Thesis

A Dissertation

Presented to

The Faculty of the Graduate School of Arts and Sciences

Brandeis University

Michtom School of Computer Science

James A. Storer, Advisor

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Aaditya Prakash

May, 2019

This dissertation, directed and approved by Aaditya Prakash's committee, has been accepted and approved by the Graduate Faculty of Brandeis University in partial fulfillment of the requirements for the degree of:

DOCTOR OF PHILOSOPHY

Eric Chasalow, Dean of Arts and Sciences

Dissertation Committee:

James A. Storer, Chair

Antonella DiLillo

Sadid Hasan

Abstract

Thesis

A dissertation presented to the Faculty of
the Graduate School of Arts and Sciences of
Brandeis University, Waltham, Massachusetts

by Aaditya Prakash

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Introduction to Deep Learning

1.2 Convolutional Neural Networks

1.2.1 Object localization

1.3 Limitations of Convolutional Networks

1.3.1 Adversarial Images

1.3.2 Overalapping Filters

Chapter 2

Semantic Image Compression

2.1 Introduction

Several attempts have been made to improve upon the lossy image compression offered by JPEG [GPG12] [Tod+16]. Despite these efforts, JPEG continues to be the standard image file format on the web. Because of the status of JPEG as the default standard and its wide adoption, it seems unlikely that the new formats will get any traction. We propose an image compression technique to improve the visual quality of standard JPEG by using a higher bitrate to encode image regions flagged by our model as containing an object of interest and lowering the bitrate elsewhere in the image. The compressed output of our method can be decoded by standard JPEG implementations.

The JPEG algorithm uses a scaling factor Q in order to scale the quantization matrix to achieve a variety of compression ratios. However, this ratio is an image level property and all the 8×8 blocks of a given image are compressed using the same scaling factor. Natural images are heterogeneous with respect to frequency, and contain both regions of primarily low frequencies and of primarily high frequencies. While low frequency regions are more tolerant

of higher compression ratios, high frequency regions are not [CE99]. Therefore, variable quantization in JPEG would aim to provide optimum perceptual quality across the image by compressing different blocks at different ratios. Previous approaches to use variable quantization have employed DCT analysis on each block to determine the frequency components [KT98] as well as classification of each block according to a pre-determined look-up table [MT00]. Soon et al [TPN96] proposed classifying blocks as textures, edges or flat regions, and adjusting the quantization matrix for each block accordingly. Adaptive Quantization techniques have also been applied to video coding [Xia+14].

These methods are limited in their ability to improve the perceptual quality of an image because frequency analysis and image metrics do not correlate with human perception [KSC92]. Therefore, we propose variable quantization of JPEG in which the choice of scaling factor is informed by semantic knowledge of the image. Human vision naturally focuses on familiar semantic objects, and is particularly sensitive to distortions of these objects as compared to distortions of background details. Our goal is to develop a technique which improves the visual quality of an image by improving the signal to noise ratio within these semantic objects while keeping the overall visual quality close to that of standard JPEG. Measuring visual quality is an ongoing research area and there is no consensus among researchers on the proper metric. We evaluate our model on a variety of metrics, `SSIM`[Wan+04], `MS-SSIM`[WSB03], `VIFP`[SB06], `PSNR-HVS`[Egi+06] and `PSNR-HVSM`[Pon+07]; these robust metrics have been shown to correlate better with subjective quality measurement than pixel-level metrics such as `MSE` or `PSNR` [Pon+07]. Yuri et al [KVR15] showed that `PSNR` as an image comparison metric has severe limitations.

Convolutional Neural Networks (CNNs) have been successfully applied to a variety of computer vision tasks [He+16] [KSH12]. Their feature extraction and transfer learning capabilities are now well known[ZF14] . CNNs, well known for their ability to classify images

by their most prominent object, and have also been used to draw a bounding box around that object [Gir+14]. This method is capable of providing a binary map for the presence of the most salient object. Some success has been obtained in predicting the visual saliency map of a given image [Jia+15] [KTB14].

We propose a deep convolution network designed to locate several semantic objects within a single image. Our model differs from traditional object detection models like [Dai+16] [Gir+14] as these models are restricted to detecting a single salient object in an image, typically an instance of a pre-computed class for that image. Previous work has shown that semantic object detection has a variety of advantages over saliency maps [MHG+14] [Zün+13]. Semantic detection models recognize discrete objects and are thus able to generate maps that are more coherent for human perception. Visual saliency models are based on human eye fixations, and thus produce results which do not capture object boundaries. This is particularly evident in the results obtained by Stella et al [SL09], in which image compression is guided by a multi-scale saliency map, and the obtained images show blurred edges and soft focus. Our proposed model captures the structure of the depicted scene and thus maintains the integrity of semantic objects, unlike results produced using human eye fixations [Liu+15].

Our proposed model produces a single class-invariant feature map by learning separate feature maps for each of a set of object classes and then summing over the top features. Our model trades lower confidence of bounding edges for the ability to find all salient regions of an image in a single pass, as opposed to standard object-detection CNNs, which require multiple passes over the image to identify and locate all the objects. In comparison to grid-based features as described by Yuri et al [Rez+13] our features are scale- and transformation-invariant, which allows application of the model to a wider class of images.

We employ a Convolutional Neural Network (CNN) tailored to the specific task of semantic

image understanding to achieve higher visual quality in lossy image compression. We focus on the JPEG standard, which remains the dominant image representation on the internet and in consumer electronics. Several attempts have been made to improve upon its lossy image compression, for example WebP [GPG12] and Residual-GRU [Tod+16], but many of these require custom decoders and are not sufficiently content-aware.

We improve the visual quality of standard JPEG by using a higher bit rate to encode image regions flagged by our model as containing content of interest and lowering the bit rate elsewhere in the image. With our enhanced JPEG encoder, the quantization of each region is informed by knowledge of the image content. Human vision naturally focuses on familiar objects, and is particularly sensitive to distortions of these objects as compared to distortions of background details [Jia+15]. By improving the signal-to-noise ratio within multiple regions of interest, we improve visual quality of those regions, while preserving overall `PSNR` and compression ratio. A second encoding pass produces a final JPEG encoding that may be decoded with any standard off-the-shelf JPEG decoder. Measuring visual quality is an ongoing area of research and there is no consensus among researchers on the proper metric. Yuri et al [KVR15] showed that `PSNR` has severe limitations as an image comparison metric. Richter et al [RK09][Ric11] addressed structural similarity (`SSIM`[Wan+04] and `MS-SSIM`[WSB03]) for JPEG and JPEG 2000. We evaluate on these metrics, as well as `VIFP`[SB06], `PSNR-HVS`[Egi+06] and `PSNR-HVSM`[Pon+07], which have been shown to correlate with subjective visual quality. Figure 2.1 compares close-up views of a salient object in a standard JPEG and our new content-aware method.

Figure 2.1: Comparison of compression of semantic objects in standard JPEG[left] and our model [right]

CNNs have been successfully applied to a variety of computer vision tasks [KSH12]. The feature extraction and transfer learning capabilities of CNNs are well known [ZF14], as are their ability to classify images by their most prominent object [He+16], and compute a bounding box [Gir+14]. Some success has been obtained in predicting the visual saliency map of a given image [Jia+15],[KTB14]. Previous work has shown that semantic object detection has a variety of advantages over saliency maps [MHG+14], [Zün+13]. Semantic detection recognizes discrete objects and is thus able to generate maps that are more coherent for human perception. Visual saliency models are based on human eye fixations, and thus produce results which do not capture object boundaries [KTB14]. This is evident in the results obtained by Stella et al [SL09], in which image compression is guided by a multi-scale saliency map, and the obtained images show blurred edges and soft focus.

We present a CNN designed to locate multiple regions of interest (ROI) within a single image. Our model differs from traditional object detection models like [Dai+16], [Gir+14] as these models are restricted to detecting a single salient object in an image. It captures the

structure of the depicted scene and thus maintains the integrity of semantic objects, unlike results produced using human eye fixations [Liu+15]. We produce a single class-invariant feature map by learning separate feature maps for each of a set of object classes and then summing over the top features. Because this task does not require precise identification of object boundaries, our system is able to capture multiple salient regions of an image in a single pass, as opposed to standard object detection CNNs, which require multiple passes over the image to identify and locate multiple objects. Model training need only be done offline, and encoding with our model employs a standard JPEG encoder combined with efficient computation of saliency maps (over 60 images per second for 1920x1080 using a Titan X Maxwell GPU). A key advantage of our approach is that its compressed output can be decoded by any standard off-the-shelf JPEG implementation. It serves to maintain the existing decoding complexity, the primary issue for distribution of electronic media.

Section 2 reviews CNN techniques used for object localization, semantic segmentation and class activation maps. We also discuss merits of using our technique over these methods. Section 3 presents our new model which can generate a map showing multiple regions of interest. In Section 4 show how we combine this map to make JPEG semantically aware. Sections 5 presents experimental results on a variety of image datasets and metrics. Section 6 concludes with future areas for research.

2.2 Review of localization using CNNs

CNNs are multi-layered feed-forward architectures where the learned features at each level are the weights of the convolution filters to be applied to the output of the previous level. Learning is done via gradient-based optimization [LB95]. CNNs differ from fully connected neural networks in that the dimensions of the learned convolution filters are, in general,

much smaller than the dimensions of the input image, so the learned features are forced to be localized in space. Also, the convolution operation uses the same weight kernel at every image location, so feature detection is spatially invariant.

Given an image x, and a convolution filter of size $n \times n$, then a convolutional layer performs the operation shown in equation 2.1, where $\mathbf{W}$ is the learned filter.

$$y_{ij} = \sum_{a=0}^{n} \sum_{b=0}^{n} \mathbf{W}_{ab}\, x_{(i+a)(j+b)} \tag{2.1}$$

In practice, multiple filters are learned in parallel within each layer, and thus the output of a convolution layer is a 3-d feature map, where the depth represents the number of filters. The number of features in a given layer is a design choice, and may differ from layer to layer. CNNs include a max pooling [LB95] step after every or every other layer of convolution, in which the height and width of the feature map (filter response) are reduced by replacing several neighboring activations (coefficients), generally within a square window, with a single activation equal to the maximum within that window. This pooling operation is strided, but the size of the pooling window can be greater than the stride, so windows can overlap. This results in down-sampling of input data, and filters applied to such a map will have a larger receptive field (spatial support in the pixel space) for a given kernel size, thus reducing the number of parameters of the CNN model and allowing the training of much deeper networks. This does not change the depth of the feature map, but only its width and height. In practice, pooling windows are typically of size 2×2 or 4×4, with a stride of two, which reduces the number activations by 75%. CNNs apply some form of non-linear operation such as sigmoid $(1 - e^{-x})^{-1}$ or linear rectifier $max(0, x)$ on the output of each convolution operation.

Region-based CNNs use a moving window to maximize the posterior of the presence of an object [Gir15]. Faster RCNNs [Ren+15] have been proposed, but they are still computa-

tionally expensive and are limited to determining the presence or absence of a single class of object within the entire image. Moving-window methods are able to produce rectangular bounding boxes, but cannot produce object silhouettes. In contrast, recent deep learning models proposed for semantic segmentation [LSD15], [Gir+14], [Zhe+15] are very good at drawing a close border around the objects. However, these methods do not scale well to more than a small number of object categories (e.g. 20) [Eve+10]. Segmentation methods typically seek to produce a hard boundary for the object, in which every pixel is labeled as either part of the object or part of the background. In contrast, class activation mapping produces a fuzzy boundary and is therefore able to capture pixels which are not part of any particular object, but are still salient to some spatial interaction between objects. Segmentation techniques are also currently limited by the requirement for strongly-labeled data for training. Obtaining training data where the locations of all the objects in the images are tagged is expensive and not scalable [Eve+10]. Our approach only requires image-level labels of object classes, without pixel-level annotation or bounding-box localization.

In a traditional CNN, there are two fully-connected (non-convolutional) layers as the final layers of the network. The final layer has one neuron for every class in the training data, and the final step in the inference is to normalize the activations of the last layer to sum to one. The second to last layer, however, is fully connected to the last convolution layer, and a non-linearity is applied to its activations. The authors of [Oqu+15], [Zho+16] modify this second to last layer to allow for class localization. In their architecture, the second to last layer is not learned, but consists of one neuron for each feature map, which has fixed equally-weighted connections to each activation of its corresponding map. No non-linearity is applied to the outputs of these neurons, so that each activation in this layer represents the global spatial average of one feature map from the previous layer. Since the output of this layer is connected directly to the classification layer, each class will in essence learn a weight

for each feature map from the final convolution layer. Thus, given an image and a class, the classification weights for that class can be used to re-weight the layers of activations of the final convolution layer on that image. These activations can be collapsed along the feature axis to create a class activation map, spatially localizing the best evidence for that class within that image. This is used to make Class Activation Maps (CAM), which shows the probability distribution of the most likely class.

While standard max pooling combines spatially local activations for each feature independently, 'global max pooling' instead combines the activations of all features for each spatial location. If the activation of a convolution layer is of size $M \times N \times D$ where M and N are the spatial dimensions of feature activations and D denotes the number of features in the feature map, then after 'global max pooling', the size of activations becomes $M \times N$. It is important to note that global average pooling is performed in lieu of a final fully-connected layer (or perceptron), which is often used in image classification tasks [SZ14]. Fully-connected layers have dense connections across all neurons and thus do not preserve spatial locality of activations. The substitution of a global max pooling layer retains this locality property in exchange for a fractional loss of classification accuracy[Oqu+15]. Pooling of feature maps is probabilistic [LRH15], which means the most salient object is not always the one with highest activation across the feature map. If we plot the activation of each feature on the 2-dimensional map and overlay them on each other, we find significant overlap of regions of high activations. This means taking a 'global max pooling' would lead to missing some. Zhou [Zho+16] reported that they obtained better localization by performing 'global average pooling' instead of 'global max pooling'. This is because taking an average across the feature map ensures each detected object has consistently high activations across many features, instead of a single maximal activation of one feature. Figure 2.2 (c) shows an example of such a map, the equation of which is given by

$$M_c(x, y) = \sum_{d \,\epsilon\, \mathbf{D}} w_d^c \, f_d(x, y) \tag{2.2}$$

where w_d^c is the learned weight of class c for feature map d. Training for CAM minimizes the cross entropy between objects' true probability distribution over classes (all mass given to the true class) and the predicted distribution, which is obtained as

$$P(c) = \frac{\exp(\sum_{xy} M_c(x, y))}{\sum_c \exp(\sum_{xy} M_c(x, y))} \tag{2.3}$$

Since CAMs are trained to maximize posterior probability for the class, they tend to only highlight a single most prominent object. This makes them useful for studying the output of CNNs, but not well suited to more general semantic saliency, as real world images typically contain multiple objects of interest. This has practical applications in understanding and visualizing convolution neural networks but is less applicable to real world images, which typically contain multiple objects of interest. Without using standard resolution, color-space or types of object classes it is hard to use such systems to extract semantic information from the images.

2.3 Multi-Structure Region of Interest

We have developed a variant of CAM which balances the activation for multiple objects and thus does not suffer from the issues of global average pooling. Our method, *Multi-Structure Region of Interest* (MS-ROI), allows us to effectively train on localization tasks independent of the number of classes. For the purposes of semantic compression, obtaining a tight bound on the objects is not important. However, identifying and approximately locating all the objects is critical. We propose a set of 3D feature maps in which each feature map is learned

for an individual class, and is learned independently of the maps for other classes. For $\mathbf{L}$ layers, where each layer l contains d_l features, an image of size $n \times n$, and with $\mathbf{C}$ classes, this results in a total activation size of

$$\sum_{l \,\epsilon\, \mathbf{L}} d_l \times \mathbf{C} \times \frac{n}{k^l} \times \frac{n}{k^l}$$

where k is the max pooling stride size. This is computationally very expensive, and not practical for real world data. CNNs designed for full-scale color images have many filters per layer and are several layers deep. For such networks, learning a model with that many parameters would be unfeasible in terms of computational requirements. We propose two techniques to make this idea feasible for large networks:

i Reduce the number of classes and increase the inter-class variance by combining similar classes

ii Share feature maps across classes to jointly learn lower level features

Most CNN models are built for classification on the Large Scale Visual Recognition Challenge, commonly known as ImageNet. [1]. ImageNet has one thousand classes and many of the classes are fine-grained delineations of various types of animals, flowers and other objects. We significantly reduce the number of classes by collapsing these sets of similar classes to a single, more general class. This is desirable because, for the purpose of selecting a class invariant 'region of interest,' we do not care about the differences between Siberian husky and Eskimo dog or between Lace-flower and Tuberose. As long as objects of these combined classes have similar structure and are within the same general category, the map produced will be almost identical. Details of the combined classes used in our model are provided in the Experimental Results section.

[1] http://image-net.org/challenges/LSVRC/

It is self-evident that most images contain only a few classes and thus it is computationally inefficient to build a separate feature map for every class. More importantly, many classes have similar lower-level features, even when the number of classes is relatively small. The first few layers of a CNN learn filters which recognize small edges and motifs [ZF14], which are found across a wide variety of object classes. Therefore, we propose parameter sharing across the feature maps for different classes. This reduces the number of parameters and also allows for the joint learning of these shared, low-level features.

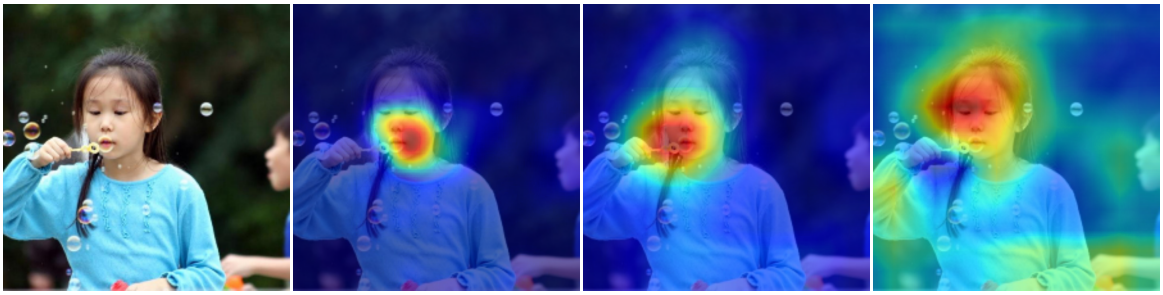

Figure 2.2: Comparison of various methods of detecting objects in an image

Although we do not restrict ourselves to a single most-probable class, it is desirable to eliminate the effects of activations for classes which are not present in the image. In order to accomplish this, we propose a thresholding operation which discards those classes whose learned features do not have a sufficiently large total activation when summed across all features and across the entire image. Let Z_l^c denote the total sum of the activations of layer ℓ for all feature maps for a given class c. Since our feature map is a 4-dimensional tensor, Z_l^c can be obtained by summation of this tensor over the three non-class dimensions.

$$Z_l^c = \sum_{d \,\epsilon\, \mathbf{D}} \sum_{x,y} f_d^c(x, y) \tag{2.4}$$

Next, we use Z_l^c to filter the classes. Computation of the multi-structure region of interest

is shown below.

$$\hat{M}(x,y) = \sum_{c \,\epsilon\, \mathbf{c}} \begin{cases} \sum_d f_d^c(x,y), & \text{if } Z_l^c > T \\ \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

We use the symbol $\hat{M}$ to denote the multi-structure map generated by our proposed model in order to contrast it with the map generated using standard CAM models, M. $\hat{M}$ is a sum over all classes with total activations Z_l^c beyond a threshold value T. T is determined during the training or chosen as a hyper-parameter for learning. In practice, it is sufficient to *argsort* Z_l^c and pick the top five classes and combine them via a sum weighted by their rank. It should be noted that, because $\hat{M}$ is no longer a reflection of the class of the image, we use the term 'region of interest'.

A comparison of our model (MS-ROI) with CAM and human fixation is shown in Figure 2.2. Only our model identifies the face of the boy on the right as well the hands of both children at the bottom. When doing compression, it is important that we do not lower the quality of body extremities or other objects which other models may not identify as critical to the primary object class of the image. If a human annotator were to paint the areas which should be compressed at better quality, we believe the annotated area would be closer to that captured by our model.

To train the model to maximize the detection of all objects, instead of using a softmax function as in equation 2.3, we use sigmoid, which does not marginalize the posterior over the classes. Thus the likelihood of a class c is given by equation 2.6.

$$P(c) = \frac{1}{1 + \exp(Z_l^c)} \tag{2.6}$$

2.4 Integrating MS-ROI map with JPEG

We obtain from MS-ROI a saliency value for each pixel in the range [0,1], where 1 indicates maximum saliency. Then discretize these saliency values into k levels, where k is a tuneable hyper-parameter. The lowest level contains pixels of saliency $[0, 1/k]$, the second level contains pixels of saliency $(1/k, 2/k]$ and so forth. We next select a range of JPEG quality levels, Q_l to Q_h. Each saliency level will be compressed using a Q value drawn from this range, corresponding to that level. In other words, saliency level n, with saliency range $[n/k, (n+1)/k]$ will be compressed using

$$Q_n = Q_l + \frac{n * (Q_h - Q_l)}{k} \tag{2.7}$$

For each level $l \leq n \leq h$, we obtain a decoded JPEG of the image after encoding at quality level Q_n. For each 8×8 block of our output image, we select the block of color values obtained by the JPEG corresponding to that block's saliency level. This mosiac of blocks is finally compressed using a standard JPEG encoder with the desired output quality to produce a file which can be decoded by any off-the-shelf JPEG decoder.

Details of our choices for k, Q_l and Q_h, as well as target image sizes are provided in the next section. A wider range of Q_l and Q_h will tend to produce stronger results, but at the

expense of very poor quality in non-salient regions.

Algorithm 1: JPEG Encoding with MSROI

Input : Image: I

Output: Image: O, same dimensions as I

Let $Q_{l,h}$ be N fixed values in range l to h

for $b \leftarrow 8 \times 8$ *blocks in* I **do**

    $Q_i = \hat{M}$[b] #Nearest quantized level

    I'[b] $\leftarrow$ JPEG {I[b], Q_i}

end

O $\leftarrow$ JPEG{I', Q_f}

Encoding of image using MSROI on a JPEG standard can be summarized as shown by algorithm 1.

Figure 2.3 shows the steps of taking an image and using the MSROI map to obtain the final encoded image.

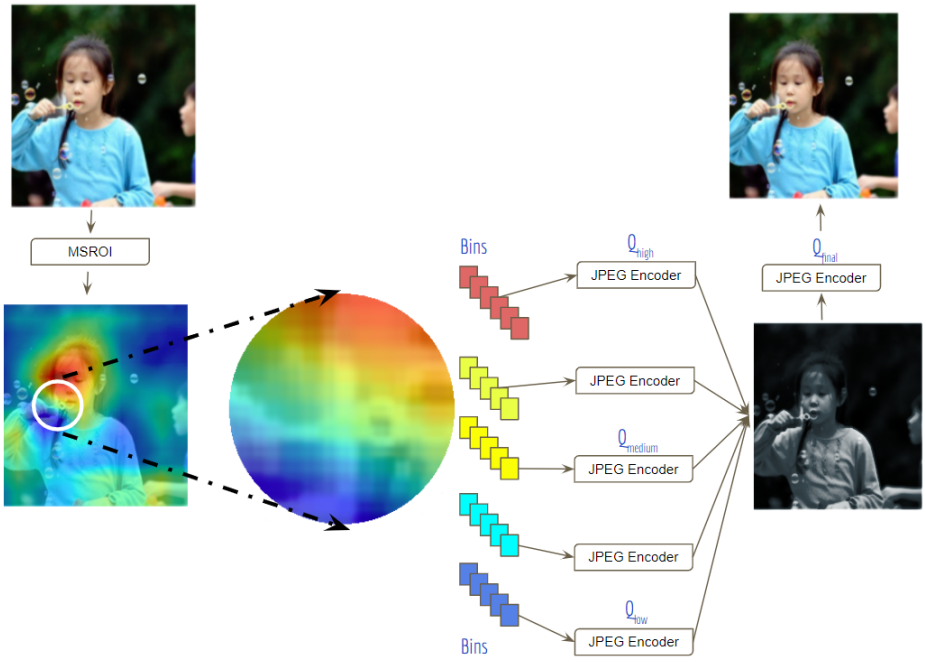

Figure 2.3: Process of combining MSROI maps to obtain final encoded image

2.5 Experimental Results

We trained our model with the Caltech-256 dataset [GHP07], which contains 256 classes of man-made and natural objects, common plants and animals, buildings, etc.

We believe this offers a good balance between covering more classes as compared to CIFAR-100 which contains only 100 classes, and avoiding overly finely-grained classes as in ImageNet with 1000 classes [Den+09]. For the results reported here, we experimented with several stacked layers of convolution as shown in the diagram below:

$$\text{IMAGE} \longmapsto \Big[\big[\text{CONV} \to \text{RELU}\big]^2 \to \text{MAXPOOL}\Big]^5 \longmapsto \text{MS-ROI} \longmapsto \text{MAP}$$

MS-ROI refers to the operation shown in the equation 2.5. To obtain the final image we

discretize the heat-map into five levels and use JPEG quality levels Q in increments of ten from $Q_l = 30$ to $Q_h = 70$. For all experiments, the file size of the standard JPEG image and the JPEG obtained from our model were kept within $\pm 1\%$ of each other. On average, salient regions were compressed at $Q_f = 65$, and non-salient regions were compressed at $Q = 45$. The average Q for the final image generated using our model was 55, whereas for all standard JPEG samples, Q was chosen to be 50. It should be noted that even though images are at different Q they have same size.

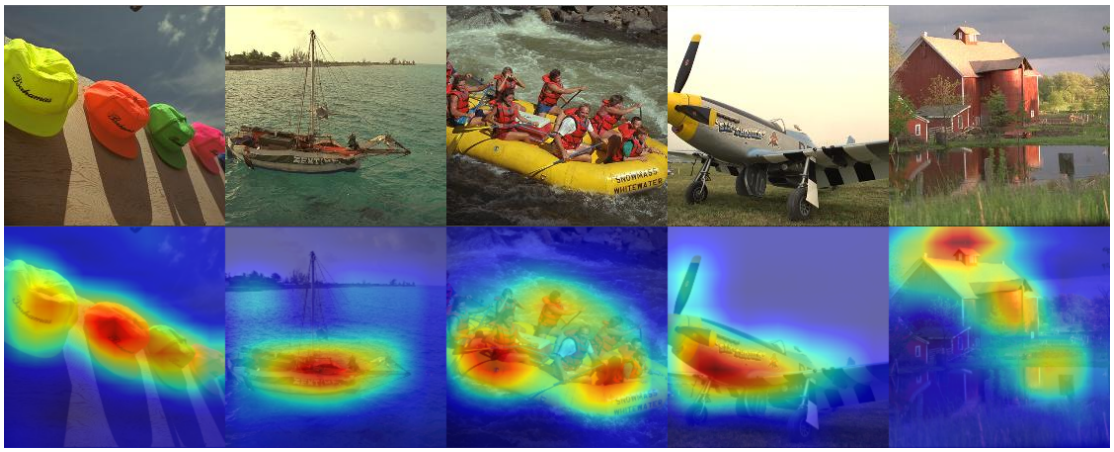

Figure 2.4: Sample of our map for five KODAK images.

We tested on the Kodak PhotoCD set (24 images) and the the MIT dataset (2,000 images). Kodak is a well known dataset consisting primarily of natural outdoor color images. Figure 2.4 shows a sample of five of these images, along with the corresponding heatmaps generated by our algorithm; the first four show typical results which strongly capture the salient content of the images, while the fifth is a rare case of partial failure, in which the heatmap does not fully capture all salient regions. The MIT set allows us to compare results across twenty categories. In Table 2.1 we only report averaged results across 'Outdoor Man-made' and 'Outdoor Natural' categories (200 images), as these categories are likely to contain multiple semantic objects, and are therefore appropriate for our method. Both

datasets contain images of smaller resolutions, but the effectiveness of perceptual compression is more pronounced for larger images. Therefore, we additionally selected a very large image of resolution 8705×8400, which we scale to a range of sizes to demonstrate the effectiveness of our system at a variety of resolutions. See Figure 2.5 for the image sizes used in this experiment. Both Figure 2.5 and Figure 2.6 show the `PSNR-HVS` difference between our model and standard JPEG. Positive values indicate our model has higher performance compared to standard JPEG. In addition to an array of standard quality metrics, we also report a `PSNR` value calculated only for those regions our method has identified as salient, which we term `PSNR-S`. By examining only regions of high semantic saliency, this metric demonstrates that our compression method is indeed able to preserve visual quality in targeted regions, without sacrificing performance on traditional image-level quality metrics or compression ratio. It should be noted that the validity of this metric is dependent on the correctness of the underlying saliency map, and thus should only be interpreted to demonstrate the success of the final image construction in preserving details highlighted by that map.

	PSNR-S	PSNR	PSNR-HVS	PSNR-HVSM	SSIM	MS-SSIM	VIFP
Kodak PhotoCD [24 images]							
Std JPEG	33.91	34.70	34.92	42.19	0.969	0.991	0.626
Our model	39.16	34.82	35.05	42.33	0.969	0.991	0.629
MIT Saliency Benchmark [Outdoor Man-made + Natural, 200 images]							
Std JPEG	36.9	31.84	35.91	45.37	0.893	0.982	0.521
Our model	40.8	32.16	36.32	45.62	0.917	0.990	0.529
Re-sized images of a very large image, see fig: 2.5 [20 images]							
Std JPEG	35.4	27.46	33.12	43.26	0.912	0.988	0.494
Our model	39.6	28.67	34.63	44.89	0.915	0.991	0.522

Table 2.1: Results across datasets

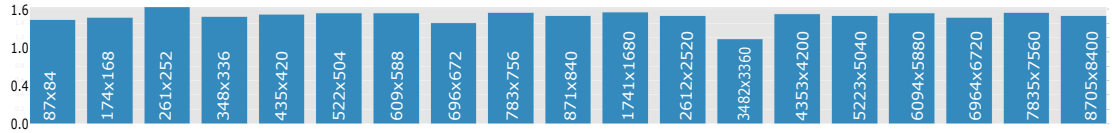

Figure 2.5: PSNR-HVS of our model - JPEG across various image size (higher is better).

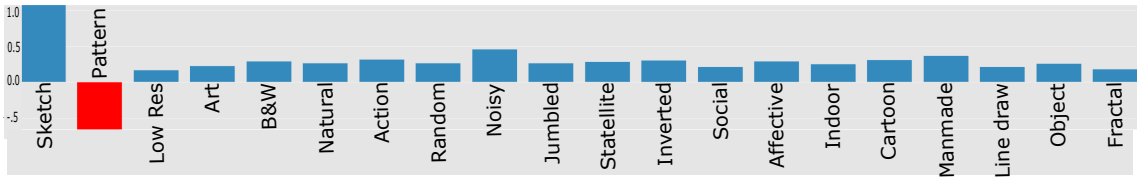

Figure 2.6: PSNR-HVS of our model - JPEG across various categories of MIT Saliency dataset (higher is better).

2.6 Discussion

While the comparison of metrics in Table 2.1 on standard JPEG and variable quality JPEG using our generated map is similar, it should be noted that these metrics still lack ability to judge human perception. It is evident from the results that the salient objects have significantly higher `PSNR-S`, yet maintains overall `PSNR` and the file size. We believe we have effectively created a model to make JPEG semantically aware without sacrificing overall image quality. We know from the JPEG standard that images compressed with higher Q better, and by design our model compresses the salient objects with higher Q; see Figure 2.1 for qualitative evaluation.

The results in Table 2.1 show the success of our method in maintaining or improving performance on traditional image quality metrics. Further, given the efficacy of our method in identifying multiple regions of interest, the `PSNR-S` measurements demonstrate the power of our method to produce superior visual quality in subjectively important regions.

When we look at Figure 2.1 we see the difference in quality of the salient object. Our model consistently outperforms standard JPEG on `SSIM` and `MS-SSIM` metric. These metrics measure the inherent structures in the image and Richter et al [RK09] [Ric11] has shown efficacy of these metric for image compression.

Figure 2.6 shows the performance of our model across all categories of the MIT dataset. Performance was strongest in categories like 'Outdoor Natural', 'Outdoor Man Made', 'Action' and 'Object', while categories like 'Line Drawing', 'Fractal' and 'Low Resolution' showed the least improvement. Not surprisingly, the category 'Pattern', which lacks semantic objects, is the only category where our model did not improve upon standard JPEG. Figure 2.5 shows results on the same image scaled to different sizes. Because our model benefits from

the scale-invariance of CNNs, we are able to preserve performance across a wide range of input sizes. Performance of our model on `PSNR-HVS` is consistent across all tested sizes. CNNs are able to learn scale-invariant features, and therefore the same object at any size should always be classified similarly. Our model preserves this feature even when looking for multiple objects.

2.7 Conclusion

We have presented a model which can learn to detect multiple objects at any scale and generate a map of multiple semantically salient image regions. This provides sufficient information to perform variable-quality image compression, without providing a precise semantic segmentation. Unlike region-based models, our model does not have to iterate over many windows. We sacrifice exact localization for the ability to detect multiple salient objects. Our variable compression improves upon visual quality without sacrificing compression ratio. Encoding requires a single inference over the pre-trained model, the cost of which is reasonable when performed using a GPU, along with a standard JPEG encoder. The cost of decoding, which employs a standard, off-the-shelf JPEG decoder remains unchanged. We believe it will be possible to incorporate our approach into other lossy compression methods such as JPEG 2000 and vector quantization, a subject of future work. Improvements to the power of our underlying CNN, addressing evolving visual quality metrics, and other applications such as video compression, are also potential areas of future work.

Chapter 3

Adversarial Images

Chapter 4

Pixel Deflection

Chapter 5

Efficient Training

Bibliography

[CE99] Surendar Chandra and Carla Schlatter Ellis. "JPEG compression metric as a quality aware image transcoding". In: *2nd USENIX Symposium on Internet Technologies and Systems (USITS99)*. 1999.

[Dai+16] Jifeng Dai et al. "R-FCN Object Detection via Region-based Fully Convolutional Networks". In: *arXiv preprint arXiv1605.06409* (2016).

[Den+09] J. Deng et al. "ImageNet A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009.

[Egi+06] Karen Egiazarian et al. "New full-reference quality metrics based on HVS". In: *CD-ROM proceedings of the second international workshop on video processing and quality metrics, Scottsdale, USA*. Vol. 4. 2006.

[Eve+10] Mark Everingham et al. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2 (2010).

[GHP07] Gregory Griffin, Alex Holub, and Pietro Perona. "Caltech-256 object category dataset". In: (2007).

[Gir+14] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[Gir15] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[GPG12] Giaime Ginesu, Maurizio Pintus, and Daniele D Giusto. "Objective assessment of the WebP image coding algorithm". In: *Signal Processing Image Communication* 27.8 (2012).

[He+16] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[Jia+15] Ming Jiang et al. "SALICON Saliency in context". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2015.

[KSC92] Stanley A Klein, D Amnon Silverstein, and Thom Carney. "Relevance of human vision to JPEG-DCT compression". In: *SPIE/IS&T 1992 Symposium on Electronic Imaging Science and Technology*. International Society for Optics and Photonics. 1992.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012.

[KT98] Konstantinos Konstantinides and Daniel Tretter. "A method for variable quantization in JPEG for improved text quality in compound documents". In: *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*. Vol. 2. IEEE. 1998.

[KTB14] Matthias Kúmmerer, Lucas Theis, and Matthias Bethge. "Deep gaze i Boosting saliency prediction with feature maps trained on imagenet". In: *arXiv preprint arXiv1411.1045* (2014).

[KVR15] Louis Kerofsky, Rahul Vanam, and Yuriy Reznik. "Perceptual adaptation of Objective video quality metrics". In: *Proc. Ninth International Workshop on Video Processing and Quality Metrics (VPQM)*. 2015.

[LB95] Yann LeCun and Yoshua Bengio. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995).

[Liu+15] Nian Liu et al. "Predicting eye fixations using convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[LRH15] Chi Li, Austin Reiter, and Gregory D Hager. "Beyond spatial pooling fine-grained representation learning in multiple domains". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[MHG+14] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. "Recurrent models of visual attention". In: *Advances in Neural Information Processing Systems*. 2014.

[MT00] Nasir D Memon and Daniel R Tretter. "Method for variable quantization in JPEG for improved perceptual quality". In: *Electronic Imaging*. International Society for Optics and Photonics. 2000.

[Oqu+15] Maxime Oquab et al. "Is object localization for free?-weakly-supervised learning with convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[Pon+07] Nikolay Ponomarenko et al. "On between-coefficient contrast masking of DCT basis functions". In: *Proceedings of the third international workshop on video processing and quality metrics*. Vol. 4. 2007.

[Ren+15] Shaoqing Ren et al. "Faster R-CNN Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015.

[Rez+13] Yuriy Reznik et al. *Coding of feature location information*. US Patent 8,571,306. 2013.

[Ric11] Thomas Richter. "SSIM as global quality metric a differential geometry view". In: *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*. IEEE. 2011.

[RK09] Thomas Richter and Kil Joong Kim. "A ms-ssim optimal jpeg 2000 encoder". In: *2009 Data Compression Conference*. IEEE. 2009.

[SB06] Hamid R Sheikh and Alan C Bovik. "Image information and visual quality". In: *IEEE Transactions on Image Processing* 15.2 (2006).

[SL09] X Yu Stella and Dimitri A Lisin. "Image compression based on visual saliency at individual scales". In: *International Symposium on Visual Computing*. Springer. 2009.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv1409.1556* (2014).

[Tod+16] George Toderici et al. "Full Resolution Image Compression with Recurrent Neural Networks". In: *arXiv preprint arXiv1608.05148* (2016).

[TPN96] Soon Hie Tan, Khee K Pang, and King N Ngan. "Classified perceptual coding with adaptive quantization". In: *IEEE transactions on circuits and systems for video technology* 6.4 (1996).

[Wan+04] Zhou Wang et al. "Image quality assessment from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004).

[WSB03] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. "Multiscale structural similarity for image quality assessment". In: *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. Vol. 2. Ieee. 2003.

[Xia+14] Guoqing Xiang et al. "An Adaptive Perceptual Quantization Algorithm Based on Block-Level JND for Video Coding". In: *Pacific Rim Conference on Multimedia*. Springer. 2014.

[ZF14] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European Conference on Computer Vision*. Springer. 2014.

[Zhe+15] Shuai Zheng et al. "Conditional random fields as recurrent neural networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[Zho+16] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[Zün+13] Fabio Zünd et al. "Content-aware compression using saliency-driven image retargeting". In: *2013 IEEE International Conference on Image Processing*. IEEE. 2013.