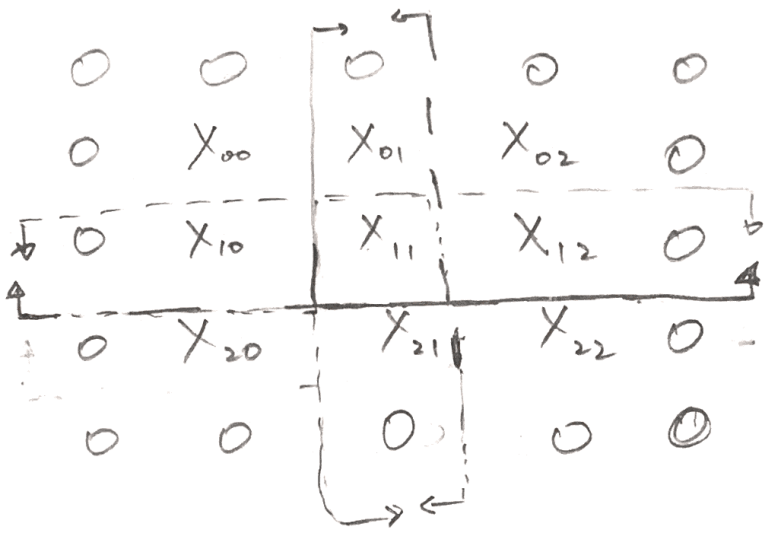
CS 4803 | PS2

Reagan Kan | rkan3 | 903 404 746

1.1



zero-pad size = 1

stride = 2

y shape = (4, 1)

X shape = (9, 1)

→ A shape = (4, 9)

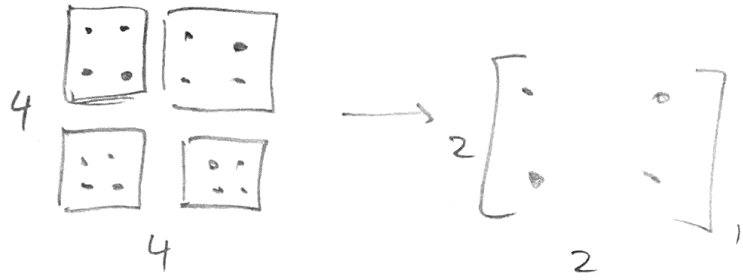
$$A = \begin{bmatrix} X_{00} & X_{01} & X_{02} & X_{10} & X_{11} & X_{12} & X_{20} & X_{21} & X_{22} \\ W_{11} & W_{12} & 0 & W_{21} & W_{22} & 0 & 0 & 0 & 0 \\ 0 & W_{10} & W_{11} & 0 & W_{20} & W_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & W_{01} & W_{02} & 0 & W_{11} & W_{12} & 0 \\ 0 & 0 & 0 & 0 & W_{00} & W_{01} & 0 & W_{10} & W_{11} \end{bmatrix}$$

So, $y = A \cdot X$

$$\begin{bmatrix} y_{00} \\ y_{01} \\ y_{10} \\ y_{11} \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} & 0 & W_{21} & W_{22} & 0 & 0 & 0 & 0 \\ 0 & W_{10} & W_{11} & 0 & W_{20} & W_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & W_{01} & W_{02} & 0 & W_{11} & W_{12} & 0 \\ 0 & 0 & 0 & 0 & W_{00} & W_{01} & 0 & W_{10} & W_{11} \end{bmatrix} \begin{bmatrix} X_{00} \\ X_{01} \\ X_{02} \\ X_{10} \\ X_{11} \\ X_{12} \\ X_{20} \\ X_{21} \\ X_{22} \end{bmatrix}$$

1-2 $W_{shape} = (2,2)$ $X_{shape} = (2,2)$. stride 2, no pad

So a forward convolution takes (r, c), does stride 2 w/ 2x2 kernel, no pad, to get 2x2 output.



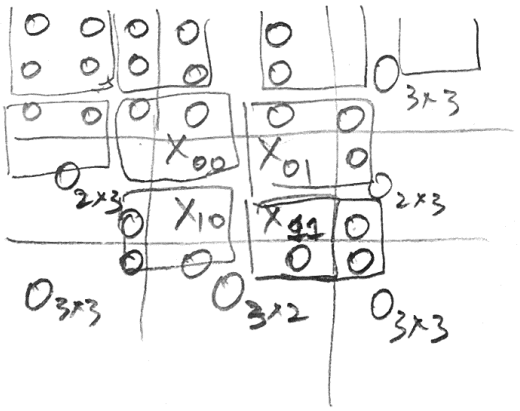
, so $Y_{shape} = (4,4)$.

Must use same kernel $_{2\times2}$ and stride $_{2}$ to convert input dim $(2+2p', 2+2p')$ to $(4,4)$

$4 = ((2+2p') - 2)/2 + 1$ from $((W+2P) - F)/S + 1$

$3 = p'$ So zero pad X w/ 3.

y = Convolve 2x2, w/ 2 stride on the following:



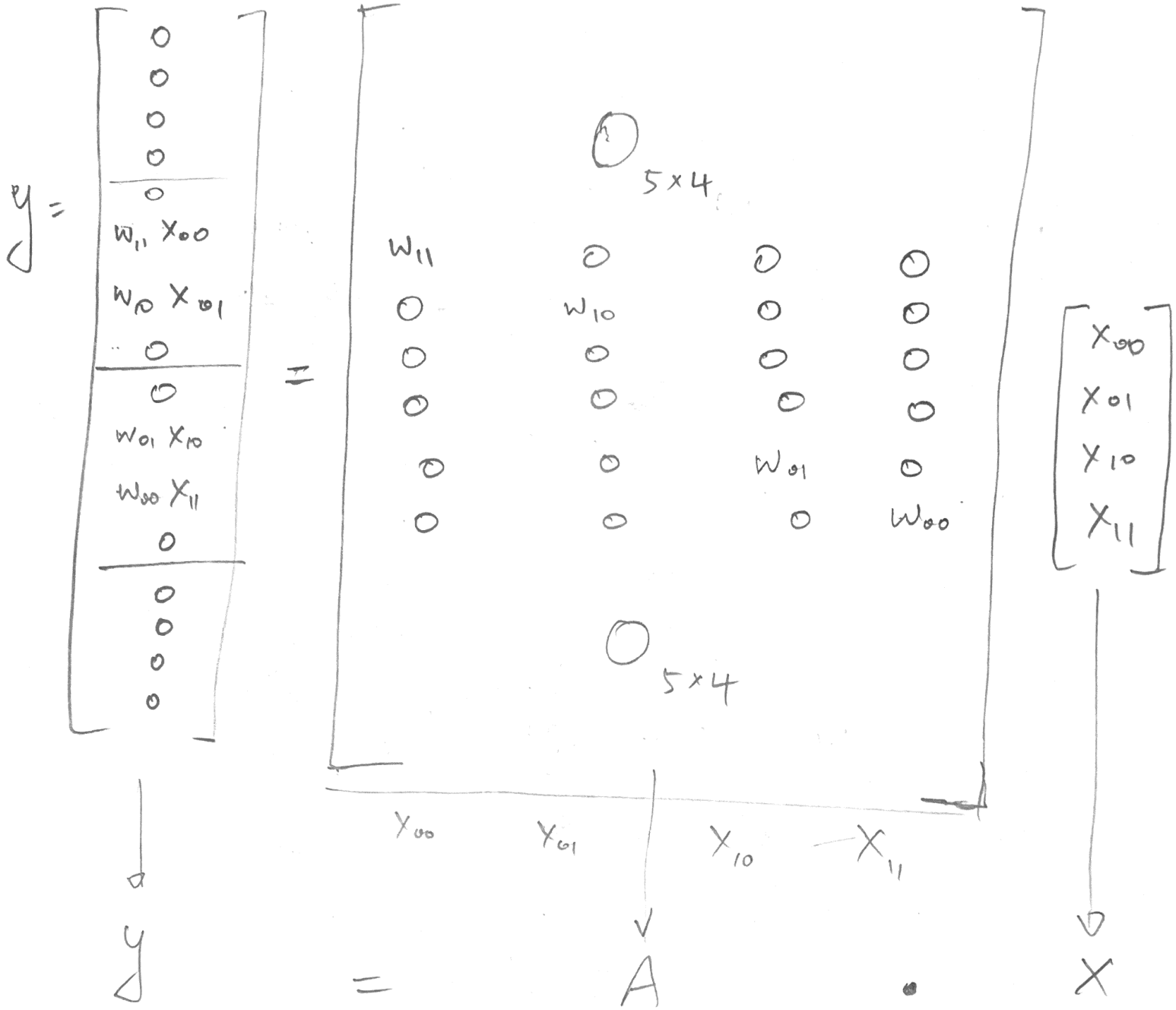
row major

$4\times4 \rightarrow 16\times1$

$$y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & X_{00}W_{11} & X_{01}W_{10} & 0 \\ 0 & X_{10}W_{01} & X_{11}W_{00} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Now A has shape (16×4)

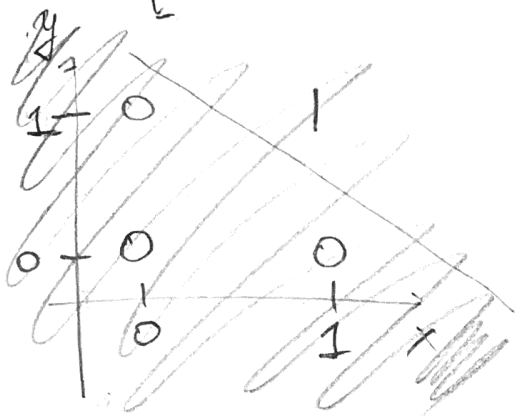
Need to find A s.t. $y_{row} = A \begin{bmatrix} X_{00} \\ X_{01} \\ X_{10} \\ X_{11} \end{bmatrix}$ →


y =
0
0
0
0
0
$w_{11} x_{00}$
$w_{10} x_{01}$
0
0
$w_{01} x_{10}$
$w_{00} x_{11}$
0
0
0
0
0
=
w_{11}
w_{10}
w_{01}
w_{00}
5×4
5×4
x_{00}
x_{01}
x_{10}
x_{11}
y = A . X

2-1] $w \in \mathbb{R}^2$, $b \in \mathbb{R}$, $X \in \{0,1\}^2$

$$f(x) = \begin{cases} 1 & w^T x + b \geq 0 \longrightarrow w^T x \geq -b \\ 0 & w^T x + b < 0 \longrightarrow w^T x < -b \end{cases}$$

f_{AND}



$w_1(1) + w_2(1) \geq -b$ (a)

$w_1(0) + w_2(0) < -b$ (b)

$w_1(0) + w_2(1) < -b$ (c)

$w_1(1) + w_2(0) < -b$ (d)

from (b), $0 < -b$, so $b < 0$. Let's fix $b = -1$

then (a) becomes $w_1 + w_2 \geq 1$

(c) " $w_2 < 1$

(d) $w_1 < 1$

Letting $w_1 = w_2 = 0.6$ satisfies these 3 inequalities.

$$W_{AND} = \begin{pmatrix} 0.6 \\ 0.6 \end{pmatrix} \quad b_{AND} = -1.0$$

f_{OR}

$w_1(0) + w_2(0) < -b \rightarrow 0 < -b$, $b < 0$, Again fix $b = -1$

$w_1(0) + w_2(1) \geq 1 \rightarrow w_2 \geq 1$ (a)

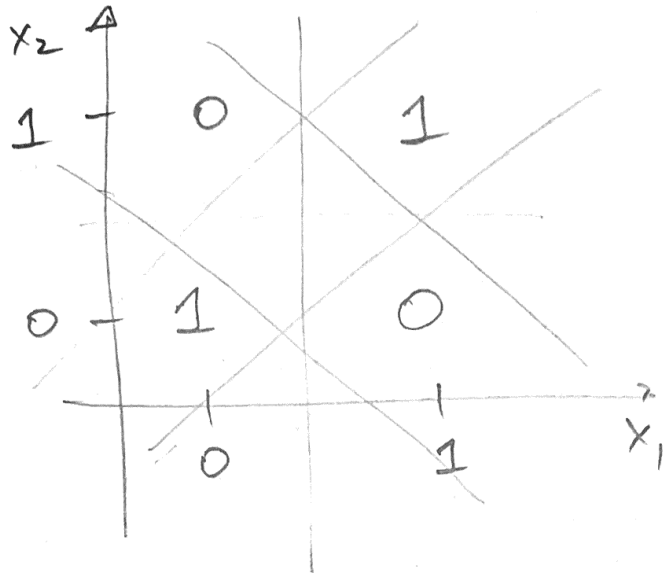
$w_1(1) + w_2(0) \geq 1 \rightarrow w_1 \geq 1$ (b)

$w_1(1) + w_2(1) \geq 1 \rightarrow w_1 + w_2 \geq 1$ (c)

Let $w_1 = w_2 = 1$, satisfies a, b, c.

$$W_{OR} = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \quad b_{OR} = -1.0$$

2.2



$w_1(0) + w_2(0) \geq -b \rightarrow 0 \geq -b \rightarrow b \geq 0$ (a)

$w_1(1) + w_2(1) \geq -b \rightarrow w_1 + w_2 \geq -b$ *

$w_1(1) + w_2(0) < -b \rightarrow w_1 < -b$ **

$w_1(0) + w_2(1) < -b \rightarrow w_2 < -b$ ***

from **, ***, $w_1 + w_2 < -2b$

$\Rightarrow w_1 + w_2 + 2b < 0.$ (b)

from *, $w_1 + w_2 + b \geq 0$

Let $x = w_1 + w_2 + b$, then $x \geq 0.$ (c)

from (b), $(w_1 + w_2 + b) + b < 0, \Rightarrow x + b < 0$

from (a), (b), we know $x \geq 0$, $b \geq 0$, so $x + b \geq 0.$

this contradicts (b) $(x + b < 0)$

So "$\Leftrightarrow$" cannot be represented using linear model of given form.

3.1 $\sigma(\cdot) = f_1(\cdot) = \left|\vec{W}^{(1)}\vec{x} + \vec{b}\right| = \left|2\underset{d\times d}{I}\cdot\underset{d\times 1}{\vec{x}} + \underset{d\times 1}{\vec{b}}\right|_{(d\times 1)}$

Look at each element in $|2I\vec{x} + \vec{b}|$

call it $|(2(1)x_i + b_i)| \quad \forall i \in \{1 \dots d\}$.

$b_i = -1$. (this is given).

We only care about $O = (0,1)^d$. ↳ open range

So $|2x_i - 1| < 1$.

$-1 < 2x_i - 1 < 1$

$\frac{0}{2} < \frac{2x_i}{2} < \frac{2}{2}$

$0 < x_i < 1$

So, for each element, $(0, 1)$ is the only input region that can be mapped to output region $(0,1)$ with given $W^{(1)}$, $b^{(1)}$.

each element of $\sigma(\cdot)$ has 1 input regions.

So, total of 1^d input regions = **1 input region**

$\sigma(\cdot)$ is a bijection, since $\sigma^{-1}(\cdot)$ exists.

$$\sigma^{-1}(\cdot) = \frac{1}{2}I(\vec{x} - \vec{b})$$

3.2] $f \cdot g(\cdot)$ identifies $n_g \cdot n_f$ regions onto $(0,1)^d$

3.3] from the explanation at the top of section 3-Depth, each layer $h^{(\cdot)}$ has d elements, each of which identifies 2 region inputs, so each layer has 2^d regions that are identified.

Since the entire net has L, layers, and

from the result of 3.2, composition of functions identifies a number of regions equal to the product of the number of regions identified by each composed function,

Then the L layer net is a composition of L functions and each layer has $\underbrace{2 \cdot 2 \cdots \cdot 2}_{d}$ or 2^d identified regions

$$L \text{ layer net} = \underbrace{2^d \cdot 2^d \cdot \cdots \cdot 2^d}_{L} = 2^{Ld}$$

Number 5: Paper Review

1. [2 points] The paper shows that training with very weak (noisy) labels still leads to robust learning of features that generalize well. Why is deep learning so robust to noise? Provide some conjectures, relating it to what you know about machine/deep learning and optimization.

We discussed the softmax layer before in class. For classification problems, this layer helps factor in the relative scores of each class label. The inter class relationships may be difficult for noise to affect significantly.

Also, deep neural nets are widely trained using some form of gradient descent optimization, which cases where the loss is non-convex, converges to local minima, or an estimation of the true global minima. Estimations can be more invariant to noise. For example, taking the mean of a sample can reduce the effects of noise.

Noise could also be viewed as simply part of the overall problem. The noise just raises the complexity of the problem, thus raising the threshold for the model capacity to sufficiently solve the problem. Deep networks have many non-linear layers, which result in a model with very high capacity. It is possible that modern deep networks are deep enough to have a high enough model complexity to describe problems that incorporate noise.

2. [2 points] Another finding is that learning features using similar label spaces (i.e. on categories that overlap with what you are hoping to generalize to) is more successful than learning features on dissimilar label spaces. Why might that be the case?

Often times, label spaces can be converted into numerical vector spaces. So the problem can be rephrased as why learning features from a “similar” or “overlapping” vector space is more successful than learning on “non-overlapping” vector spaces.

From a linear algebra perspective, the essence of a space is encoded in its basis so the notion of “overlapping” could be represented by intersecting bases. This means that there is a subspace which is spanned by the vectors that belong to both bases. This subspace may contain information that is shared across both vector spaces. Learning features related to this subspace may reveal useful information about both spaces, which includes the target space. Dissimilar label spaces will not share this subspace and useful overlapping information cannot be learned.

Network Visualization (45 Points)

In the first part of the notebook we will explore the use of different type of attribution algorithms - both gradient and perturbation - for images, and understand their differences using the Captum model interpretability tool for PyTorch.

Link to the Website: https://captum.ai/ (https://captum.ai/) Link to the Github page: https://github.com/pytorch/captum (https://github.com/pytorch/captum)

As an exercise you'll be also asked to implement Saliency Maps from scratch.

- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.
- Mukund Sundararajan, Ankur Taly, Qiqi Yan, "Axiomatic Attribution for Deep Networks", ICML, 2017
- Matthew D Zeiler, Rob Fergus, "Visualizing and Understanding Convolutional Networks", Visualizing and Understanding Convolutional Networks, 2013.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2016

For the full list of available attribution algorithms please check out: https://captum.ai/api/ (https://captum.ai/api/)

In the second and third parts we will focus on generating new images, by studying and implementing key components in two papers:

- Szegedy et al, "Intriguing properties of neural networks", ICLR 2014
- Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML 2015 Deep Learning Workshop

You will need to first read the papers, and then we will guide you to understand them deeper with some problems.

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.

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

This notebook is the first part of homework 2. We will explore four different techniques:

1. **Saliency Maps**: Saliency maps are a quick way to tell which part of the image influenced the classification decision made by the network.
2. **GradCAM**: GradCAM is a way to show the focus area on an image for a given label.
3. **Fooling Images**: We can perturb an input image so that it appears the same to humans, but will be misclassified by the pretrained network.
4. **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.

We will use **PyTorch 1.4** to finish the problems in this notebook, which has been tested with Python3.6 on Linux and Mac.

Suppose you have already installed the dependencies in the last homework. **Before you start this one, here are some preparation work you need to do**:

- Download the imagenet_val_25 dataset

```
cd cs7643/datasets
bash get_imagenet_val.sh
```

- To install captum (please install the latest version 0.2.0 from source - for more information see https://github.com/pytorch/captum (https://github.com/pytorch/captum)),

```
git clone https://github.com/pytorch/captum.git
cd captum
pip install -e .
```

- The total credit for this notebook is 45 points - 10 points for each section, and 5 points for the captum attribution calls.
- Although we will run your notebook in grading, you still need to **submit the notebook with all the outputs you generated, in pdf form**. Sometimes it will inform us if we get any inconsitent results with respect to yours.

In [1]:
```
import torch
from torch.autograd import Variable
from torch.autograd import Function as TorchFunc
import torchvision
import torchvision.transforms as T
import random

import numpy as np
from scipy.ndimage.filters import gaussian_filter1d
import matplotlib.pyplot as plt
import matplotlib
from cs7643.image_utils import SQUEEZENET_MEAN, SQUEEZENET_STD
from PIL import Image
import captum

%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'
print(captum.__version__)
print(torch.__version__)
```

```
0.2.0
1.4.0
```

Helper Functions

Our pretrained model was trained on images that had been preprocessed by subtracting the per-color mean and dividing by the per-color standard deviation. We define a few helper functions for performing and undoing this preprocessing.

You don't need to do anything in this cell. Just run it.

In [2]:
```
def preprocess(img, size=224):
    transform = T.Compose([
        T.Resize(size),
        T.ToTensor(),
        T.Normalize(mean=SQUEEZENET_MEAN.tolist(),
                    std=SQUEEZENET_STD.tolist()),
        T.Lambda(lambda x: x[None]),
    ])
    return transform(img)

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

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

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

def visualize_attr_maps(attributions, titles, attr_preprocess=lambda attr: attr.permute(1, 2, 0).detach().numpy(),
                        cmap='viridis', alpha=0.7):
    '''
    A helper function to visualize captum attributions for a list of captum attribution algorithms.

    attributions(A list of torch tensors): Each element in the attributions list corresponds to an
                    attribution algorithm, such an Saliency, Integrated Gradient, Perturbation, etc.
                    Each row in the attribution tensor contains
    titles(A list of strings): A list of strings, names of the attribution algorithms corresponding to each element in
                    the `attributions` list. len(attributions) == len(titles)
    '''
    N = attributions[0].shape[0]
    plt.figure()
    for i in range(N):
        axs = plt.subplot(len(attributions) + 1, N + 1, i+1)
        plt.imshow(X[i])
        plt.axis('off')
        plt.title(class_names[y[i]])

    plt.subplot(len(attributions) + 1, N + 1 , N + 1)
    plt.text(0.0, 0.5, 'Original Image', fontsize=14)
    plt.axis('off')
    for j in range(len(attributions)):
        for i in range(N):
            plt.subplot(len(attributions) + 1, N + 1 , (N + 1) * (j + 1) + i + 1)
            print(f"attributions shape: {attributions[j][i].shape}")
            attr = np.array(attr_preprocess(attributions[j][i]))
            attr = (attr - np.mean(attr)) / np.std(attr).clip(1e-20)
            attr = attr * 0.2 + 0.5
            attr = attr.clip(0.0, 1.0)
            plt.imshow(attr, cmap=cmap, alpha=alpha)
            plt.axis('off')
        plt.subplot(len(attributions) + 1, N + 1 , (N + 1) * (j + 1) + N + 1)
        plt.text(0.0, 0.5, titles[j], fontsize=14)
        plt.axis('off')

    plt.gcf().set_size_inches(20, 13)
    plt.show()

def compute_attributions(algo, inputs, **kwargs):
    '''
    A common function for computing captum attributions
    '''
    return algo.attribute(inputs, **kwargs)
```

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, 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 the experiments in this notebook on a CPU machine. You are encouraged to use a larger model to finish the rest of the experiments if GPU resouces are not a problem for you, but please highlight the backbone network you use in your implementation if you do it.

Switching a backbone network is quite easy in pytorch. You can refer to [torchvision model zoos (https://github.com/pytorch/vision/tree/v0.2.1/torchvision/models)](https://github.com/pytorch/vision/tree/v0.2.1/torchvision/models) for more information.

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

In [3]:
```
# Download and load the pretrained SqueezeNet model.
model = torchvision.models.squeezenet1_1(pretrained=True)

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

Load some ImageNet images

If you have not execute the downloading script. Here is a reminder that you have to do it now. We have provided a few example images from the validation set of the ImageNet ILSVRC 2012 Classification dataset.

To download these images run

```
cd cs7643/datasets/
bash get_imagenet_val.sh
```

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

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

In [4]:
```
from cs7643.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()
```

Saliency Maps (10 pts)

Using this pretrained model, we will compute class saliency maps as described in the paper:

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014. (https://arxiv.org/abs/1312.6034)

We will also review this paper in the paper presentation.

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.

Hint: PyTorch `gather` method

Recall when you need 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 <= y[i] < C`, then `s[np.arange(N), y]` is a numpy array of shape `(N,)` which selects one element from each element in `s` using the indices in `y`.

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

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

will be a PyTorch Tensor (or Variable) 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 (http://pytorch.org/docs/torch.html#torch.gather) and the squeeze method (http://pytorch.org/docs/torch.html#torch.squeeze).

In [5]:
```
# Example of using gather to select one entry from each row in PyTorch.
# I added the max_indices and gather code.
def gather_example():
    N, C = 4, 5
    s = torch.randn(N, C)
    indices = torch.argmax(s, dim=1)
    print(f" max indices {indices.view(-1, 1).shape}")
    print(f" gradient {s.shape}")
    print(s.gather(1, indices.view(-1, 1)).shape)

    s = torch.randn(5,3,234,234)
    indices = torch.argmax(s, dim=1)
    y = torch.LongTensor([1, 2, 1, 3])
    print(f" max indices {indices.shape}")
    print(f" gradient {s.shape}")
    (a, b, c) = indices.shape
    print(s.gather(1, indices.view(a, -1, b, c)).shape)
#     for i in range(5):
#         print(f" max indices {indices[i].shape}")
#         print(f" gradient {s[i].shape}")
#         print(s[i].gather(1, indices[i]))
gather_example()
```

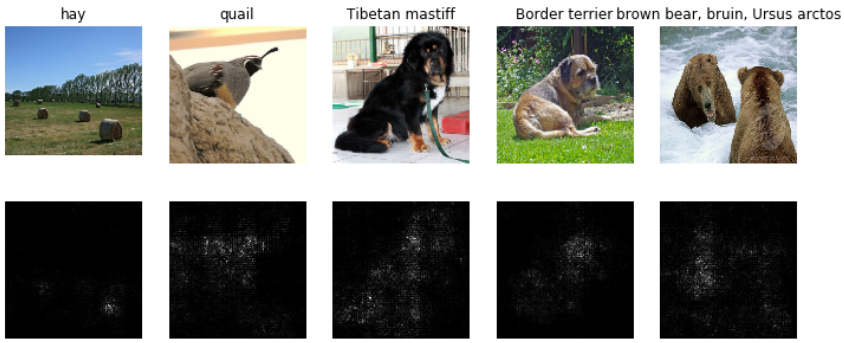
```
 max indices torch.Size([4, 1])
 gradient torch.Size([4, 5])
torch.Size([4, 1])
 max indices torch.Size([5, 234, 234])
 gradient torch.Size([5, 3, 234, 234])
torch.Size([5, 1, 234, 234])
```

In [6]:
```
def compute_saliency_maps(X, y, model):
    """
    Compute a class saliency map using the model for images X and labels y.

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

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

    # Wrap the input tensors in Variables
    X_var = Variable(X, requires_grad=True)
    y_var = Variable(y, requires_grad=False)
    saliency = None

    lam = 1e3 # This is the regularization parameter when you need it

    ##############################################################################
    # TODO: Implement this function. Perform a forward and backward pass through #
    # the model to compute the gradient of the correct class score with respect  #
    # to each input image. You first want to compute the loss over the correct   #
    # scores, and then compute the gradients with a backward pass.               #
    ##############################################################################
    # 1. compute scores with a forward pass
    scores_all_classes = model.forward(X_var)

    # 1.5 get scores for specific class label
    scores = scores_all_classes.gather(1, y_var.view(-1, 1)).squeeze()
    print(list(scores.shape))
    print(scores)
    print(y)

    # 2. backprop to find gradScores w.r.t Image. i.e. dScore/dX
    scores.backward(torch.ones_like(scores)) #do the backpass
    gradient = X_var.grad #this is w in the paper
    abs_gradient = gradient.abs()

    # 3. use gradient to pick i,j pixel index.
    # self-hint(?): use gather to take max across channels.
    max_indices = torch.argmax(abs_gradient, dim=1)


    print(f"gradient shape: {abs_gradient.shape}")
    print(f"max_indices shape (should be gradShape.drop(1)): {max_indices.shape}")

    (a, b, c) = max_indices.shape
    saliency = abs_gradient.gather(1, max_indices.view(a, -1, b, c)).squeeze()
    print(f"gradient shape: {abs_gradient.shape}")
    print(f"max_indices shape (should be gradShape.drop(1)): {max_indices.shape}")
    print(f"saliency shape: {saliency.shape}")
    print(f"y_var: {y}")
    ##############################################################################
    #                               END OF YOUR CODE                             #
    ##############################################################################
    return saliency
```

Once you have completed the implementation in the cell above, run the following to visualize some class saliency maps on our example images from the ImageNet validation set. You can compare to the figure 2 in the refered paper as a comparison for your results.

In [7]:
```
def show_saliency_maps(X, y):
    # Convert X and y from numpy arrays to Torch Tensors
    X_tensor = torch.cat([preprocess(Image.fromarray(x)) for x in X], dim=0)
    print(y)
    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)
        print(f"X[{i}]: {X[i].shape}")
        plt.imshow(X[i])
        plt.axis('off')
        plt.title(class_names[y[i]])
        plt.subplot(2, N, N + i + 1)
        print(f"saliency[{i}]: {saliency[i].shape}")
        plt.imshow(saliency[i], cmap=plt.cm.gray)
        plt.axis('off')
        plt.gcf().set_size_inches(12, 5)
    plt.show()

show_saliency_maps(X, y)
```

```
[958  85 244 182 294]
[5]
tensor([24.1313, 25.1475, 38.8825, 25.4514, 30.2723],
       grad_fn=<SqueezeBackward0>)
tensor([958,  85, 244, 182, 294])
gradient shape: torch.Size([5, 3, 224, 224])
max_indices shape (should be gradShape.drop(1)): torch.Size([5, 224, 224])
gradient shape: torch.Size([5, 3, 224, 224])
max_indices shape (should be gradShape.drop(1)): torch.Size([5, 224, 224])
saliency shape: torch.Size([5, 224, 224])
y_var: tensor([958,  85, 244, 182, 294])
X[0]: (224, 224, 3)
saliency[0]: (224, 224)
X[1]: (224, 224, 3)
saliency[1]: (224, 224)
X[2]: (224, 224, 3)
saliency[2]: (224, 224)
X[3]: (224, 224, 3)
saliency[3]: (224, 224)
X[4]: (224, 224, 3)
saliency[4]: (224, 224)
```



Captum (1 pt)

As a final step, we will show you how simple it is to use Saliency Maps using Captum instead.

Captum offers a number of attribution algorithms for PyTorch models that are very easy to use. Let's apply those algorithms to our model and the five images that we loaded from imagenet.

We have created a generic helper visualization function that will allow you to visualize and compare the attributions of different algorithms next to each other for each image. (`compute_attributions` and `visualize_attr_maps` , found in the helper functions section)

Let's apply the saliency maps attribution algorithm on the images and observe how the attributions differ.

To do so we need to import those algorithms from the Captum library, create instances of the corresponding algorithms, call our compute_attribute function on these instances, and finally visualize using our helper function.

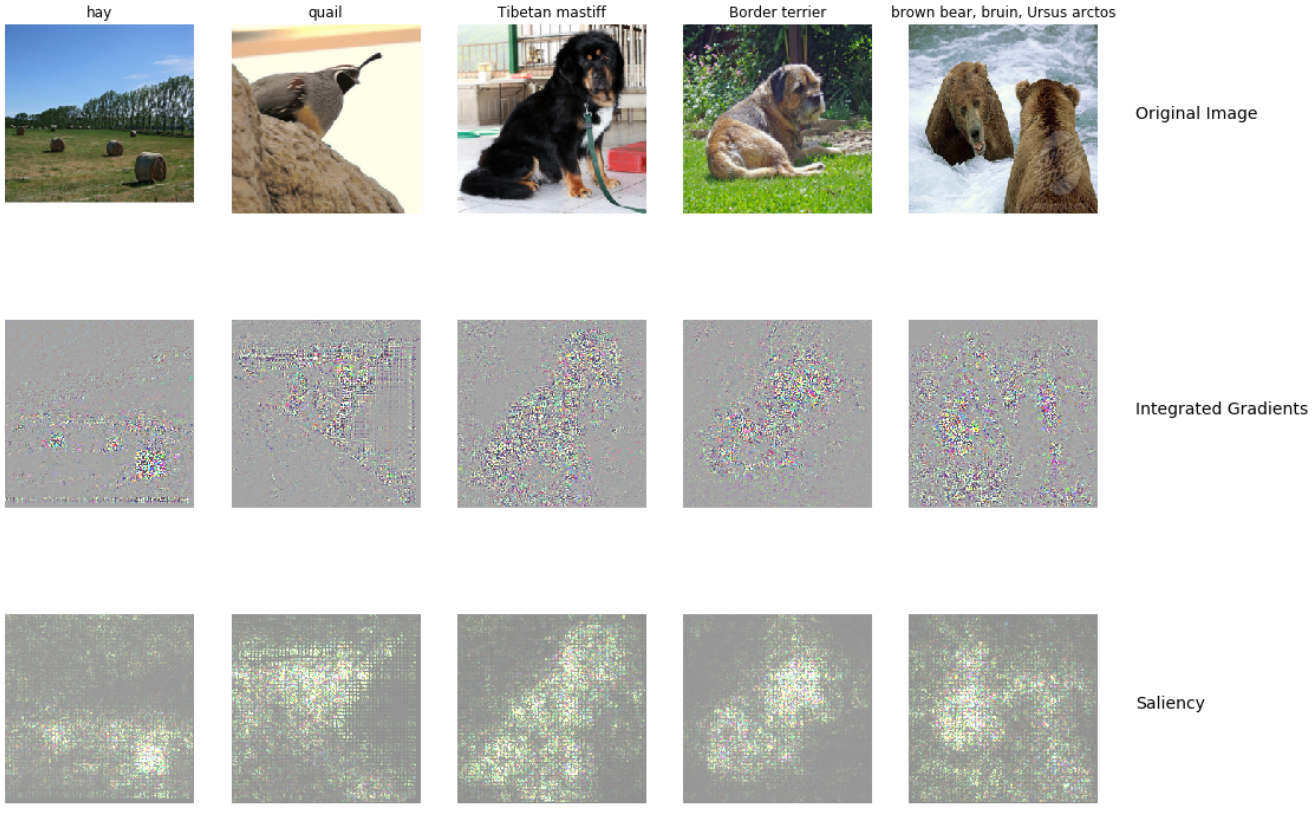
We have included an example of how we can apply a number of attribution algorithms on our model and images.

Feel free to try other algorithms and compare their results.

Please, be aware that some of the algorithms, such as perturbation algorithms, might take longer to execute and might have higher memory requirments. In case you run into OOM issues for integrated gradients you can try to reduce the number of integral approximation steps (n_steps) or set the value of internal_batch_size input argument to a small number. The value of `ablations_per_eval` or `perturbation_per_eval` can be adjusted also for all perturbation algorithms in order to reduce memory footprint. This might lead to a slower execution runtime, however it will help to avoid OOM.

In [8]:
```
from captum.attr import IntegratedGradients, Saliency


# 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)


# Computing Integrated Gradient
int_grads = IntegratedGradients(model)
attr_ig = compute_attributions(int_grads, X_tensor, target=y_tensor, n_steps=10)

################################################################################
# TODO: Compute/Visualize Saliency using captum.                               #
################################################################################
sal_grads = Saliency(model)
attr_sal = compute_attributions(sal_grads, X_tensor, target=y_tensor)
visualize_attr_maps([attr_ig, attr_sal], ['Integrated Gradients', 'Saliency'])
################################################################################
#                             END OF YOUR CODE                                 #
################################################################################
```

```
/Users/reagankan/captum/captum/attr/_utils/gradient.py:32: UserWarning: Input Tensor 0 did not already require gradients, required_grads has been s
et automatically.
  "required_grads has been set automatically." % index

attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
```



GradCAM (10 pts)

GradCAM (which stands for Gradient Class Activation Mapping) is a technique that tells us where a convolutional network is looking when it is making a decision on a given input image. There are three main stages to it:

- **Guided Backprop (Changing ReLU Backprop Layer, https://arxiv.org/abs/1412.6806 (https://arxiv.org/abs/1412.6806))**
- **GradCAM (Manipulating gradients at the last convolutional layer, https://arxiv.org/abs/1610.02391 (https://arxiv.org/abs/1610.02391))**
- **Guided GradCAM (Pointwise multiplication of above stages)**

In this section, you will be implementing these three stages to recreate the full GradCAM pipeline. At each stage, you can visualize what the current output is to see exactly what is happening.

We begin with Guided Backprop. We encourage you to read the paper above first, to gain an understanding of what Guided Backprop is trying to do. From the paper, we have that:

- The 'deconvolution' is equivalent to a backward pass through the network, except that when propagating through a nonlinearity, its gradient is solely computed based on the top gradient signal, ignoring the bottom input. In case of the ReLU nonlinearity this amounts to setting to zero certain entries based on the top gradient. We propose to combine these two methods: rather than masking out values corresponding to negative entries of the top gradient ('deconvnet') or bottom data (backpropagation), we mask out the values for which at least one of these values is negative.

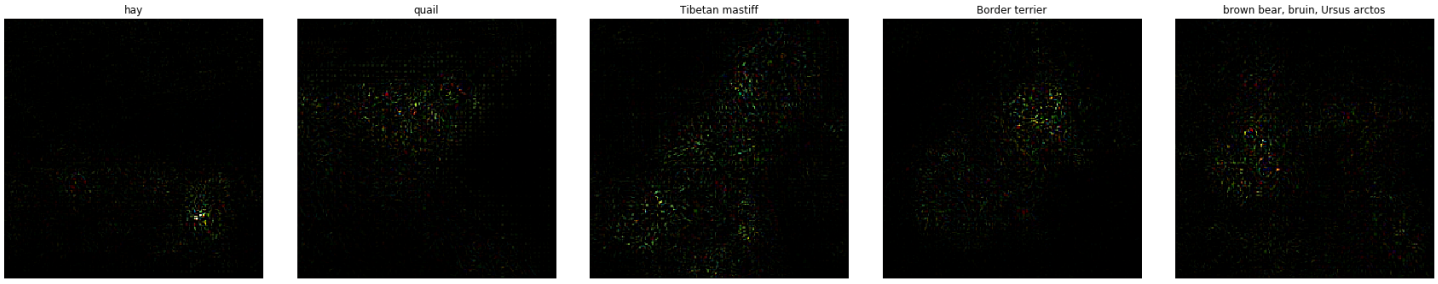
In [9]:
```
# FOR THIS SECTION ONLY, we need to use gradients. We introduce a new model we will use explicitly for GradCAM for this.
gc_model = torchvision.models.squeezenet1_1(pretrained=True)
for param in gc_model.parameters():
    param.requires_grad = True
```

In [10]:
```
class CustomReLU(TorchFunc):
    """
    Define the custom change to the standard ReLU function necessary to perform guided backpropagation.
    We have already implemented the forward pass for you, as this is the same as a normal ReLU function.
    """

    @staticmethod
    def forward(self, x):
        #print((x > 0).type_as(x))
        output = torch.addcmul(torch.zeros(x.size()), x, (x > 0).type_as(x))
        self.save_for_backward(x, output)
#         print(self.saved_tensors)
        return output

    @staticmethod
    def backward(self, y):
        ################################################################################
        # TODO: Implement this function. Perform a backwards pass as described in    #
        # paper above. Note: torch.addcmul might be useful, and you can access       #
        # the input/output from the forward pass with self.saved_tensors.            #
        ################################################################################
#         print(f"forward io: {self.saved_tensors}")
        forwardIn, forwardOut = self.saved_tensors
#         print(f"seven: {seven}")
        bottomMask = (forwardIn > 0).type_as(forwardIn)
        topMask = (y > 0).type_as(y)
        assert(forwardIn.size() == y.size())
        applyBottomMask = torch.addcmul(torch.zeros(forwardIn.size()), y, bottomMask)
        applyTopMask = torch.addcmul(torch.zeros(forwardIn.size()), applyBottomMask, topMask)

        superMask = torch.addcmul(torch.zeros(bottomMask.size()), bottomMask, topMask)
        superOutput = torch.addcmul(torch.zeros(bottomMask.size()), y, superMask)
#         assert(torch.eq(superOutput, applyTopMask))
        assert(torch.all(torch.eq(superOutput, applyTopMask)))
        return applyTopMask

        ################################################################################
        #                              END OF YOUR CODE                                #
        ################################################################################
```

To test your implementation, run the code below.

In [11]:
```
for idx, module in gc_model.features._modules.items():
    if module.__class__.__name__ == 'ReLU':
        gc_model.features._modules[idx] = CustomReLU.apply

def guided_backprop(X_tensor,y_tensor):
    ##########################################################################
    # TODO: Implement guided backprop as described in paper.                 #
    # (Hint): Now that you have implemented the custom ReLU function, this   #
    # method will be similar to a single training iteration.                 #
#     ##########################################################################
    print(f"y tensor: {y_tensor}")
    print(f"ones_like type: {torch.ones_like(y_tensor).float()}")
    scores = gc_model.forward(X_tensor).gather(1, y_tensor.view(-1, 1)).squeeze()
    scores.backward(torch.ones_like(y_tensor).float())
    gradient1 = X_tensor.grad.permute(0,2,3,1)

    scores = gc_model.forward(X_tensor)
#     print(scores.shape)
    init_grad = torch.zeros_like(scores).float()
    init_grad[list(range(5)), y_tensor] = 1.0
#     print(torch.argmax(init_grad[0]))
#     print(y_tensor)
    scores.backward(init_grad)
    gradient = X_tensor.grad
    print(f"gradient shape: {gradient.shape}")
    gradient = gradient.permute(0,2,3,1) #do not use .view, which changes ordering of elements.
    print(f"gradient shape: {gradient.shape}")
#     print(torch.ones_like(gradient) == (gradient >= 0).type_as(gradient))
#     print(gradient)
    assert(torch.all(torch.eq(gradient, gradient1)))
#     return gradient1
    return gradient #this is slightly better than gradient1. not sure if my eyes are bad since they are equal.
    ##########################################################################
    #                              END OF YOUR CODE                          #
    ##########################################################################

X_tensor = torch.cat([preprocess(Image.fromarray(x)) for x in X], dim=0).requires_grad_(True)
y_tensor = torch.LongTensor(y)
gbp_result = guided_backprop(X_tensor,y_tensor)

plt.figure(figsize=(24, 24))
for i in range(gbp_result.shape[0]):
    plt.subplot(1, 5, i + 1)
    c, w, h = gbp_result[i].shape
#     print(gbp_result[i].shape)
#     plt.imshow(gbp_result[i].view(w, h, c))
    plt.imshow(gbp_result[i])
    plt.title(class_names[y[i]])
    plt.axis('off')
plt.gcf().tight_layout()
```

```
y tensor: tensor([958,  85, 244, 182, 294])
ones_like type: tensor([1., 1., 1., 1., 1.])

Clipping input data to the valid range for imshow with RGB data ([0..1] for floats or [0..255] for integers).
Clipping input data to the valid range for imshow with RGB data ([0..1] for floats or [0..255] for integers).
Clipping input data to the valid range for imshow with RGB data ([0..1] for floats or [0..255] for integers).
Clipping input data to the valid range for imshow with RGB data ([0..1] for floats or [0..255] for integers).
Clipping input data to the valid range for imshow with RGB data ([0..1] for floats or [0..255] for integers).

gradient shape: torch.Size([5, 3, 224, 224])
gradient shape: torch.Size([5, 224, 224, 3])
```



Next, we can implement GradCAM. We have given you which module(=layer) that we need to capture gradients from, which you can see in `conv_module` variable below - feel free to play around with this to see visualizations for different layers, but in your final submission keep it to what we gave you. We have already provided a gradient and activation hook for you, and the gradient value of the module you choose will be stored in the `gradient_value` variable (similarly `activation_value` will hold the layer activation). The rest of the implementation of GradCAM is up to you.

In [12]:
```
# Reset model from any changes made during Guided Backprop
gc_model = torchvision.models.squeezenet1_1(pretrained=True)
for param in gc_model.parameters():
    param.requires_grad = True
```

In [13]:
```
conv_module = gc_model.features[12]

gradient_value = None # Stores gradient of the module you chose above during a backwards pass.
activation_value = None # Stores the activation of the module you chose above during a forwards pass.

def gradient_hook(a,b,gradient):
    global gradient_value
    gradient_value = gradient[0]

def activation_hook(a,b,activation):
    global activation_value
    activation_value = activation

conv_module.register_forward_hook(activation_hook)
conv_module.register_backward_hook(gradient_hook)
```

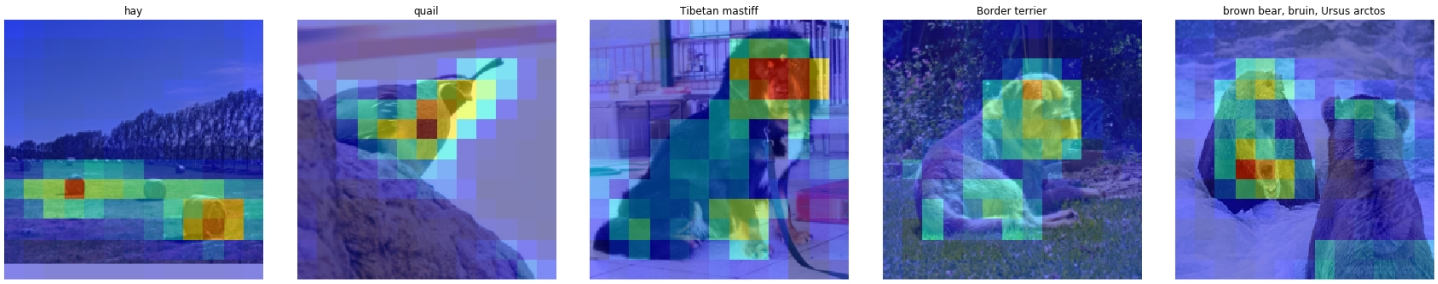
Out[13]:
```
<torch.utils.hooks.RemovableHandle at 0x12e2b35f8>
```

In [14]:
```
def grad_cam(X_tensor, y_tensor):
    ##############################################################################
    # TODO: Implement GradCam as described in paper.                             #
    ##############################################################################
    cam = None

    #forward pass image. activation hook should've saved desired activation map into activation_value.
    scores = gc_model.forward(X_tensor)

    #prep backward: zero out unwanted classes in gradient
    init_grad = torch.zeros_like(scores).float()
    init_grad[list(range(y_tensor.shape[0])), y_tensor] = 1.0

    #backpropagate. gradient hook should've saved dScore/dActivationMap into gradient_value.
    scores.backward(init_grad)
#     dScoreDActivation = gradient_value

    #alpha: neuron importance weights
    alpha = torch.mean(gradient_value, (2,3))
    print(f"alpha size: {alpha.shape}")
    print(f"alpha: {alpha}")
    print(f"gradient_value size: {gradient_value.shape}")
#     print(f"gradient_value: {gradient_value}")

    #cam: ReLU(sum[k](alpha[k] * grad[k]))
    # [WRONG]in this case there is only 1 k value ?????
    # [GOOD] or is it k in [0,512) since activation_value size: torch.Size([5, 512, 13, 13])
    print(f"activation_value size: {activation_value.shape}")
    cam = alpha[:, :, None, None] * activation_value
    cam = torch.sum(cam, axis=1)
    print(f"cam size: {cam.shape}")

    cam = torch.addcmul(torch.zeros(cam.size()), cam, (cam > 0).type_as(cam)) #ReLU
    print(f"image size: {X_tensor.shape}")
    print(f"cam size: {cam.shape}")
    cam = cam.detach().numpy() #code below requires numpy
    ##############################################################################
    #                             END OF YOUR CODE                               #
    ##############################################################################

    # Rescale GradCam output to fit image.
    cam_scaled = []
    for i in range(cam.shape[0]):
        cam_scaled.append(np.array(Image.fromarray(cam[i]).resize(X_tensor[i,0,:,:].shape)))
    cam = np.array(cam_scaled)
    cam -= np.min(cam)
    cam /= np.max(cam)
    return cam
```

To test your implementation, run the code below.

In [15]:
```
X_tensor = torch.cat([preprocess(Image.fromarray(x)) for x in X], dim=0).requires_grad_(True)
y_tensor = torch.LongTensor(y)
gradcam_result = grad_cam(X_tensor, y_tensor)

plt.figure(figsize=(24, 24))
for i in range(gradcam_result.shape[0]):
    gradcam_val = gradcam_result[i]
    img = X[i] + (matplotlib.cm.jet(gradcam_val)[:,:,:3]*255)
    img = img / np.max(img)
    plt.subplot(1, 5, i + 1)
    plt.imshow(img)
    plt.title(class_names[y[i]])
    plt.axis('off')
plt.gcf().tight_layout()
```

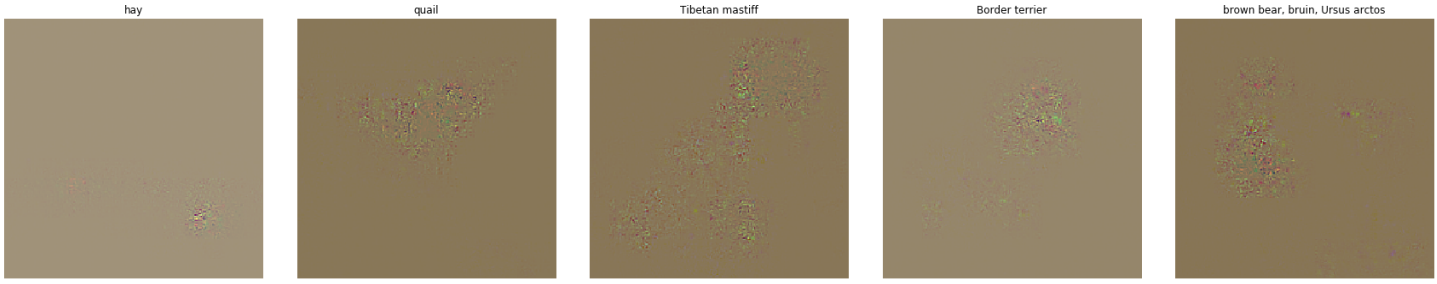
```
alpha size: torch.Size([5, 512])
alpha: tensor([[ 9.3862e-05, -3.7522e-05,  6.3260e-05,  ...,  2.9678e-04,
          1.1335e-04,  5.9996e-05],
        [ 3.8974e-05, -2.4997e-04,  2.7565e-05,  ..., -3.6877e-05,
          1.4532e-04, -2.4953e-04],
        [-3.5052e-04, -6.6962e-05, -3.5692e-05,  ..., -1.7163e-04,
         -2.0609e-04,  3.9557e-04],
        [ 4.8211e-04, -1.9298e-04,  1.1444e-04,  ...,  2.6151e-04,
          1.2922e-04,  2.1928e-04],
        [-3.7134e-04, -2.4911e-04, -2.6240e-04,  ..., -2.8875e-04,
         -2.1763e-04, -2.0061e-05]])
gradient_value size: torch.Size([5, 512, 13, 13])
activation_value size: torch.Size([5, 512, 13, 13])
cam size: torch.Size([5, 13, 13])
image size: torch.Size([5, 3, 224, 224])
cam size: torch.Size([5, 13, 13])
```



As a final step, we can combine GradCam and Guided Backprop to get Guided GradCam.

In [16]:
```
X_tensor = torch.cat([preprocess(Image.fromarray(x)) for x in X], dim=0).requires_grad_(True)
y_tensor = torch.LongTensor(y)
gradcam_result = grad_cam(X_tensor, y_tensor)
gbp_result = guided_backprop(X_tensor,y_tensor)
print(type(gradcam_result))
print(type(gbp_result))
# assert(False)

plt.figure(figsize=(24, 24))
for i in range(gradcam_result.shape[0]):
    gbp_val = gbp_result[i]
    gbp_val /= torch.max(gbp_val)
    gradcam_val = (matplotlib.cm.jet(gradcam_result[i])[:,:,:3]*255)

    ################################################################################
    # TODO: Pointwise multiplication and normalization of the gradcam and guided #
    #       backprop results (2 lines)                                           #
    ################################################################################
    gradcam_val = torch.from_numpy(gradcam_val)
    img = gradcam_val * gbp_val #pointwise multiplication
    img /= torch.max(img) #normalized
    print("differences btw TA and mine are likely come from diff in GradCam heatmaps.")
#     print("hello")
#     print(gradcam_val.shape); print(type(gradcam_val))
#     print(gbp_val.shape); print(type(gbp_val))
#     print(f"img dim: {img.shape}")
    ################################################################################
    #                              END OF YOUR CODE                                #
    ################################################################################

    img = np.expand_dims(img.permute(2, 0, 1),axis=0)#originally: img.transpose(2, 0, 1)
    img = np.float32(img)
    img = torch.from_numpy(img)
    img = deprocess(img)
    plt.subplot(1, 5, i + 1)
    plt.imshow(img)
    plt.title(class_names[y[i]])
    plt.axis('off')
plt.gcf().tight_layout()
```

```
alpha size: torch.Size([5, 512])
alpha: tensor([[ 9.1515e-05, -3.3862e-05,  5.7375e-05,  ...,  3.4427e-04,
          1.1777e-04,  6.1617e-05],
        [ 4.4631e-05, -1.9032e-04,  3.2914e-05,  ..., -3.7946e-05,
          1.8517e-04, -2.1581e-04],
        [-3.6628e-04, -7.0948e-05, -4.0033e-05,  ..., -1.6166e-04,
         -2.1363e-04,  3.3037e-04],
        [ 4.4332e-04, -2.1099e-04,  1.1184e-04,  ...,  2.6752e-04,
          1.2144e-04,  1.8830e-04],
        [-3.5254e-04, -2.4595e-04, -2.3428e-04,  ..., -3.1570e-04,
         -2.3795e-04, -1.9247e-05]])
gradient_value size: torch.Size([5, 512, 13, 13])
activation_value size: torch.Size([5, 512, 13, 13])
cam size: torch.Size([5, 13, 13])
image size: torch.Size([5, 3, 224, 224])
cam size: torch.Size([5, 13, 13])
y tensor: tensor([958,  85, 244, 182, 294])
ones_like type: tensor([1., 1., 1., 1., 1.])
gradient shape: torch.Size([5, 3, 224, 224])
gradient shape: torch.Size([5, 224, 224, 3])
<class 'numpy.ndarray'>
<class 'torch.Tensor'>
differences btw TA and mine are likely come from diff in GradCam heatmaps.
differences btw TA and mine are likely come from diff in GradCam heatmaps.
differences btw TA and mine are likely come from diff in GradCam heatmaps.
differences btw TA and mine are likely come from diff in GradCam heatmaps.
differences btw TA and mine are likely come from diff in GradCam heatmaps.
```



Captum (1 pt)

As a final step, implement GradCam and GuidedBackprop exactly as you did for saliency maps above and compare your visualizations with the ones using Captum (note: **These visualization will look significantly different, as Captum has different pre/post processing steps for images**).

In [17]:
```
from captum.attr import GuidedGradCam, GuidedBackprop


# 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)

################################################################################
# TODO: Compute/Visualize GuidedBackprop and GradCAM as well.                  #
################################################################################
ggc_grads = GuidedGradCam(gc_model, conv_module)
attr_ggc = compute_attributions(ggc_grads, X_tensor, target=y_tensor)

gbp_grads = GuidedBackprop(gc_model)
attr_gbp = compute_attributions(gbp_grads, X_tensor, target=y_tensor)


print(type(attr_ggc))
print(attr_ggc.shape)
visualize_attr_maps([attr_ggc, attr_gbp], ['GuidedGradCam', 'GuidedBackprop'])
# visualize_attr_maps([], [ ''])
################################################################################
#                              END OF YOUR CODE                                #
################################################################################
```

```
/Users/reagankan/captum/captum/attr/_core/guided_backprop_deconvnet.py:58: UserWarning: Setting backward hooks on ReLU activations.The hooks will b
e removed after the attribution is finished
  "Setting backward hooks on ReLU activations."

<class 'torch.Tensor'>
torch.Size([5, 3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
attributions shape: torch.Size([3, 224, 224])
```

Visualizaing layers and neurons using Captum (3 pts)

Let's try to attribute to a selected layer and visualize the attribution for any selected channel. We can choose to change the layers and channels and observe how the attribution changes.

Consider also using https://captum.ai/api/_modules/captum/attr/_utils/attribution.html#LayerAttribution.interpolate (https://captum.ai/api/_modules/captum/attr/_utils/attribution.html#LayerAttribution.interpolate) to interpolate layer dimensions to given input dimensions.

Please, be aware of the memory limitation and how you can overcome those using the techinques described in the previous section.

In [18]:

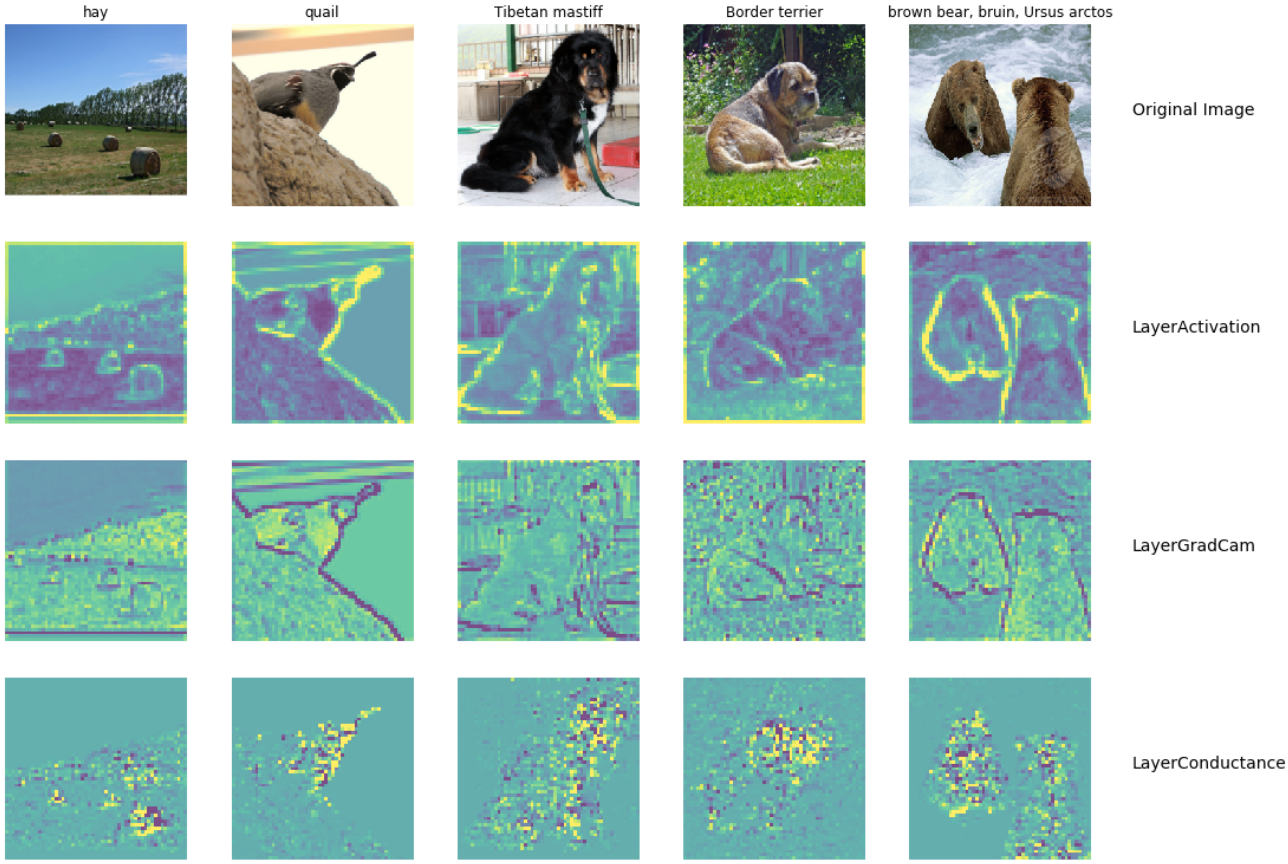
```
from captum.attr import LayerActivation, LayerConductance, LayerGradCam
layer = model.features[3]

attr1 = compute_attributions(LayerConductance(model, layer), X_tensor, target=y_tensor)
attr2 = compute_attributions(LayerGradCam(model, layer), X_tensor, target=y_tensor)
print(captum.__version__)
```

```
0.2.0
```

In [19]:

```
from captum.attr import LayerActivation, LayerConductance, LayerGradCam

# Try out different layers and see observe how the attributions change
layer = model.features[3]

layer_act = LayerActivation(model, layer)
layer_act_attr = compute_attributions(layer_act, X_tensor)
layer_act_attr_sum = layer_act_attr.mean(axis=1, keepdim=True)
##############################################################################
# TODO: Visualize Individual Layer Gradcam and Layer Conductance (similar    #
# to what we did for the other captum sections, using our helper methods),   #
# but with some preprocessing calculations.                                  #
##############################################################################
from captum.attr._utils.attribution import LayerAttribution
interpolate = LayerAttribution.interpolate

def adjustShape(attr):
    print(layer_act_attr_sum.shape)
    attr = interpolate(attr, (224, 224))
    print(attr.shape)
    attr = attr.permute(0,2,3,1)
    print(attr.shape)
    attr = interpolate(attr, (224, 3))
    print(attr.shape)
    attr = attr.permute(0, 3, 1, 2)
    print(attr.shape)
    return attr

attr0 = layer_act_attr_sum
print(attr0.shape)
# print(attr1.shape)
# print(attr2.shape)
# attr0 = adjustShape(layer_act_attr_sum)
# attr1 = adjustShape(attr1)
# # attr2 = adjustShape(attr2)
# print(attr0.squeeze().shape)
# print(attr1.shape)
# print(attr2.shape)


# visualize_attr_maps([attr0], ['LayerActivation'], attr_preprocess=lambda attr: attr[0,:,:].detach().numpy())
visualize_attr_maps([attr0, attr2, attr1], ['LayerActivation', 'LayerGradCam', 'LayerConductance'], attr_preprocess=lambda attr: attr[0,:,:].detach()
.numpy())
##############################################################################
#                            END OF YOUR CODE                                #
##############################################################################
```

```
torch.Size([5, 1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([1, 55, 55])
attributions shape: torch.Size([128, 55, 55])
attributions shape: torch.Size([128, 55, 55])
attributions shape: torch.Size([128, 55, 55])
attributions shape: torch.Size([128, 55, 55])
attributions shape: torch.Size([128, 55, 55])
```



Fooling Images (10 pts)

We can also use the similar concept of image gradients to study the stability of the network. Consider a state-of-the-art deep neural network that generalizes well on an object recognition task. We expect such network to be robust to small perturbations of its input, because small perturbation cannot change the object category of an image. However, [2] find that applying an imperceptible non-random perturbation to a test image, it is possible to arbitrarily change the network's prediction.

[2] Szegedy et al, "Intriguing properties of neural networks", ICLR 2014 (https://arxiv.org/abs/1312.6199)

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. We term the so perturbed examples "adversarial examples".

Read the paper, and then implement the following function to generate fooling images.

In [20]:
```
def make_fooling_image(X, target_y, model):
    """
    Generate a fooling image that is close to X, but that the model classifies
    as target_y.

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

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

    model.eval()

    # Initialize our fooling image to the input image, and wrap it in a Variable.
    X_fooling = X.clone()
    X_fooling_var = Variable(X_fooling, requires_grad=True)

    # We will fix these parameters for everyone so that there will be
    # comparable outputs

    learning_rate = 1 # learning rate is 1. Note. changed the value from 10 to 1 to match comment.
    max_iter = 100 # maximum number of iterations

    #for it in range(max_iter):
    not_fooled = True
    it = 0
    while it < max_iter and not_fooled:
    ################################################################################
    # TODO: Generate a fooling image X_fooling that the model will classify as   #
    # the class target_y. You should perform gradient ascent on the score of the #
    # target class, stopping when the model is fooled.                           #
    # When computing an update step, first normalize the gradient:               #
    #   dX = learning_rate * g / ||g||_2                                         #
    #                                                                            #
    # Inside of this loop, write the update rule.                                #
    #                                                                            #
    # HINT:                                                                      #
    # You can print your progress (current prediction and its confidence score)  #
    # over iterations to check your gradient ascent progress.                    #
    ################################################################################
        scores = model.forward(X_fooling_var)
        init_grad = torch.zeros_like(scores).float()
        init_grad[[0], target_y] = 1.0
        scores.backward(init_grad)
#         print(f"scores shape: {scores.shape}")
        grad = X_fooling_var.grad
        X_fooling_var.data += learning_rate * grad/grad.norm(2)

        it += 1
        if target_y == scores.data.max(1)[1][0]:
            not_fooled = False
            print(f"early stop at iteration: {it - 1}")
    ################################################################################
    #                              END OF YOUR CODE                                #
    ################################################################################

    X_fooling = X_fooling_var.data

    return X_fooling
```

Now you can run the following cell to **generate a fooling image**. You will see the message 'Fooled the model' when you succeed.

In [21]:
```
idx = 0
target_y = 6 # target label. Change to a different label to see the difference.

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(Variable(X_fooling))

if target_y == scores.data.max(1)[1][0]:
    print('Fooled the model!')
else:
    print('The model is not fooled!')
```

```
early stop at iteration: 10
Fooled the model!
```

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

In [22]:
```
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()
```

Class visualization (10 pts)

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 [1]; [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.

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014 (https://arxiv.org/abs/1312.6034)

[3] Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML 2015 Deep Learning Workshop (http://yosinski.com/media/papers/Yosinski__2015__ICML_DL__Understanding_Neural_Networks_Through_Deep_Visualization__.pdf)

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

In [23]:
```
def jitter(X, ox, oy):
    """
    Helper function to randomly jitter an image.

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

    Returns: A new PyTorch Tensor of shape (N, C, H, W)
    """
    if ox != 0:
        left = X[:, :, :, :-ox]
        right = X[:, :, :, -ox:]
        X = torch.cat([right, left], dim=3)
    if oy != 0:
        top = X[:, :, :-oy]
        bottom = X[:, :, -oy:]
        X = torch.cat([bottom, top], dim=2)
    return X
```

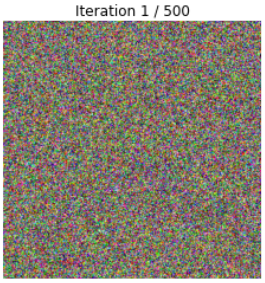
In [24]:
```
def create_class_visualization(target_y, model, dtype, **kwargs):
    """
    Generate an image to maximize the score of target_y under a pretrained model.

    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.eval()

    model.type(dtype)
    l2_reg = kwargs.pop('l2_reg', 1e-9) #original 1e-3
    learning_rate = kwargs.pop('learning_rate', 0.1) #original 25
    num_iterations = kwargs.pop('num_iterations', 500)#1000) #original 100
    blur_every = kwargs.pop('blur_every', 10) #original 10
    max_jitter = kwargs.pop('max_jitter', 16) #original 16
    show_every = kwargs.pop('show_every', 100) #original 25

    # Randomly initialize the image as a PyTorch Tensor, and also wrap it in
    # a PyTorch Variable.
    img = torch.randn(1, 3, 224, 224).mul_(1.0).type(dtype)
    img_var = Variable(img, requires_grad=True)

    for t in range(num_iterations):
        # Randomly jitter the image a bit; this gives slightly nicer results
        ox, oy = random.randint(0, max_jitter), random.randint(0, max_jitter)
        img.copy_(jitter(img, ox, oy))

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

#         init_grad = torch.zeros_like(scores).float()
#         init_grad[[0], target_y] = 1.0
#         scores.backward(init_grad)

        specific_class_score = torch.Tensor([target_y]).long()

        scores = scores.gather(1, specific_class_score.view(-1, 1)).squeeze()
        scores.backward(torch.ones_like(scores)) #do the backpass

        grad = img_var.grad - 2 * l2_reg * img_var#.norm(2)
        img_var.data += learning_rate * grad #ASCENT
#         img_var.data -= learning_rate * (img_var.grad) - l2_reg * img_var.norm(2)
        #grad = img_var.grad + 2 * l2_reg * img_var.norm(2)
        #img_var.data += learning_rate * grad
        #img_var.data += learning_rate * grad - learning_rate * l2_reg * img_var.norm(2) * img_var.norm(2)
        #img_var.data += learning_rate * grad - l2_reg * img_var.norm(2)


        ########################################################################
        #                             END OF YOUR CODE                         #
        ########################################################################

        # Undo the random jitter
        img.copy_(jitter(img, -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[:, c].clamp_(min=lo, max=hi)
        if t % blur_every == 0:
            blur_image(img, sigma=0.5)

        # Periodically show the image
        if t == 0 or (t + 1) % show_every == 0 or t == num_iterations - 1:
#             print(f"score of target class: {scores[[0], target_y].data}")
#             print(f"target class {target_y}")
#             print(f"class with max score {scores.data.max(1)}")
            scores = model(Variable(X_fooling))
            plt.imshow(deprocess(img.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.cpu())
```

Once you have completed the implementation in the cell above, run the following cell to generate images of several classes. Show the generated images when you submitted your notebook.

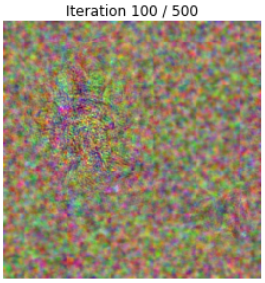
In [25]:

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

# You can use a single class during your debugging session,
# but please show all the generated outputs in your submitted notebook

# 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)


targets = [76, 78, 187, 683, 366, 604]

for target in targets:
    out = create_class_visualization(target, model, dtype)
```

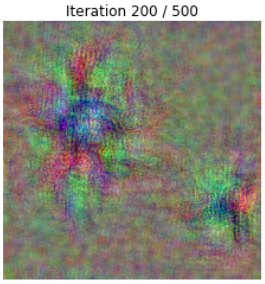
tarantula
Iteration 1 / 500



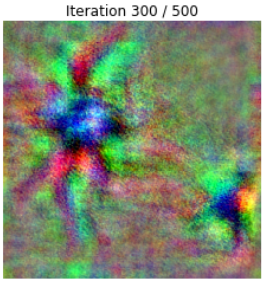
tarantula
Iteration 100 / 500



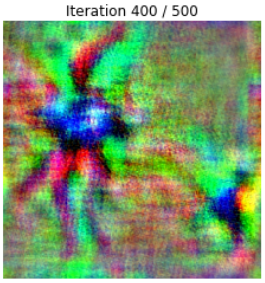
tarantula
Iteration 200 / 500



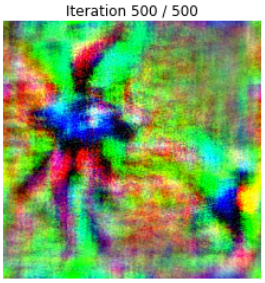
tarantula
Iteration 300 / 500



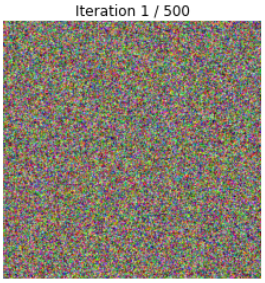
tarantula
Iteration 400 / 500



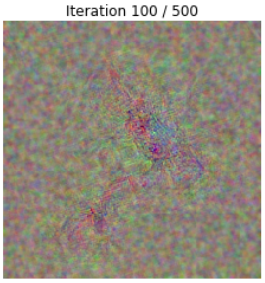
tarantula
Iteration 500 / 500



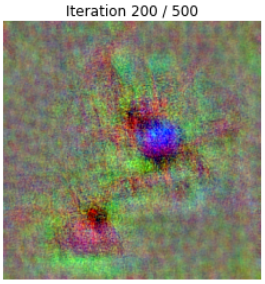
tick
Iteration 1 / 500



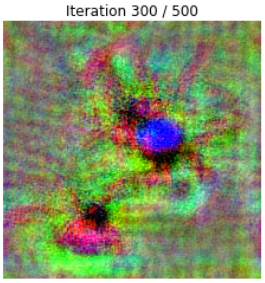
tick
Iteration 100 / 500



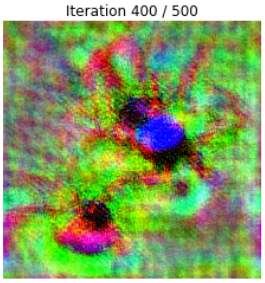
tick
Iteration 200 / 500



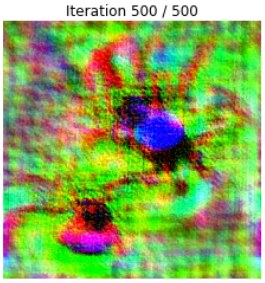
tick
Iteration 300 / 500



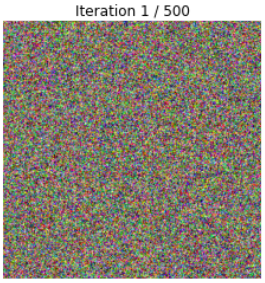
tick
Iteration 400 / 500



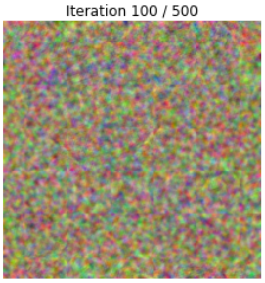
tick
Iteration 500 / 500



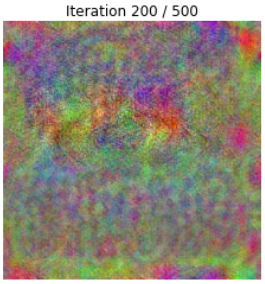
Yorkshire terrier
Iteration 1 / 500



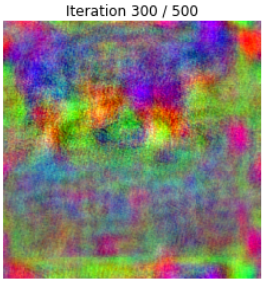
Yorkshire terrier
Iteration 100 / 500



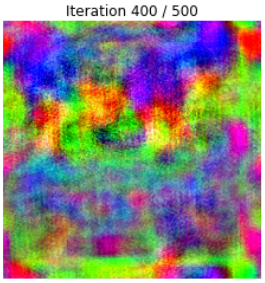
Yorkshire terrier
Iteration 200 / 500



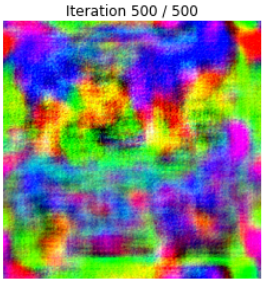
Yorkshire terrier
Iteration 300 / 500



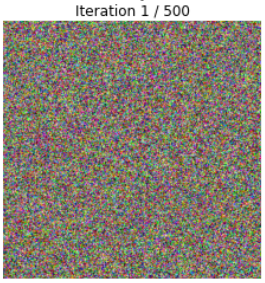
Yorkshire terrier
Iteration 400 / 500



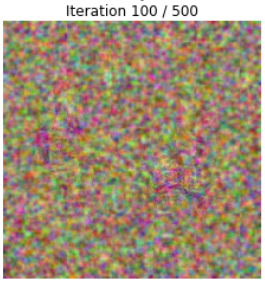
Yorkshire terrier
Iteration 500 / 500



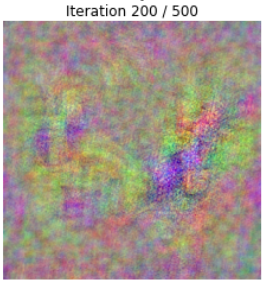
oboe, hautboy, hautbois
Iteration 1 / 500



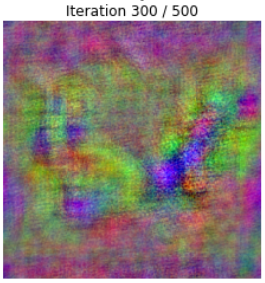
oboe, hautboy, hautbois
Iteration 100 / 500



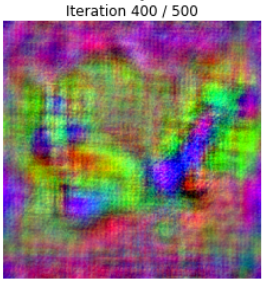
oboe, hautboy, hautbois
Iteration 200 / 500



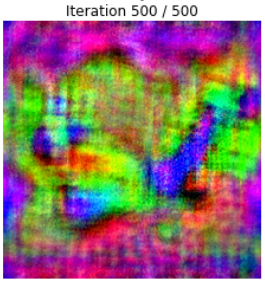
oboe, hautboy, hautbois
Iteration 300 / 500



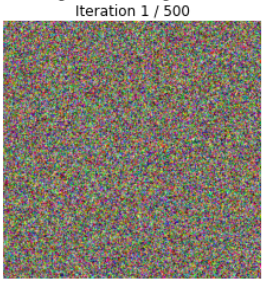
oboe, hautboy, hautbois
Iteration 400 / 500



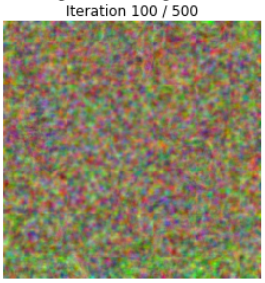
oboe, hautboy, hautbois
Iteration 500 / 500



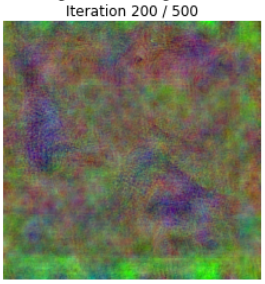
gorilla, Gorilla gorilla
Iteration 1 / 500



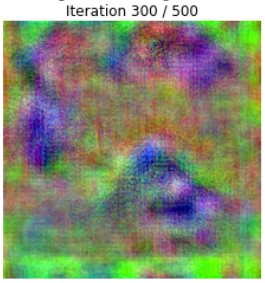
gorilla, Gorilla gorilla
Iteration 100 / 500



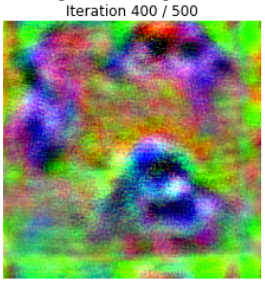
gorilla, Gorilla gorilla
Iteration 200 / 500



gorilla, Gorilla gorilla
Iteration 300 / 500

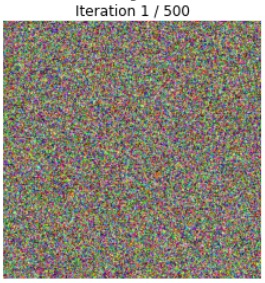


gorilla, Gorilla gorilla
Iteration 400 / 500

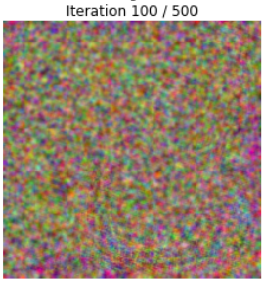


gorilla, Gorilla gorilla
Iteration 500 / 500

hourglass
Iteration 1 / 500

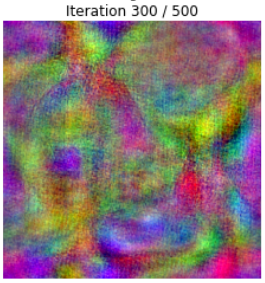


hourglass
Iteration 100 / 500

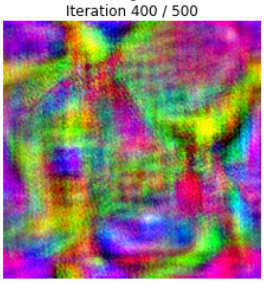


hourglass
Iteration 200 / 500

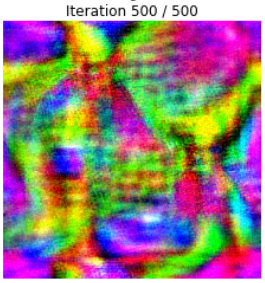
hourglass
Iteration 300 / 500



hourglass
Iteration 400 / 500



hourglass
Iteration 500 / 500



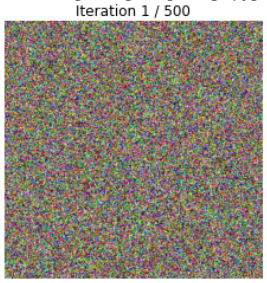
Try out your class visualization on other classes! You should also feel free to play with various hyperparameters to try and improve the quality of the generated image, but this is not required.

In [26]:

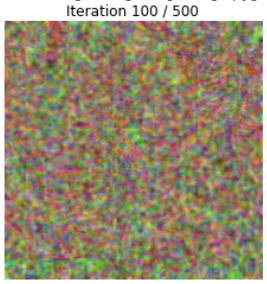
```
# target_y = 78 # Tick
# target_y = 187 # Yorkshire Terrier
# target_y = 683 # Oboe
# target_y = 366 # Gorilla
# target_y = 604 # Hourglass
target_y = np.random.randint(1000)
print(class_names[target_y])
out = create_class_visualization(target_y, model, dtype)
```

```
orangutan, orang, orangutang, Pongo pygmaeus
```

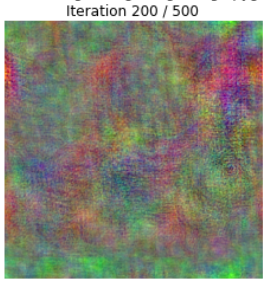
orangutan, orang, orangutang, Pongo pygmaeus
Iteration 1 / 500



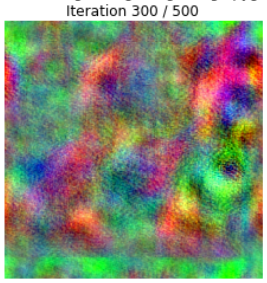
orangutan, orang, orangutang, Pongo pygmaeus
Iteration 100 / 500



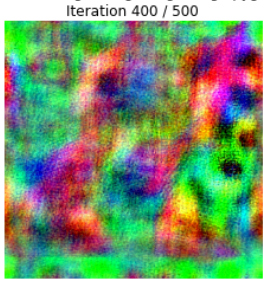
orangutan, orang, orangutang, Pongo pygmaeus
Iteration 200 / 500



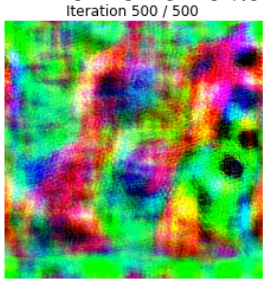
orangutan, orang, orangutang, Pongo pygmaeus
Iteration 300 / 500



orangutan, orang, orangutang, Pongo pygmaeus
Iteration 400 / 500



orangutan, orang, orangutang, Pongo pygmaeus
Iteration 500 / 500



In [ ]:

Style Transfer (20 Points)

Another task closely related to image gradient is style transfer. This has become a cool application in deep learning with computer vision. In this notebook we will study and implement the style transfer technique from:

- "Image Style Transfer Using Convolutional Neural Networks" (Gatys et al., CVPR 2015) (http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf).

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

In this notebook, we will also use SqueezeNet (https://arxiv.org/abs/1602.07360) as our feature extractor which can easily work on a CPU machine. Similarly, if computational resources are not any problem for you, you are encouraged to try a larger network, which may give you benefits in the visual output in this homework.

Note for grading:

- The total credits for this notebook are 20 points. For each of the loss function, **you will need to pass the unit test to receive full credits, otherwise it will be 0.** For the final output you will be expected to generate the images similar to the output to receive the full credits.
- Although we will not run your notebook in grading, you still need to **submit the notebook with all the outputs you generated**. Sometimes it will inform us if we get any inconsitent results with respect to yours.

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

Excited? Let's get started!

First, run the setup cells which provide the utility functions you will need later.

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

import numpy as np

from scipy.misc import imread
from collections import namedtuple
import matplotlib.pyplot as plt

from cs7643.image_utils import SQUEEZENET_MEAN, SQUEEZENET_STD
%matplotlib inline
```

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

In [2]:
```
def preprocess(img, size=512):
    transform = T.Compose([
        T.Resize(size),
        T.ToTensor(),
        T.Normalize(mean=SQUEEZENET_MEAN.tolist(),
                    std=SQUEEZENET_STD.tolist()),
        T.Lambda(lambda x: x[None]),
    ])
    return transform(img)

def deprocess(img):
    transform = T.Compose([
        T.Lambda(lambda x: x[0]),
        T.Normalize(mean=[0, 0, 0], std=[1.0 / s for s in SQUEEZENET_STD.tolist()]),
        T.Normalize(mean=[-m for m in SQUEEZENET_MEAN.tolist()], std=[1, 1, 1]),
        T.Lambda(rescale),
        T.ToPILImage(),
    ])
    return transform(img)

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

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

def features_from_img(imgpath, imgsize):
    img = preprocess(PIL.Image.open(imgpath), size=imgsize)
    img_var = Variable(img.type(dtype))
    return extract_features(img_var, cnn), img_var

# Older versions of scipy.misc.imresize yield different results
# from newer versions, so we check to make sure scipy is up to date.
def check_scipy():
    import scipy
    vnums = list(map(int, scipy.__version__.split('.')))
    assert vnums[1] >= 16 or vnums[0] >= 1, "You must install SciPy >= 0.16.0 to complete this notebook."

check_scipy()

answers = np.load('style-transfer-checks.npz')
```

As in the last notebook, we need to set the dtype to select either the CPU or the GPU

In [3]:
```
dtype = torch.FloatTensor
# Uncomment out the following line if you're on a machine with a GPU set up for PyTorch!
# dtype = torch.cuda.FloatTensor
```

In [4]:
```
# Load the pre-trained SqueezeNet model.
cnn = torchvision.models.squeezenet1_1(pretrained=True).features
cnn.type(dtype)

# Fix the weights of the pretrained network
for param in cnn.parameters():
    param.requires_grad = False

# We provide this helper code which takes an image, a model (cnn), and returns a list of
# feature maps, one per layer.
def extract_features(x, cnn):
    """
    Use the CNN to extract features from the input image x.

    Inputs:
    - x: A PyTorch Variable of shape (N, C, H, W) holding a minibatch of images that
      will be fed to the CNN.
    - cnn: A PyTorch model that we will use to extract features.

    Returns:
    - features: A list of feature for the input images x extracted using the cnn model.
      features[i] is a PyTorch Variable of shape (N, C_i, H_i, W_i); recall that features
      from different layers of the network may have different numbers of channels (C_i) and
      spatial dimensions (H_i, W_i).
    """
    features = []
    prev_feat = x
    for i, module in enumerate(cnn._modules.values()):
        next_feat = module(prev_feat)
        features.append(next_feat)
        prev_feat = next_feat
    return features
```

Implementation: Computing Loss

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

Content loss (3 pts)

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

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

Then the content loss is given by:

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

In [5]:
```
def content_loss(content_weight, content_current, content_original):
    """
    Compute the content loss for style transfer.

    Inputs:
    - content_weight: Scalar giving the weighting for the content loss.
    - content_current: features of the current image; this is a PyTorch Tensor of shape
      (1, C_l, H_l, W_l).
    - content_target: features of the content image, Tensor with shape (1, C_l, H_l, W_l).

    Returns:
    - scalar content loss
    """

    ################################################################################
    # TODO: Implement content loss function                                        #
    # Please pay attention to use torch tensor math function to finish it.         #
    # Otherwise, you may run into the issues later that dynamic graph is broken    #
    # and gradient can not be derived.                                             #
    ################################################################################
#     print(f'content_original shape = {content_original.shape}')
#     print(f'content_current shape = {content_current.shape}')
#     print(f'content_weight = {content_weight}')
    diff = content_current - content_original
    diff_sqr = diff.pow(2)
    loss = content_weight * diff_sqr.sum()
    return loss
    ################################################################################
    #                             END OF YOUR CODE                                 #
    ################################################################################
```

Test your content loss function. You should see errors less than 0.001 (normally it should be exactly 0).

In [6]:
```
def content_loss_test(correct):
    content_image = 'styles/tubingen.jpg'
    image_size =  192
    content_layer = 3
    content_weight = 6e-2

    c_feats, content_img_var = features_from_img(content_image, image_size)

    bad_img = Variable(torch.zeros(*content_img_var.data.size()))
    feats = extract_features(bad_img, cnn)
#     print(f'correct shape = {correct.shape}')
    student_output = content_loss(content_weight, c_feats[content_layer], feats[content_layer]).data.numpy()
    error = rel_error(correct, student_output)
    print('Maximum error is {:.3f}'.format(error))

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

```
Maximum error is 0.000
```

Style loss (3 pts for Gram matrix + 3 pts for loss)

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

First, compute the Gram matrix G which represents the correlations between the responses of each filter, where F is as above. The Gram matrix is an approximation to the covariance matrix -- we want the activation statistics of our generated image to match the activation statistics of our style image, and matching the (approximate) covariance is one way to do that. There are a variety of ways you could do this, but the Gram matrix is nice because it's easy to compute and in practice shows good results.

Given a feature map F^ℓ of shape $(1, C_\ell, M_\ell)$, the Gram matrix has shape $(1, C_\ell, C_\ell)$ and its elements are given by:

$$G^\ell_{ij} = \sum_k F^\ell_{ik} F^\ell_{jk}$$

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

$$L^\ell_s = w_\ell \sum_{i,j} \left(G^\ell_{ij} - A^\ell_{ij}\right)^2$$

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

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

Begin by implementing the Gram matrix computation below:

In [7]:
```
def gram_matrix(features, normalize=True):
    """
    Compute the Gram matrix from features.

    Inputs:
    - features: PyTorch Variable of shape (N, C, H, W) giving features for
      a batch of N images.
    - normalize: optional, whether to normalize the Gram matrix
        If True, divide the Gram matrix by the number of neurons (H * W * C)

    Returns:
    - gram: PyTorch Variable of shape (N, C, C) giving the
      (optionally normalized) Gram matrices for the N input images.
    """
    ################################################################################
    # TODO: Implement content loss function                                        #
    # Please pay attention to use torch tensor math function to finish it.         #
    # Otherwise, you may run into the issues later that dynamic graph is broken    #
    # and gradient can not be derived.                                             #
    #                                                                              #
    # HINT: you may find torch.bmm() function is handy when it comes to process    #
    # matrix product in a batch. Please check the document about how to use it.    #
    ################################################################################
#     print(f'features shape = {features.shape}')
    [N, C, H, W] = features.shape
    #vectorized the feature maps
    features_vectorized = features.view(N,C,H*W)
    #transpose, [N,C,H*W] to [N,H*W,C] such that we can use matrix multiplication
    features_vec_trans = features_vectorized.transpose(1,2)
    gram_mat = torch.bmm(features_vectorized,features_vec_trans)
    if normalize:
        divisor = C*H*W
        gram_mat /= divisor
    return gram_mat
    ################################################################################
    #                               END OF YOUR CODE                               #
    ################################################################################
```

Test your Gram matrix code. You should see errors less than 0.001 (normally it should be exactly 0).

In [8]:
```
def gram_matrix_test(correct):
    style_image = 'styles/starry_night.jpg'
    style_size = 192
    feats, _ = features_from_img(style_image, style_size)
#     print(f'correct shape = {correct.shape}')
    student_output = gram_matrix(feats[5].clone()).data.numpy()
    error = rel_error(correct, student_output)
#     print('Maximum error is {:.3f}'.format(error))

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

Next, implement the style loss:

In [9]:
```
# Now put it together in the style_loss function...
def style_loss(feats, style_layers, style_targets, style_weights):
    """
    Computes the style loss at a set of layers.

    Inputs:
    - feats: list of the features at every layer of the current image, as produced by
      the extract_features function.
    - style_layers: List of layer indices into feats giving the layers to include in the
      style loss.
    - style_targets: List of the same length as style_layers, where style_targets[i] is
      a PyTorch Variable giving the Gram matrix the source style image computed at
      layer style_layers[i].
    - style_weights: List of the same length as style_layers, where style_weights[i]
      is a scalar giving the weight for the style loss at layer style_layers[i].

    Returns:
    - style_loss: A PyTorch Variable holding a scalar giving the style loss.
    """

    ################################################################################
    # TODO: Implement content loss function                                        #
    # Please pay attention to use torch tensor math function to finish it.         #
    # Otherwise, you may run into the issues later that dynamic graph is broken    #
    # and gradient can not be derived.                                             #
    #                                                                              #
    # Hint:                                                                        #
    # you can do this with one for loop over the style layers, and should not be   #
    # very much code (~5 lines). Please refer to the 'style_loss_test' for the     #
    # actual data structure.                                                       #
    #                                                                              #
    # You will need to use your gram_matrix function.                              #
    ################################################################################
#     print(len(feats))
#     for i in range(len(feats)):
#         print(f'i = {i}')
#         print(f'feats shape = {feats[i].shape}')
    total_loss = torch.zeros(1,)
    for i in range(len(style_layers)):
#         print(f'i = {i}')
#         print(f'style_layers = {style_layers[i]}')
#         print(f'style_targets shape = {style_targets[i].shape}')
#         print(f'style_weights = {style_weights[i]}')
        gram_mat = gram_matrix(feats[style_layers[i]], normalize=True)
        loss = content_loss(style_weights[i], gram_mat, style_targets[i])
        total_loss = total_loss + loss
    return total_loss

    ################################################################################
    #                              END OF YOUR CODE                                #
    ################################################################################
```

Test your style loss implementation. The error should be less than 0.001 (normally it should be exactly 0).

In [10]:
```
def style_loss_test(correct):
    content_image = 'styles/tubingen.jpg'
    style_image = 'styles/starry_night.jpg'
    image_size =  192
    style_size = 192
    style_layers = [1, 4, 6, 7]
    style_weights = [300000, 1000, 15, 3]

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

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


style_loss_test(answers['sl_out'])
```

```
Error is 0.000
```

Total-variation regularization (3 pts)

It turns out that it's helpful to also encourage smoothness in the image. We can do this by adding another term to our loss that penalizes wiggles or **total variation** in the pixel values. This concept is widely used in many computer vision task as a regularization term.

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

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

You may not see this loss function in this particular reference paper, but you should be able to implement it based on this equation. In the next cell, fill in the definition for the TV loss term.

You need to provide an efficient vectorized implementation to receive the full credit, your implementation should not have any loops. Otherwise, penalities will be given according to the actual implementation.

In [11]:
```
def tv_loss(img, tv_weight):
    """
    Compute total variation loss.

    Inputs:
    - img: PyTorch Variable of shape (1, 3, H, W) holding an input image.
    - tv_weight: Scalar giving the weight w_t to use for the TV loss.

    Returns:
    - loss: PyTorch Variable holding a scalar giving the total variation loss
      for img weighted by tv_weight.
    """

    ################################################################################
    # TODO: Implement content loss function                                        #
    # Please pay attention to use torch tensor math function to finish it.         #
    # Otherwise, you may run into the issues later that dynamic graph is broken    #
    # and gradient can not be derived.                                             #
    ################################################################################
#     print(f'img shape = {img.shape}')
#     print(f'tv_weight = {tv_weight}')
    [N, C, H, W] = img.shape

    loss = torch.zeros(1,)
    for n in range(N):
        for c in range(C):
            curr = img[n][c]
#             print(f'curr shape = {curr.shape}')
#             print(f'curr dim names = {curr.names}')

            curr_h0 = curr[0:H-1,:]
            curr_h1 = curr[1:H,:]
#             print(f'curr_h0 shape = {curr_h0.shape}')

            loss = loss + content_loss(tv_weight, curr_h1, curr_h0)

            curr_w0 = curr[:,0:W-1]
            curr_w1 = curr[:,1:W]
#             print(f'curr_w0 shape = {curr_w0.shape}')

            loss = loss + content_loss(tv_weight, curr_w1, curr_w0)

    return loss
    ################################################################################
    #                             END OF YOUR CODE                                 #
    ################################################################################
```

Test your TV loss implementation. Error should be less than 0.001 (normally it should be exactly 0).

In [12]:
```
def tv_loss_test(correct):
    content_image = 'styles/tubingen.jpg'
    image_size =  192
    tv_weight = 2e-2

    content_img = preprocess(PIL.Image.open(content_image), size=image_size)
    content_img_var = Variable(content_img.type(dtype))

    student_output = tv_loss(content_img_var, tv_weight).data.numpy()
    error = rel_error(correct, student_output)
    print('Error is {:.3f}'.format(error))

tv_loss_test(answers['tv_out'])
```

```
Error is 0.000
```

Implement style transfer (6 pts)

You have implemented all the loss functions in the paper. Now we're ready to string it all together. Please read the entire function: figure out what are all the parameters, inputs, solvers, etc. **The update rule in the following block is hold out for you to finish.**

In [13]:
```
def style_transfer(content_image, style_image, image_size, style_size, content_layer, content_weight,
                   style_layers, style_weights, tv_weight, init_random = False):
    """
    Run style transfer!

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

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

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

    # Initialize output image to content image or noise
    if init_random:
        img = torch.Tensor(content_img.size()).uniform_(0, 1)
    else:
        img = content_img.clone().type(dtype)

    # We do want the gradient computed on our image!
    img_var = Variable(img, requires_grad=True)

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

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

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

    for t in range(200):
        if t < 190:
            img.clamp_(-1.5, 1.5)
        feats = extract_features(img_var, cnn)

        ################################################################################
        # TODO: Implement this update rule with by forwarding it to criterion        #
        # functions and perform the backward update.                                 #
        #                                                                            #
        # HINTS: all the weights, loss functions are defined. You don't need to add  #
        # any other extra weights for the three loss terms.                          #
        # The optimizer needs to clear its grad before backward in every step.       #
        ################################################################################
        optimizer.zero_grad()
        loss = content_loss(content_weight, feats[content_layer].clone(), content_target)
        loss = loss + style_loss(feats, style_layers, style_targets, style_weights)
        loss = loss + tv_loss(img_var, tv_weight)
        loss.backward()
        optimizer.step()

        ################################################################################
        #                             END OF YOUR CODE                                 #
        ################################################################################

        if t % 100 == 0:
            print('Iteration {}'.format(t))
            plt.axis('off')
            plt.imshow(deprocess(img.cpu()))
            plt.show()
    print('Iteration {}'.format(t))
    plt.axis('off')
    plt.imshow(deprocess(img.cpu()))
    plt.show()
```

Generate some pretty pictures!

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

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

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

In [14]:
```
# Composition VII + Tubingen
params1 = {
    'content_image' : 'styles/tubingen.jpg',
    'style_image' : 'styles/composition_vii.jpg',
    'image_size' : 192,
    'style_size' : 512,
    'content_layer' : 3,
    'content_weight' : 5e-2,
    'style_layers' : (1, 4, 6, 7),
    'style_weights' : (20000, 500, 12, 1),
    'tv_weight' : 5e-2
}

style_transfer(**params1)
```

Content Source Img.

Style Source Img.

```
Iteration 0
```

```
Iteration 100
```

```
Iteration 199
```

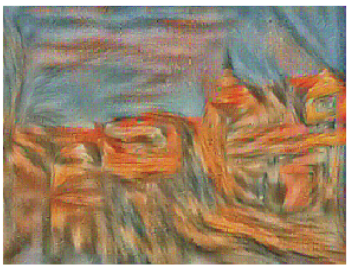
In [15]:
```
# Scream + Tubingen
params2 = {
    'content_image':'styles/tubingen.jpg',
    'style_image':'styles/the_scream.jpg',
    'image_size':192,
    'style_size':224,
    'content_layer':3,
    'content_weight':3e-2,
    'style_layers':[1, 4, 6, 7],
    'style_weights':[200000, 800, 12, 1],
    'tv_weight':2e-2
}

style_transfer(**params2)
```

Content Source Img.

Style Source Img.

Iteration 0

Iteration 100



Iteration 199

In [16]:
```
# Starry Night + Tubingen
params3 = {
    'content_image' : 'styles/tubingen.jpg',
    'style_image' : 'styles/starry_night.jpg',
    'image_size' : 192,
    'style_size' : 192,
    'content_layer' : 3,
    'content_weight' : 6e-2,
    'style_layers' : [1, 4, 6, 7],
    'style_weights' : [300000, 1000, 15, 3],
    'tv_weight' : 2e-2
}

style_transfer(**params3)
```

Content Source Img.

Style Source Img.

```
Iteration 0
```

```
Iteration 100
```

```
Iteration 199
```

Feature Inversion (Just run it, 2 pts)

The code you've written can do another cool thing. In an attempt to understand the types of features that convolutional networks learn to recognize, a recent paper [2] attempts to reconstruct an image from its feature representation. We can easily implement this idea using image gradients from the pretrained network, which is exactly what we did above (but with two different feature representations).

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

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

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

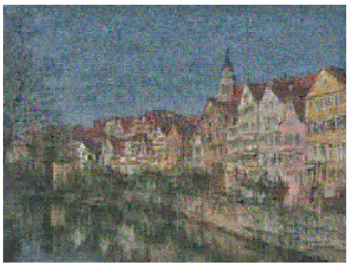
In [17]:
```
# Feature Inversion -- Starry Night + Tubingen
params_inv = {
    'content_image' : 'styles/tubingen.jpg',
    'style_image' : 'styles/starry_night.jpg',
    'image_size' : 192,
    'style_size' : 192,
    'content_layer' : 3,
    'content_weight' : 6e-2,
    'style_layers' : [1, 4, 6, 7],
    'style_weights' : [0, 0, 0, 0], # we discard any contributions from style to the loss
    'tv_weight' : 2e-2,
    'init_random': True # we want to initialize our image to be random
}

style_transfer(**params_inv)
```

Content Source Img.

Style Source Img.

Iteration 0

Iteration 100



Iteration 199

In [ ]: