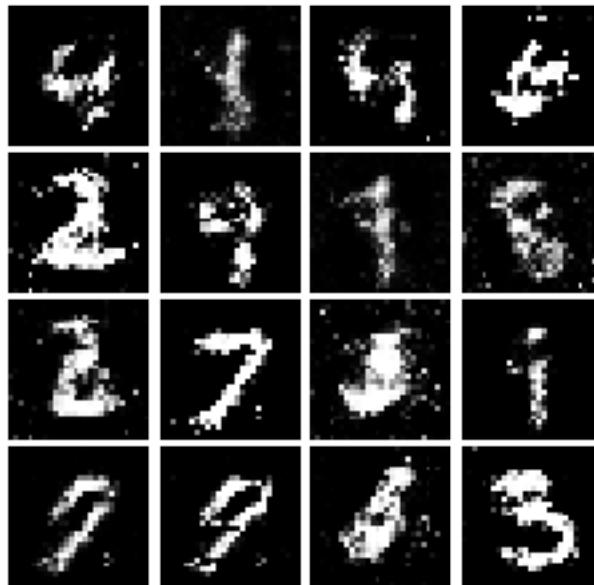


## 6.1 Inline Question 1

What does your final vanilla GAN image look like?

```
[13]: # This output is your answer.  
print("Vanilla GAN final image:")  
show_images(images[-1])  
plt.show()
```

Vanilla GAN final 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:

```
[14]: 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: 1.53171e-08
Maximum error in g_loss: 2.7837e-09
```

Run the following cell to train your model!

```
[15]: 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_generator_loss,
    loader_train
)
```

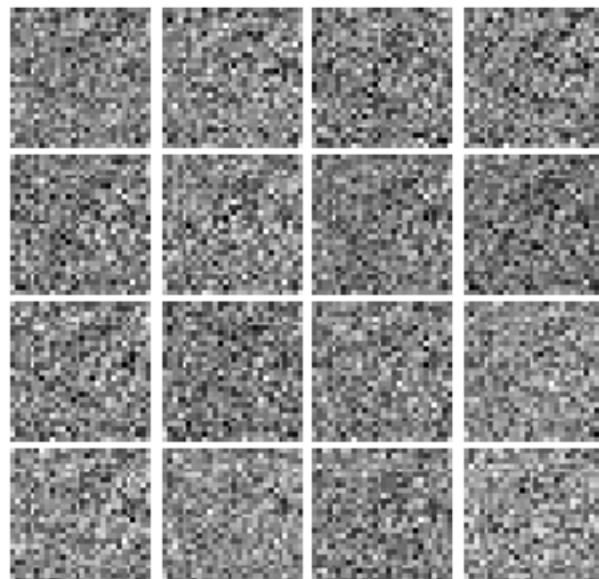
```
Iter: 0, D: 0.5689, G:0.51
Iter: 250, D: 0.2428, G:0.369
```

```
Iter: 500, D: 0.1353, G:0.3779
Iter: 750, D: 0.1522, G:0.2877
Iter: 1000, D: 0.1548, G:0.3057
Iter: 1250, D: 0.1749, G:0.2163
Iter: 1500, D: 0.2096, G:0.2132
Iter: 1750, D: 0.1978, G:0.246
Iter: 2000, D: 0.2003, G:0.2131
Iter: 2250, D: 0.2342, G:0.1764
Iter: 2500, D: 0.1925, G:0.1984
Iter: 2750, D: 0.211, G:0.1718
Iter: 3000, D: 0.2265, G:0.1615
Iter: 3250, D: 0.2109, G:0.1668
Iter: 3500, D: 0.2225, G:0.1627
Iter: 3750, D: 0.2217, G:0.1372
```

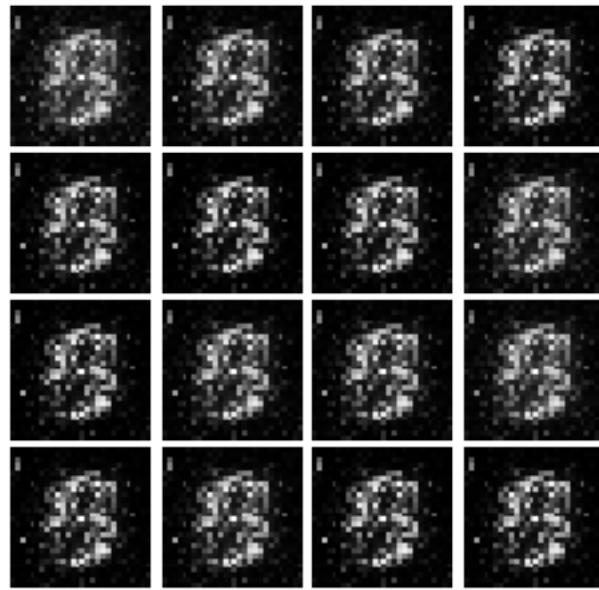
Run the cell below to show generated images.

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

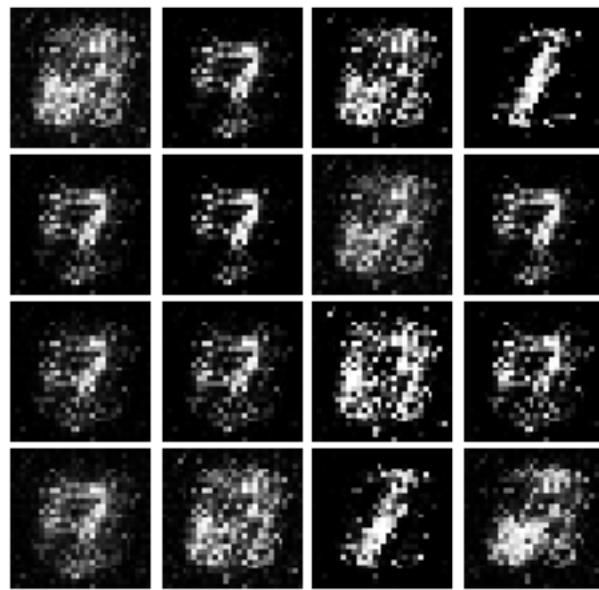
Iter: 0



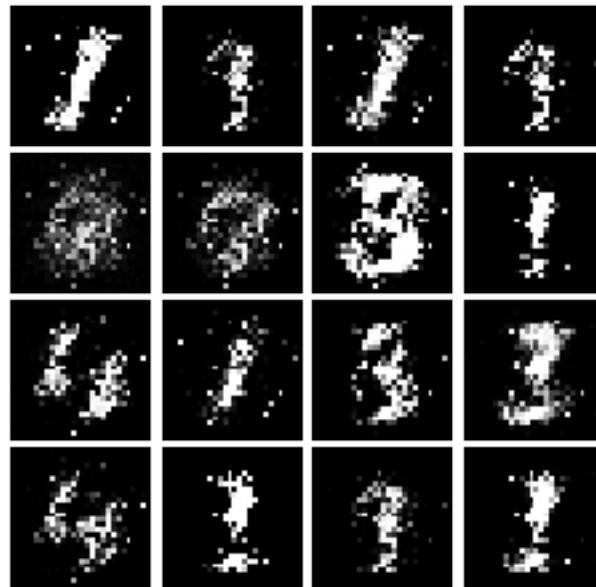
Iter: 250



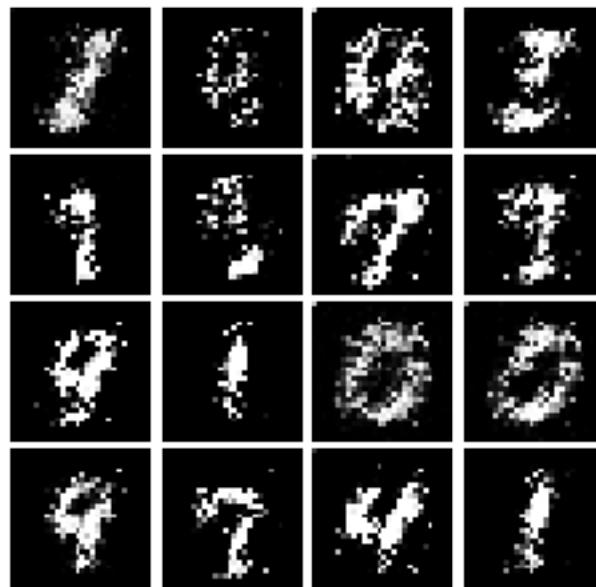
Iter: 500



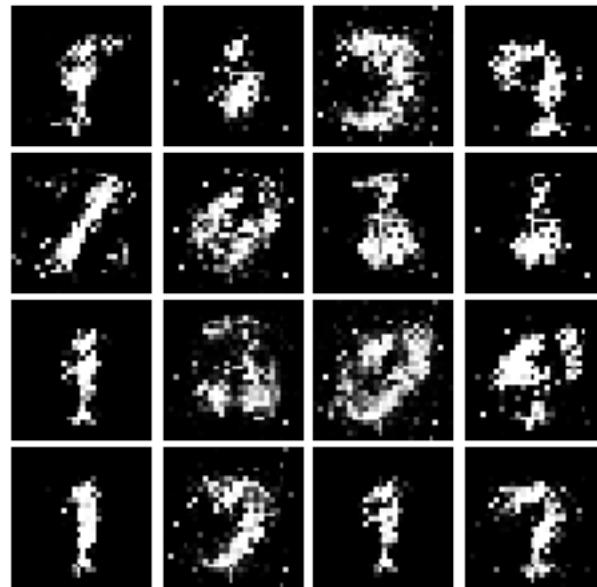
Iter: 750



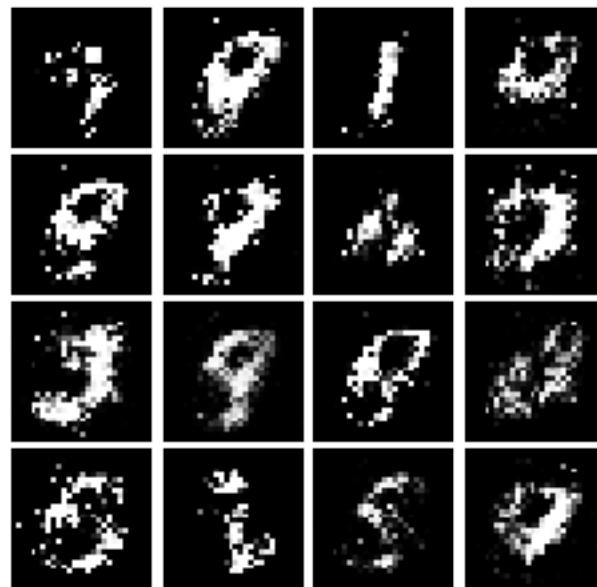
Iter: 1000



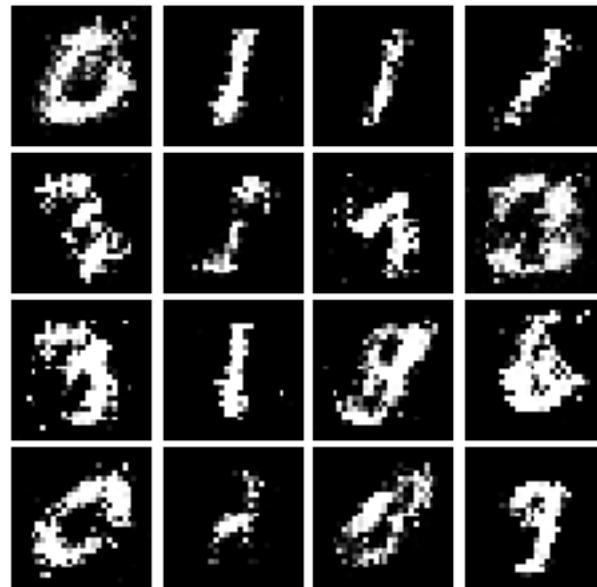
Iter: 1250



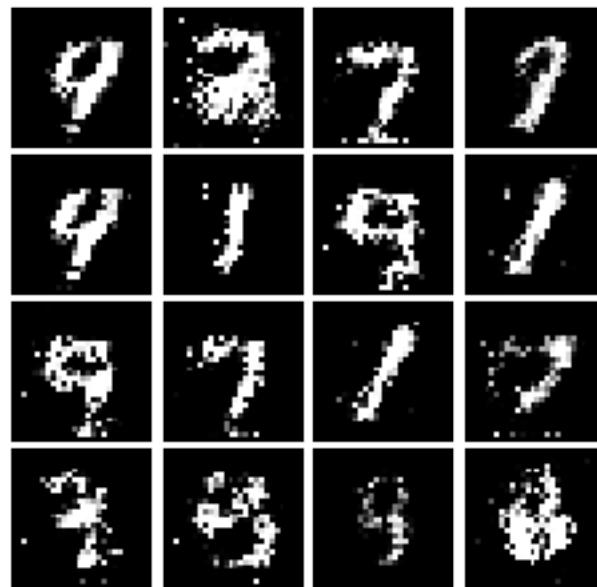
Iter: 1500



Iter: 1750



Iter: 2000



Iter: 2250



Iter: 2500



Iter: 2750



Iter: 3000



Iter: 3250



Iter: 3500



Iter: 3750



## 7.1 Inline Question 2

What does your final LSGAN image look like?

```
[17]: # This output is your answer.  
print("LSGAN final image:")  
show_images(images[-1])  
plt.show()
```

LSGAN final 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`

```
[18]: 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:

```
[19]: 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\u202a
→architecture.')
    else:
        print('Correct number of parameters in generator.')

test_dc_classifier()
```

Correct number of parameters in generator.

**Generator** For the generator, we will copy the architecture exactly from the [InfoGAN paper](#). See Appendix C.1 MNIST. See the documentation for [tf.nn.conv2d\\_transpose](#). We are always "training" in GAN mode. \* Fully connected with output size 1024 \* ReLU \* BatchNorm \* Fully connected with output size  $7 \times 7 \times 128$  \* ReLU \* BatchNorm \* Reshape into Image Tensor of shape  $7, 7, 128$  \* Conv2D $^T$  (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`

```
[20]: 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()
```

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

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

```
[21]: 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\u202a
→architecture.')
    else:
        print('Correct number of parameters in generator.')

test_dc_generator()
```

Correct number of parameters in generator.

```
[22]: 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)

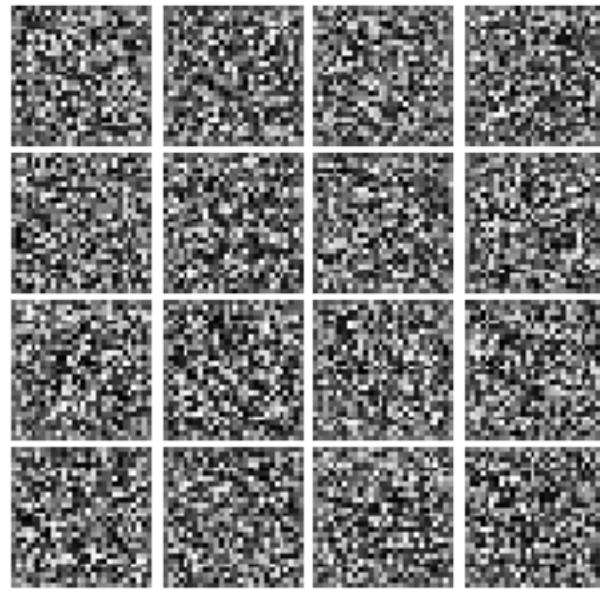
images = run_a_gan(
    D_DC,
    G_DC,
    D_DC_solver,
    G_DC_solver,
    discriminator_loss,
    generator_loss,
    loader_train,
    num_epochs=5
)
```

Iter: 0, D: 1.448, G:1.464  
 Iter: 250, D: 1.327, G:0.626  
 Iter: 500, D: 1.214, G:1.128  
 Iter: 750, D: 1.109, G:0.9492  
 Iter: 1000, D: 1.206, G:1.049  
 Iter: 1250, D: 1.274, G:0.8732  
 Iter: 1500, D: 1.106, G:0.8999  
 Iter: 1750, D: 1.009, G:1.207

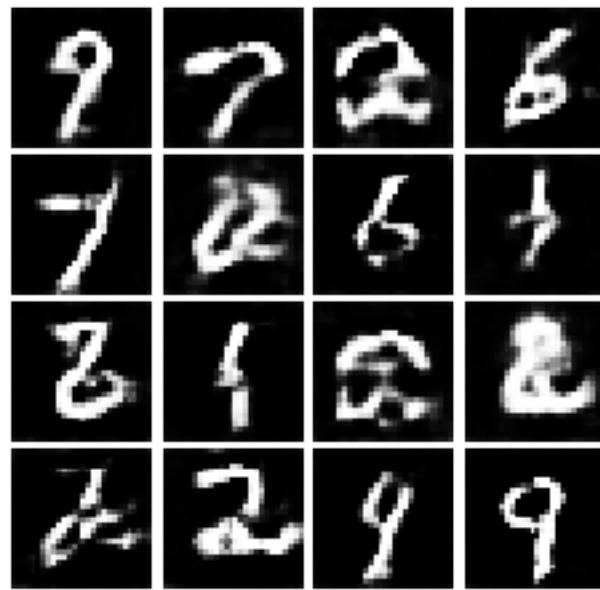
Run the cell below to show generated images.

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

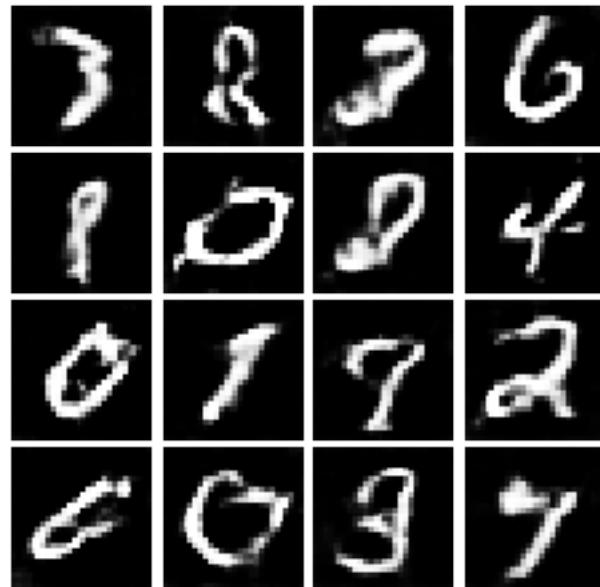
Iter: 0



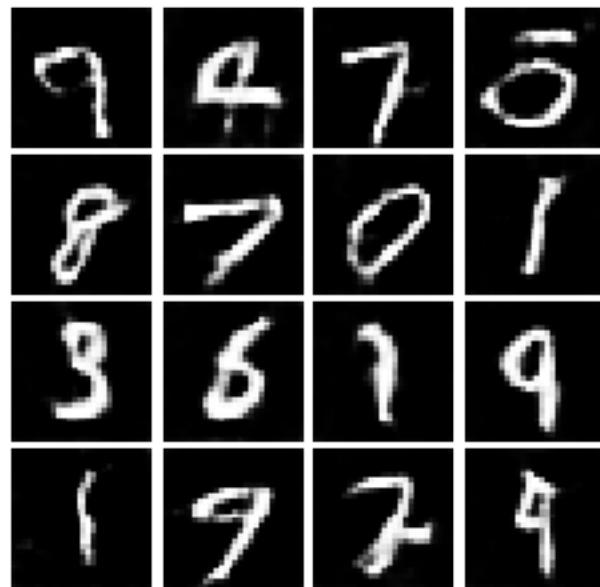
Iter: 250



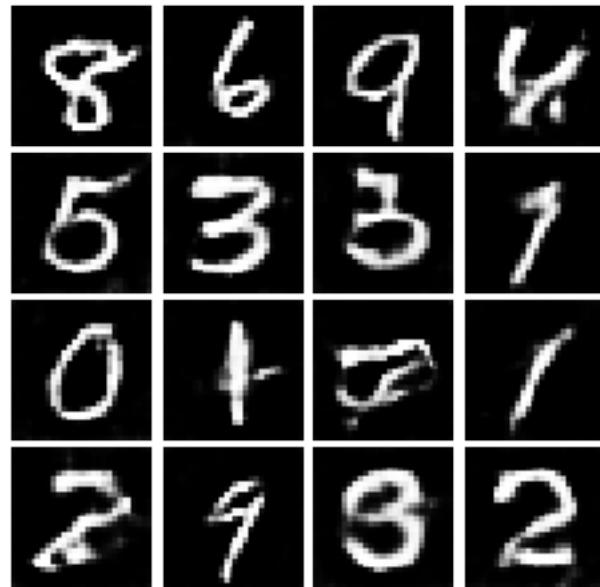
Iter: 500



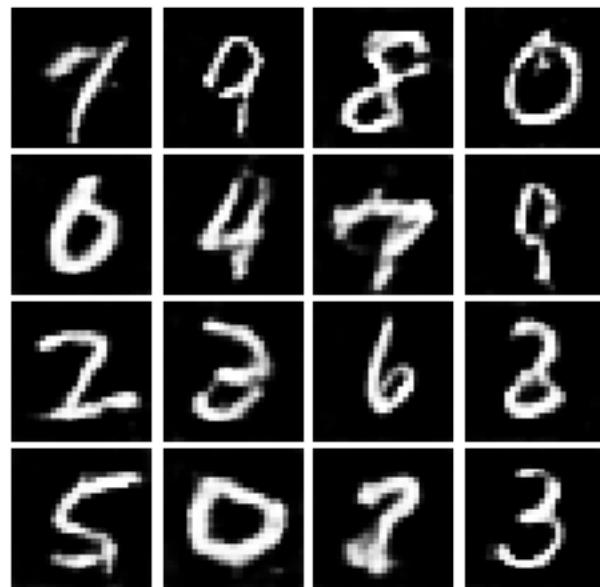
Iter: 750



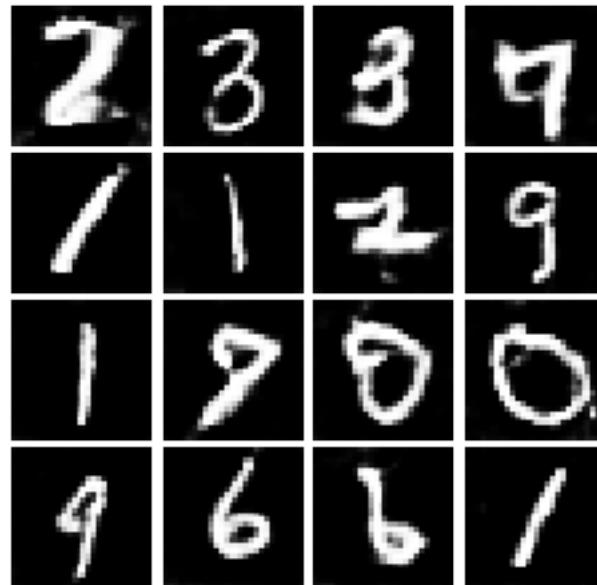
Iter: 1000



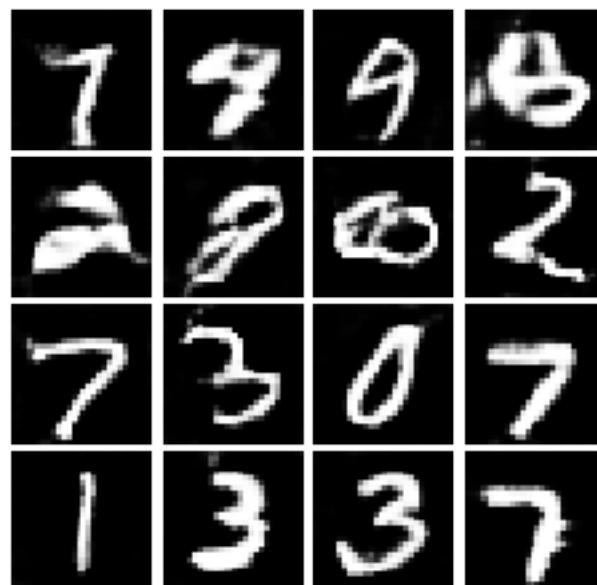
Iter: 1250



Iter: 1500



Iter: 1750

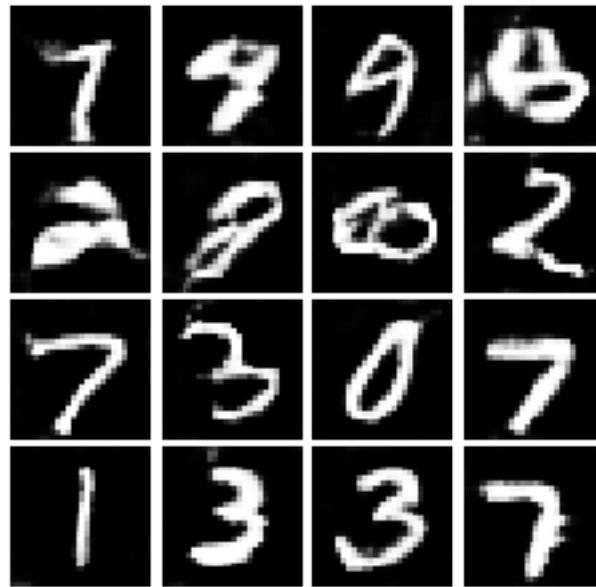


### 8.1 Inline Question 3

What does your final DCGAN image look like?

```
[24]: # This output is your answer.
print("DCGAN final image:")
show_images(images[-1])
plt.show()
```

DCGAN final image:



## 8.2 Inline Question 4

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.

Briefly 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.

### 8.2.1 Your answer:

The value of the function is going up and down.

$y_0$	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$
1	2	2	1	1	-1	-1

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

### 8.3 Inline Question 5

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

#### 8.3.1 Your answer:

No. As for a certain point, the gradient about  $x$  is  $y$  and the gradient about  $y$  is  $x$ , so  $x$  can always goes down and  $y$  can always goes up.

### 8.4 Inline Question 6

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.

#### 8.4.1 Your answer:

No. There may be a situation that the generator and the discriminator are both bad so that even the generator generated a bad picture, the discriminator still regard the picture as a real picture.

[24]: