

KRYPTON: Accelerating Occlusion based Deep CNN Explainability Workloads

University of California, San Diego
@eng.ucsd.edu

ABSTRACT

Deep Convolution Neural Networks (CNN) have revolutionized the field of computer vision with even surpassing human level accuracy in some of the image recognition tasks. Thus they are now being deployed in many real-world use cases ranging from health care, autonomous vehicles, and e-commerce applications. However one of the major criticisms pointed against Deep CNNs is the black-box nature of how they make predictions. This is a critical issue when applying CNN based approaches to critical applications such as in health care where the explainability of the predictions is also very important. For interpreting CNN predictions several approaches have been proposed and one of the widely used method in image classification tasks is occlusion experiments. In occlusion experiments one would mask the regions of the input image using a small gray or black patch and record the change in the predicted label probability. By systematically changing the position of the patch location, a sensitivity map can be generated from which the regions in the input image which influence the predicted class label most can be identified. However, this method requires performing multiple forward passes of CNN inference for explaining a single prediction and hence is very time consuming. We present KRYPTON, the first data system to elevate occlusion experiments to a declarative level and enable database inspired automated *incremental* and *approximate* inference optimizations. Experiments with real-world datasets and deep CNNs show that KRYPTON can enable up to 10x speedups.

ACM Reference Format:

. 2018. KRYPTON: Accelerating Occlusion based Deep CNN Explainability Workloads. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

Deep convolutional neural networks (CNNs) [1–4] have revolutionized the computer vision field with even surpassing human level accuracies for some image recognition challenges such as ImageNet challenge [5]. Many of these successful pre-trained CNNs from computer vision challenges have been successfully re-purposed to be used in other real-world image recognition tasks, using a paradigm called *transfer learning* [6]. In transfer learning, instead of training a CNN from scratch, one uses a pre-trained Deep CNN, e.g., ImageNet trained VGG, and fine tune it for the target problem using the target training dataset. This approach avoids the need for a large training datasets, computational power and time which is otherwise a bottleneck for training a CNN from scratch. As a result, this paradigm has enabled the wide adoption of deep CNN technology in variety of real world image recognition tasks in several domains including health care [7, 8], agriculture [9], security [10], and sociology [11]. Remarkably, United States Food and Drug Administration Agency has already approved the use of deep CNN based technologies for identifying diabetic retinopathy, an eye disease found in adults with diabetes. It is also expected that above type of decision support systems will help the human radiologists in fulfilling their workloads efficiently and also provide a remedy to the shortage of qualified radiologists globally [12].

However, despite their many success stories, one of the major criticisms for deep CNNs and deep neural networks in general is the black-box nature of how they make predictions. In order to apply deep CNN based techniques in critical applications such as health care, the decisions should be explainable so that the practitioners can use their human judgment to decide whether to rely on those predictions or not [13].

In order to improve the explainability of deep CNN predictions several approaches have been proposed. One of the most widely used approach in image recognition tasks is occlusion experiments [14]. In occlusion experiments, as

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

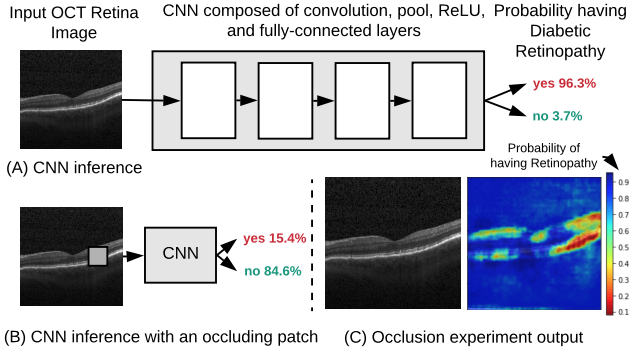


Figure 1: (A) Simplified representation depicting how OCT images are used for predicting Diabetic Retinopathy using CNNs. (B) Occluding parts of the OCT image changes the predicted probability for the disease. (C) By systematically moving patch along vertical and horizontal axes an output heat map is generated where each individual value corresponds to the predicted disease probability when the occlusion patch was placed on that position.

shown in Figure 1 (B), a square patch usually of gray or black color is used to occlude parts of the image and record the change in the predicted label probability. By systematically striding this patch horizontally and vertically one or two pixels at a time over the image a sensitivity heat map for the predicted label similar to one in Figure 1 (C) can be generated. Using this heat map, the regions in the image which are highly sensitive (or highly contributing) to the predicted class can be identified (corresponds to red color regions in the sensitivity heat map shown in Figure 1 (C)). This localization of highly sensitive regions then enables the practitioners to get an idea of the the prediction process of the deep CNN.

Example: Consider a radiologist who is examining Optical Coherence Tomography (OCT) images of the retina to identify potential diabetic retinopathy patients. The radiologist is recently given access to a deep CNN based clinical decision support system (CDSS) to identify potential images with diabetic retinopathy. It predicts the probability whether a retina image depicts a diabetic retinopathy case. She uses the CDSS for two main purposes: 1) as a cross checker while manually inspecting the retinal images, and 2) to prioritize potentially sever cases from a backlog of retina images. In both situations in addition to predicting the existence of the disease, the radiologist would like to have an explanation for the basis on which the CDSS makes the prediction, using the occlusion based explainability approach, to decide whether the pathological regions identified by the CDSS are correct and to ultimately whether to rely on the

CDSS decision. Similar examples arise in number of other heal care applications such as chest X-ray examination for identifying pneumonia cases and X-ray based child bone age assessment.

However, occlusion experiments are highly compute intensive and time consuming as each patch super imposition is treated as a new image and requires a separate CNN inference. In this work our goal is to apply database inspired optimizations to the occlusion based explainability workload to reduce both the computational cost and runtime taken for an experiment. This will also make occlusion experiments more amenable for interactive diagnosis of CNN predictions. Our main motivation is based on the observation that when performing CNN inference corresponding to each individual patch position, there are lot of redundant computations which can be avoided. To avoid redundant computations we introduce the notion of *incremental inference* of deep CNNs which is inspired by the incremental view maintenance approach which is heavily studied in the context of relational databases.

Due to the overlapping nature of how a convolution kernel would operate (details to follow in Section 3), the size of the modified patch will start growing as it progress through more layers in a CNN and reduce the amount of redundant computations. However at deeper layers the effect over patch coordinates which are radially further away from the center of the patch position will be diminishing. Our second optimization is based on this observation where we apply a form of *approximate inference* which applies a *propagation threshold* to limit the growth of the updating patch. We show that by applying propagation thresholds, a significant amount of computation redundancy can be retained.

The third optimization is also a form of *approximate inference* which we refer as *adaptive drill-down*. In most occlusion experiment use cases, such as in medical imaging, the object or pathological region of interest is contained in a relatively small region of the image. In such situations it is unnecessary to inspect the original image at the same high resolution of striding the occluding patch one or two pixels at a time, at all image locations. In adaptive drill-down approach, first a low resolution heatmap is generated with a larger occluding patch and a larger stride with relatively low computational cost and only the interested regions will be inspected further with a smaller occluding patch and a smaller stride to produce a higher resolution output. This two stage process also reduces the runtime of occlusion experiments.

Unlike the *incremental inference* approach which is exact, *propagation thresholding* and *adaptive drill-down* are approximate approaches. They essentially trade-off accuracy of the generated sensitivity heat map compared to the original, in favor of faster runtime. These changes in accuracy in the

generated heat map will be visible all the way from quality differences which are almost indistinguishable to the human eye to drastic structural differences, depending on the level of approximation. This opens up an interesting trade-off space of quality/accuracy versus runtime. KRYPTON expects the user to define the required level of quality for the generated heat maps by specifying the Structural Similarity Index (SSIM) (explained in Section 3) quality metric. The system will then automatically tune its *approximate inference* configuration values to yield the expected quality level during an initial tuning phase.

Finally, we have implemented KRYPTON on top of PyTorch deep learning toolkit by extending it by adding custom implementations of incremental and approximate inference operators. It currently supports VGG16, ResNet18, and InceptionV3 both on CPU and GPU environments, which are three widely used deep CNN architectures for transfer learning applications. We evaluate our system on three real-world datasets, 1) diabetic retinopathy (DR), 2) chest X-ray (CX), and 3) hand & wrist X-Ray (HX), and show that KRYPTON can result in up to 10x speedups with hardly distinguishable quality differences in the generated occlusion heat maps. While we have implemented KRYPTON on top of PyTorch toolkit, our work is largely orthogonal to choice of the deep learning toolkit; one could replace PyTorch with TensorFlow, Caffe2, CNTK, MXNet, or implement from scratch using C/CUDA and still benefit from our optimizations. Overall, this paper makes the following contributions:

- To the best of our knowledge, this is the first paper to study

Outline. The rest of this paper is organized as follows.

2 BACKGROUND

Deep CNNs. Deep CNNs are a type of neural networks specialized for image data. They exploit spatial locality of information in image pixels to construct a hierarchy of parametric feature extractors and transformers organized as layers of various types: *convolutions*, which use image filters from graphics, except with variable filter weights, to extract features; *pooling*, which subsamples features in a spatial locality-aware way; *batch-normalization*, which normalizes the output of the layer; *non-linearity*, which applies a non-linear transformation (e.g., ReLU); *fully connected*, which is a multi-layer perceptron; and *softmax*, which emits predicted probabilities to each class label. In most “deep” CNN architectures, above layers are simply stacked together with ones output is simply fed as the input to the other, while adding multiple layers element-wise or stacking multiple layers together to produce a new layer is also present in

some architectures. Popular deep CNN model architectures include AlexNet [1], VGG [2], Inception [4], ResNet [3], SqueezeNet [15], and MobileNet [16].

Deep CNN Explainability With image classification models, natural question is if the model is truly identifying objects in the image or just using surrounding or other objects for making false prediction. The various approaches used to explain CNN predictions can be broadly divided into two categories, namely gradient based and perturbation based approaches. Gradient based approaches generate a sensitivity map by computing the partial derivatives of model output with respect to every input pixel via back propagation. In perturbation based approaches the output of the model is observed by masking out regions in the input image and there by identify the sensitive regions. The most popular perturbation based approach is occlusion experiments which was first introduced by Zeiler et. al. [14]. Even though gradient approaches require only a single forward inference and a single backpropagation to generate the sensitivity map, the output may not be very intuitive and hard to understand because the salient pixels tend to spread over a very large area of the input image. Also as explained in [17], the back-propagation based methods are based on the AI researchers’ intuition of what constitutes a “good” explanation. But if the focus is on explaining decision to a human observer, then the approach used to produce the explanation should have a structure that humans accept. As a result in most real world use cases such as in medical imaging, practitioners tend to use occlusion experiments as the preferred approach for explanations despite being time consuming, as they produce high quality fine grained sensitivity maps [13].

Over the years there has been several modifications proposed to the original occlusion experiment approach. More recently Zintgraf. et. al. [18] proposed a variation to the original occlusion experiment approach named *Prediction Difference Analysis*. In their method instead of masking with a gray or black patch, samples from surrounding regions in the image are chosen as occlusion patches. In our work we mainly focus on the original occlusion experiment method. But, the methods and optimizations proposed in our work are readily applicable to more advanced occlusion based explainability approaches.

3 PRELIMINARIES AND OVERVIEW

In this section first we formally state the problem and explain our assumptions. Then we formalize the internals of some of the layers in a Deep CNN for the purpose of proposing our incremental inference approach in Section 4. Finally we briefly explain the Structural Similarity Index (SSIM)

Table 1: Symbols used in the Preliminaries Section

Symbol	Meaning
$\mathcal{I}(\mathcal{I}_{img})$	Input activation volume (Input Image)
\mathcal{O}	Output activation volume
C_{in}, H_{in}, W_{in}	Depth, height, and width of Input
$C_{out}, H_{out}, W_{out}$	Depth, height, and width of Output
\mathcal{K}_{conv}	Convolution filter kernels
\mathcal{B}_{conv}	Convolution bias value vector
\mathcal{K}_{pool}	Pooling filter kernel
H_k, W_k	Height and width of filter kernel
$S(S_x, S_y)$	Filter kernel or occlusion patch striding amount (S_x and S_y corresponds to width and height dimensions)
$P(P_x, P_y)$	Padding amount (P_x and P_y corresponds to padding along width and height dimensions)
$Q(Q_{inc})$	Total FLOPS count with full (incremental) inference
L	Set of convolution layers in a CNN
R	Redundancy ratio
\mathcal{P}	Occluding patch
$\mathcal{I}_{img}^{x,y}$	Modified image by superimposing \mathcal{P} on top of \mathcal{I}_{img} such that the top left corner of \mathcal{P} is positioned at x, y location of \mathcal{I}_{img}
M	Heat map produced by the occlusion experiment
H_M, W_M	Height and width of M
f	Fine-tuned CNN
C	Class label predicted by f for the original image \mathcal{I}_{img}
$\circ_{x,y}$	Superimposition operator. $A \circ_{x,y} B$, superimposes B on top of A starting off at (x, y) position

which is used to quantify the quality of the generated sensitivity heat map with respect to the original in the context of approximate inference.

3.1 Problem Statement and Assumptions

We are given a fine-tuned CNN f , an image \mathcal{I}_{img} on which the occlusion experiment needs to be run, the predicted class label C for the image \mathcal{I}_{img} , an occluding patch \mathcal{P} in RGB format, and patch striding amounts S_x and S_y which will be used to move the occlusion patch on the image at a time in horizontal and vertical directions. The occlusion experiment workload is to generate a 2D heat map M where each individual value in M corresponds to the predicted probability for C by f when \mathcal{P} is superimposed on \mathcal{I}_{img} corresponding to the position of that particular value. More precisely, we can state the workload using the following set of logical statements:

$$W_M = \lfloor (\text{width}(\mathcal{I}_{img}) - \text{width}(\mathcal{P}) + 1) / S_x \rfloor \quad (1)$$

$$H_M = \lfloor (\text{height}(\mathcal{I}_{img}) - \text{height}(\mathcal{P}) + 1) / S_y \rfloor \quad (2)$$

$$M \in \mathbb{R}^{H_M \times W_M} \quad (3)$$

$$\forall x, y \in [0, H_M] \times [0, W_M] : \quad (4)$$

$$\mathcal{I}_{img}^{x,y} \leftarrow \mathcal{P} \circ_{x,y} \mathcal{I}_{img} \quad (5)$$

$$M[x, y] \leftarrow f(\mathcal{I}_{img}^{x,y}, C) \quad (6)$$

Step (1), and (2) calculates the dimensions of the generated heat map M which is dependent on the dimensions of \mathcal{I}_{img} , \mathcal{P} , and the values of S_x and S_y . Step (5) superimposes \mathcal{P} on \mathcal{I}_{img} with its top left corner placed on the (x, y) location of \mathcal{I}_{img} . Step (6) calculates the output value at the (x, y) location by performing CNN inference for $\mathcal{I}_{img}^{x,y}$ using f and taking the predicted probability for the label C .

We make few simplifying assumptions in this paper for tractability sake. First we assume that f is a fine-tuned CNN from a roster of well-known CNNs (currently, VGG, ResNet, and Inception). This is a reasonable start, since most recent fine-tuning based transfer learning applications used only such well-known CNNs from model zoos [19, 20]. Nevertheless, our work is orthogonal to the specifics of a particular architecture and the proposed approaches can be easily extended to any architecture. We leave support for arbitrary CNNs to future work. Second we assume that the images of a particular occlusion experiment use case are mostly similar such as in OCT retina images, chest X-ray images, and hand & wrist X-ray images. This assumption is required in the context of *approximate inference* optimizations, in which KRYPTON uses an initial set of images to tune its configuration parameters to produce subsequent occlusion heat maps with the user required quality metric. However, if the images are drastically different the chances are that the system configurations tuned for one set of images doesn't map to other images well, and in such situations KRYPTON's approximate inference optimizations can be disabled.

3.2 Deep CNN Internals

The output activations of the layers in a Deep CNN, except for fully-connected ones, are arranged into three dimensional volumes which has a width, height, and depth. For example an RGB input image of 224×224 spatial size can be considered as an input volume having a width and height of 224 and a depth of 3 (corresponding to 3 color channels). Every non fully-connected layer will take in an input activation volume and transform it into another activation volume, where as a fully-connected layer will transform an input volume into an output vector. For our purpose these transformations can be broadly divided into three subcategories based on how they spatially operate:

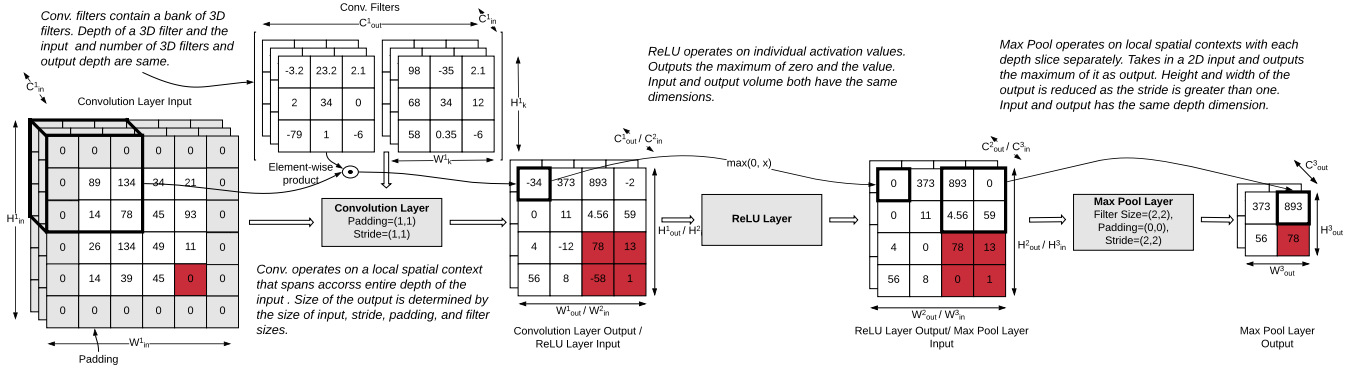


Figure 2: Simplified representation of selected layers of a Deep CNN. For simplicity sake addition of bias is not shown in the Conv. transformation. The values marked in red shows how a small spatial update in the first input would propagate through subsequent transformations. Notation used is explained in Table 1.

- Transformations that operate on individual spatial locations.
 - E.g. ReLU, Batch Normalization
- Transformations that operate on a local spatial context.
 - E.g. Convolution, Pooling
- Transformation that operate on a global spatial context.
 - E.g. Fully-Connected

With incremental spatial updates in the input, both types of transformations that operate at individual spatial locations and transformations that operate at a local spatial contexts provide opportunities for exploiting redundancy. Extending the transformations that operate at individual spatial locations to become redundancy aware is straightforward. However, with transformations that operate on a local spatial context such as convolution and pooling, this extension is non-trivial due to the overlapping nature of the spatial contexts corresponding to individual transformations. We next formally define the transformations of convolution and pooling layers and also the relationship between input and output dimensions for these layers which will be later used in Section. 4 to introduce our incremental inference approach.

Transformations that operate on a local spatial context. The two types of transformations that operate on a local spatial context in a deep CNN are convolution and max pooling layers. Convolution layers are the most important type of layer in a CNN architecture that also contributes to most of the computational cost. Each convolutional layer can have several (say C_{out}) three dimensional filter kernels organized into a four dimensional array \mathcal{K}_{conv} with each having a smaller spatial width W_k and height H_k compared to the width W_{in} and height H_{in} of the input volume \mathcal{I} , but

has the same depth C_{in} . During inference, each filter kernel is slid along the width and height dimensions of the input and a two dimensional activation map is produced by taking element-wise product between the kernel and the input and adding a bias value $\mathcal{B}[c]$ for some $c \in [0, C_{out} - 1]$. These two dimensional activation maps are then stacked together along the depth dimension to produce an output volume \mathcal{O} having the dimensions of $(C_{out}, H_{out}, W_{out})$. This can be formally defined as follows:

$$\text{Input Volume : } \mathcal{I} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}} \quad (7)$$

$$\text{Conv. Filters : } \mathcal{K}_{conv} \in \mathbb{R}^{C_{out} \times C_{in} \times H_k \times W_k} \quad (8)$$

$$\text{Conv. Bias Vector : } \mathcal{B}_{conv} \in \mathbb{R}^{C_{out}} \quad (9)$$

$$\text{Output Volume : } \mathcal{O} \in \mathbb{R}^{C_{out} \times H_{out} \times W_{out}} \quad (10)$$

$$\begin{aligned} \mathcal{O}[c, y, x] = & \sum_{k=0}^{C_{in}} \sum_{j=0}^{H_k-1} \sum_{i=0}^{W_k-1} \mathcal{K}_{conv}[c, k, j, i] \\ & \times \mathcal{I}[k, y - \lfloor \frac{H_k}{2} \rfloor + j, x - \lfloor \frac{W_k}{2} \rfloor + i] + \mathcal{B}[c] \end{aligned} \quad (11)$$

The main objective of having pooling layers in CNNs is to reduce the spatial size of output volumes. Pooling can also be thought as a convolution operation with a fixed (i.e. not learned) two dimensional filter kernel \mathcal{K}_{pool} having a width of W_k and height of H_k , which unlike convolution, operates independently on every depth slice of the input volume. The two main variations of pooling layers are max pooling (takes the maximum value from the local spatial context) and average (takes the average value from the local spatial context) pooling. A Pooling layer takes a three dimensional activation volume \mathcal{O} having a depth of C , width of W_{in} , and height

of H_{in} as input and produces another three dimensional activation volume O which has the same depth of C , width of W_{out} , and height of H_{out} as the output. Pooling kernel is generally slid with more than one pixel at a time and hence W_{out} and H_{out} are generally smaller than W_{in} and H_{in} . Pooling operation can be formally defined as follows:

$$\text{Pool Filters : } \mathcal{K}_{pool} \in \mathbb{R}^{H_k \times W_k} \quad (12)$$

$$O[c, y, x] = \sum_{j=0}^{H_k-1} \sum_{i=0}^{W_k-1} \mathcal{K}_{pool}[j, i] \times I[c, y - \lfloor \frac{H_k}{2} \rfloor + j, x - \lfloor \frac{W_k}{2} \rfloor + i] \quad (13)$$

Relationship between Input and Output Spatial Sizes.

The output volume's spatial size (W_{out} and H_{out}) is determined by the spatial size of the input volume (W_{out} and H_{out}), spatial size of the filter kernel (W_k and H_k) and two other parameters: **stride** S and **padding** P . Stride is the amount of pixel values used to slide the filter kernel at a time when producing a two dimensional activation map. It is possible to have two different values with one for the width dimension S_x and one for the height dimension S_y . Generally $S_x \leq W_k$ and $S_y \leq H_k$. Sometimes in order to control the spatial size of the output activation map to be same as the input activation map, one needs to pad the input feature map with zeros around the spatial border. Padding (P) captures the amount of zeros that needs to be added. Similar to the stride S , it is possible to have two separate values for padding with one for the width dimension P_x and one for the height dimension P_y . With these parameters defined the width and the height of the output activation volume can be defined as follows:

$$W_{out} = (W_{in} - W_k + 1 + 2 \times P_x) / S_x \quad (14)$$

$$H_{out} = (H_{in} - H_k + 1 + 2 \times P_y) / S_y \quad (15)$$

Estimating the Computational Cost of Deep CNNs

Deep CNNs are highly compute intensive and out of the different types of layers, Conv layers contributes to 90% (or more) of the computations. One of the widely used way to estimate the computational cost of a Deep CNN is to estimate the number of fused multiply add (FMA) floating point operations (FLOPs) required by convolution layers for a single forward inference and ignore the computational cost of other layers (e.g. Pooling, Fully-Connected).

For example, applying a convolution filter having the dimensions of (C_{in}^l, H_k^l, W_k^l) to a single spatial context will require $C_{in}^l \times H_k^l \times W_k^l$ many FLOPs, each corresponding to a single element-wise multiplication. Thus, the total amount

of computations Q_l required by that layer in order to produce an output O having dimensions $C_{out}^l \times H_{out}^l \times W_{out}^l$, and the total amount of computations Q required to process the entire set of convolution layers L in the CNN can be calculated as per equation. 16 and equation. 17. However, in the case incremental updates effectively only a smaller spatial patch having a width W_p^l ($W_p^l \leq W_{out}^l$) and height H_p^l ($H_p^l \leq H_{out}^l$) is needed to be recomputed. The amount of computations required for the incremental computation Q_{inc}^l and total amount of incremental computations Q_{inc} required for the entire set of convolution layers L will be smaller than the above full computation values and can be calculated as per equation. 18 and equation. 19.

Based on the above quantities we define a new metric named **redundancy ratio** R , which is the ratio between total full computational cost Q and total incremental computation cost Q_{inc} (see equation. 20). This ratio essentially acts as a surrogate for the theoretical upper-bound for computational and runtime savings that can be achieved by applying incremental computations to deep CNNs.

$$Q^l = (C_{in}^l \times H_k^l \times W_k^l) \times (C_{out}^l \times H_{out}^l \times W_{out}^l) \quad (16)$$

$$Q = \sum_{l \in L} Q^l \quad (17)$$

$$Q_{inc}^l = (C_{in}^l \times H_k^l \times W_k^l) \times (C_{out}^l \times H_p^l \times W_p^l) \quad (18)$$

$$Q_{inc} = \sum_{l \in L} Q_{inc}^l \quad (19)$$

$$R = \frac{Q}{Q_{inc}} \quad (20)$$

3.3 Estimating the Quality of Generated Approximate Heat Maps

When applying approximate inference optimizations, KRYPTON sacrifices the the accuracy of the generated heat map in favor of lesser runtime. To measure this drop of accuracy we use Structural Similarity (SSIM) Index [21], which is one of the widely used approaches to measure the *human perceived difference* between two images. When applying SSIM index, we treat the original heat map as the reference image with no distortions and the perceived image similarity of the KRYPTON generated heat map is calculated with reference to it. The generated SSIM index is a value between -1 and 1 , where 1 corresponds to perfect similarity. It is important to note that, even though SSIM index value of 1 corresponds to perfect similarity, other values doesn't necessarily imply same level of perceived quality across different image pairs. However, if the original images are closely similar, such as in chest X-ray images, it can be assumed that this condition will hold. Typically SSIM index values above 0.9 are used in

practical applications such as image compression and video encoding as at the human perception level they produce indistinguishable distortions. For more details on SSIM Index method, we refer the reader to the original SSIM Index paper [21].

4 OPTIMIZATIONS

In this section we explain *incremental inference* and the two *approximate inference* approaches, patch *propagation thresholding* and *adaptive drill-down*, in detail. In KRYPTON these optimizations are applied on top of the current dominant approach of performing CNN inference on batches of images where each image corresponds to an occluded instance of the original image. Batched inference is important as it reduces per image inference time by amortizing the fixed overheads. In our experiments we found that this simple optimization alone can give up to 1.4X speedups on CPU environments and 2X speedups on the GPU environment compared to the per image inference approach. Finally we explain how KRYPTON configures its internal system configurations for *approximate inference* by using a sample image set during the initial tuning stage.

4.1 Incremental Inference

As explained earlier, occlusion experiments in its naive form performs many redundant computations. In order to avoid these redundancies, layers in a CNN has to be change aware and operate in an incremental manner i.e. reuse previous computations as much as possible and compute only the required ones. In this section we focus on transformations that operate on a local spatial context (i.e. Convolution and Pooling) as other types either has no redundancies (global context transformations) or is trivial to make incremental (point transformations). The choice of CPUs vs GPUs for CNN inference also brings up new considerations for the batched implementation of these incremental transformations and we explain two version of implementations, one which is a naive implementation of batched incremental inference approach and the other a GPU optimized version. We also explain incremental implementations of two other linear algebra operators, element-wise addition and depth-wise concatenation.

Incremental Convolution and Pooling.

The two types of transformations in a CNN that operate on a local spatial context are Convolution and Pooling. In Section 3 we showed that both of these transformations can be cast into a form of applying a filter along the spatial dimensions of the input volume. However how each transformation operate along the depth dimension is different. For our purpose we are interested in finding the propagation of the patches in the input through the consecutive layers

Table 2: Additional symbols used in the Optimizations Section

Symbol	Meaning
x_p^I, y_p^I	Starting coordinates of the input patch
x_p^R, y_p^R	Starting coordinates of the patch that needs to be read in for the transformation
x_p^O, y_p^O	Starting coordinates of the output patch
H_p^I, W_p^I	Height, and width of the input patch
H_p^R, W_p^R	Height, and width of the patch that needs to be read in for the transformation
H_p^O, W_p^O	Height, and width of the output patch
τ	Projective field threshold

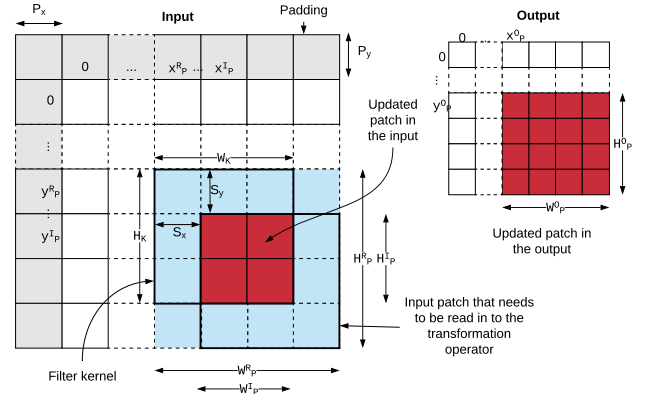


Figure 3: Simplified representation of input and output patch coordinates and dimensions of Conv. and Pool transformations.

and hence both these transformations can be treated similarly. The coordinates and the dimensions (i.e. height and width) of the modified patch in the output volume caused by a modified patch in the input volume are determined by the coordinates and the dimensions of the input patch, sizes of the filter kernel (H_k and W_k), padding values (P_x and P_y), and the strides (S_x and S_y). For example consider simplified demonstration showing a cross-section of input and output in Figure 3. We use a coordinate system whose origin is placed at the top left corner of the input. A patch is placed on the input starting off at x_p^I, y_p^I coordinates and has a height of H_p^I and width of W_p^I . The updated patch in the output starts off at x_p^O, y_p^O and has a height of H_p^O and width of W_p^O . Note that due to the overlapping nature of filter positions, to compute the output patch, transformations may have to read a slightly larger context than the input patch.

This read in context is shown by the blue shaded area in Figure 3. The starting coordinates of this read-in patch are denoted by x_p^R, y_p^R and the dimensions are denoted by W_p^R, H_p^R . The relationship between the coordinates and dimensions can be expressed as follows:

$$x_p^O = \max(\lceil (P_x + x_p^I - W_K + 1)/S_x \rceil, 0) \quad (21)$$

$$y_p^O = \max(\lceil (P_y + y_p^I - H_K + 1)/S_y \rceil, 0) \quad (22)$$

$$W_p^O = \min(\lceil (W_p^I + W_K - 1)/S_x \rceil, W_{out}) \quad (23)$$

$$H_p^O = \min(\lceil (H_p^I + H_K - 1)/S_y \rceil, H_{out}) \quad (24)$$

$$x_p^R = x_p^O \times S_x - P_x \quad (25)$$

$$y_p^R = y_p^O \times S_y - P_y \quad (26)$$

$$W_p^R = W_K + (W_p^O - 1) \times S_x \quad (27)$$

$$H_p^R = H_K + (H_p^O - 1) \times S_y \quad (28)$$

Equation 21 and 22 calculates the starting coordinates of the output patch. Use of padding effectively shifts the coordinate system and therefore P_x and P_y values are added to correct it. Due to the overlapping nature of filter kernels, the maximum affected span of the updated patch in the input will be increased by $W_K - 1$ and $H_K - 1$ amounts and hence needs to be subtracted from the input coordinates x_p^I and y_p^I (a filter of size W_K which is placed starting at $x_p^I - W_K + 1$ will see the new change at x_p^I). Dividing the above values by the stride values S_x and S_y and taking the ceil gives the starting coordinates of the output patch (essentially calculates the number of strides). Towards the left side edge of the input, where the affected span of the input cannot be extended by $W_K - 1$ or $H_K - 1$ amounts this value will be negative. Therefore the maximum of zero or the above value should be taken as the final. Equation 23 and 24 calculates the width and height of the output patches. Similar to output coordinates calculations, the span of the input patch is increased by $W_K - 1$ and $H_K - 1$ amounts. Dividing the above values by the stride values S_x and S_y and taking the ceil gives the width and height of the output patch. Since the output patch cannot grow beyond the size of the output, minimum of the output dimension or the above value should be taken as the final. Once the output patch coordinates and dimensions are calculated, it is straight forward to calculate the read-in patch coordinates as per Equations 25 and 26 and the dimensions as per Equations 27 and 28.

It is important to note that there are special situations under which the actual output patch size can be smaller than above calculated value. Consider the simplified one dimensional situation shown in Figure 4 (a), where the stride value¹ (3) is same as the filter size (3). In this situation the

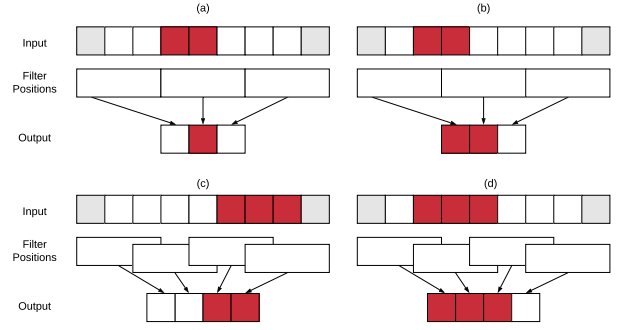


Figure 4: One dimensional representation showing special situations under which actual output size will be smaller than the values calculated by Equations 21 and 22. (a) and (b) shows a situation with filter stride being equal to the filter size. (c) and (d) shows a situation with input patch being placed at the edge of the input.

the size of the output patch is one less than the value calculated by Equation 23. However it is not the case in Figure 4 (b) which has the same input patch size but is placed at a different location. This issue arises only when the stride value is same as the filter size. Another situation arises when the input patch is placed at the edge of the input as shown in Figure 4 (c). In this situation it is not possible for the filter to move freely through all filter positions as it hits the input boundary compared to having the input patch on the middle of the input as shown in Figure 4 (c). In KRYPTON we do not distinguish these differences and use the values calculated by Equation 23 and 24 as they act as an upper bound. In case of a smaller output patch, KRYPTON simply reads off and updates slightly bigger patches to preserve uniformity. This also requires updating the starting coordinates of the patches as shown in Equations 29 and 32. Such uniform treatment is required for performing batched inference operations which out of the box gives significant speedups compared to per image inference.

If $x_p^O + W_p^O > W_{out}$:

$$x_p^O = W_{out} - W_p^O \quad (29)$$

$$x_p^I = W_{in} - W_p^I \quad (30)$$

$$x_p^R = W_{in} - W_p^R \quad (31)$$

If $y_p^O + H_p^O > H_{out}$:

$$y_p^O = H_{out} - H_p^O \quad (32)$$

$$y_p^I = H_{in} - H_p^I \quad (33)$$

$$y_p^R = H_{in} - H_p^R \quad (34)$$

¹Note that the stride value is generally less than or equal to the filter size.

Algorithm 1 Incremental Inference Algorithm

Input:

T : Transformation
 I : Pre-materialized input from original image
 $[P_1^I, \dots, P_n^I]$: Input patches
 $[(x_{P_1}^I, y_{P_1}^I), \dots, (x_{P_n}^I, y_{P_n}^I)]$: Input patch coordinates
 W_P^I, H_P^I : Input patch dimensions

Output:

$[P_1^O, \dots, P_n^O]$: Output patches
 $[(x_{P_1}^O, y_{P_1}^O), \dots, (x_{P_n}^O, y_{P_n}^O)]$: Output patch coordinates
 W_P^O, H_P^O : Output patch dimensions

```

1: procedure INCREMENTALINFERENCE
2:   Calculate  $[(x_{P_1}^O, y_{P_1}^O), \dots, (x_{P_n}^O, y_{P_n}^O)]$  and  $(W_P^O, H_P^O)$ 
3:   Calculate  $[(x_{P_1}^R, y_{P_1}^R), \dots, (x_{P_n}^R, y_{P_n}^R)]$  and  $(W_P^R, H_P^R)$ 
4:   Initialize  $R \in \mathbb{R}^{\text{depth}(I) \times H_P^R \times W_P^R}$ 
5:   for  $i$  in  $[1, \dots, n]$  do
6:      $t_1 \leftarrow I[:, x_{P_i}^R : x_{P_i}^R + W_P^R, y_{P_i}^R : y_{P_i}^R + H_P^R]$ 
7:      $t_2 \leftarrow P_i \circ_{(x_{P_i}^I - x_{P_i}^R), (y_{P_i}^I - y_{P_i}^R)} t_1$ 
8:      $R[i, :, :] \leftarrow t_2$ 
9:    $[P_1^O, \dots, P_n^O] \leftarrow T(R)$ 
10:  return  $[P_1^O, \dots, P_n^O], [(x_{P_1}^O, y_{P_1}^O), \dots, (x_{P_n}^O, y_{P_n}^O)],$ 
11:     $(W_P^O, H_P^O)$ 

```

Input:

O : Pre-materialized output from original image
 $[P_1^O, \dots, P_n^O]$: Output patches
 $[(x_{P_1}^O, y_{P_1}^O), \dots, (x_{P_n}^O, y_{P_n}^O)]$: Output patch coordinates

Output:

O' : Updated output

```

1: procedure INCREMENTALTOFULLPROJECTION
2:    $O' \leftarrow \text{copy}(O)$ 
3:   for  $i$  in  $[1, \dots, n]$  do
4:      $O'[i, :, :] \leftarrow P_i^O \circ_{x_{P_i}^O, y_{P_i}^O} O'[i, :, :]$ 
5:  return  $O'$ 

```

With all the geometric mappings defined, we now explain the complete incremental inference approach for a single transformation. Algorithm 1 presents it formally. The INCREMENTALINFERENCE procedure takes in the original transformation T , pre-materialized input for T corresponding to original image, a batch of updated patches and their geometric properties as input. First it calculates geometric properties of the output and read-in patches. A temporary input volume R is initialized to hold the input patches with their read-in contexts. The FOR loop iteratively populates R with corresponding patches. Finally T is applied on R to compute

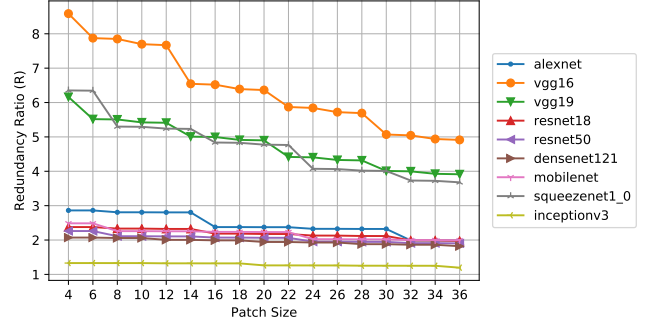


Figure 5: Maximum attainable redundancy ratios of popular CNN architectures when a square occlusion patch of different sizes is placed on the center of the image.

the output patches. In a CNN which has multiple such transformations chained together, the outputs of the INCREMENTALINFERENCE procedure are fed as input again for the incremental inference of the next transformation along with the unchanged pre-materialized input corresponding to the new transformation. However at a local context transformation and global context transformation such as in Conv. \rightarrow Fully-Connected or Pool \rightarrow Fully-Connected, full updated output has to be created as per INCREMENTALTOFULLPROJECTION procedure instead of propagating only the updated patches.

We analyze the maximum attainable redundancy ratios for popular CNN architectures when a square occlusion patch of different sizes is placed on the center² of the input image. Figure 5 shows the results. VGG 16-layer version results in the maximum redundancy ratio and Inception V3 model has the lowest redundancy ratio. Most CNN architectures result in a redundancy ratio between 2-3 except VGG 16, VGG 19, and Squeezenet 1.0 CNNs which result in higher redundancy ratios. The attainable redundancy ratio of a CNN is determined by the aspects of its internal architecture such as number of layers, the size of the filter kernels, and the filter stride values.

CPU versus GPU implementation considerations.

Through our experiments we found that even though a straight-forward implementation of *incremental inference* approach as shown in Algorithm 1 produces expected speedups for the CPU environment it performs poorly on the GPU environment. The for loop on the line 5 of Algorithm 1 is essentially preparing the input for T by copying values from one part of the memory to another sequentially. This sequential operation becomes a bottleneck for the GPU

²If the occlusion patch is placed towards to a corner of the input image the redundancy ratio will be slightly higher. But placing the occlusion on the center gives us a worst case estimate.

implementation as it cannot exploit the available parallelism of the GPU effectively. To overcome this a custom GPU kernel which performs the input preparation more efficiently by parallelly executing the memory copying for each item in the input batch is needed.

Element-wise addition and depth-wise concatenation.

Element-wise addition and depth-wise concatenation are two widely used linear algebra operators in CNNs. Element-wise addition operator requires the two input volumes to have the same dimensions and the depth-wise concatenation requires them to have same spatial dimensions. Consider a situation for these operators where the first input has incremental spatial update starting at $x_{P_1}^I, y_{P_1}^I$ coordinates with dimensions of $H_{P_1}^I$ and $W_{P_1}^I$ and for the second input starting at $x_{P_2}^I, y_{P_2}^I$ coordinates with dimensions of $H_{P_2}^I$ and $W_{P_2}^I$. Then the coordinates and the dimensions of the output and read-in patches can be expressed as follows:

$$x_P^O = x_P^R = \min(x_{P_1}^I, x_{P_2}^I) \quad (35)$$

$$y_P^O = y_P^R = \min(y_{P_1}^I, y_{P_2}^I) \quad (36)$$

$$W_P^O = W_P^R = \max(x_{P_1}^I + W_{P_1}^I, x_{P_2}^I + W_{P_2}^I) - \min(x_{P_1}^I, x_{P_2}^I) \quad (37)$$

$$H_P^O = H_P^R = \max(y_{P_1}^I + H_{P_1}^I, y_{P_2}^I + H_{P_2}^I) - \min(y_{P_1}^I, y_{P_2}^I) \quad (38)$$

4.2 Projective Field Thresholding

Projective field [22, 23] of a CNN neuron is the local region (including the depth) of the output volume which is connected to it. The term is borrowed from Neuroscience field where it is used to describe the spatiotemporal effects exerted by a retinal cell on all of the outputs of the neuronal circuitry [24]. For our work the notion of projective field is useful as it essentially determines the change propagation path for incremental changes. The three types of transformations affects the size of the projective field differently. Point transformations does not change of the projective field size while global context transformations increases it to the maximum and local context transformations increase it gradually. The amount of increase in a local context transformation is determined the filter size and stride parameters. At every transformation the size of the projective field will increase linearly by the filter size and multiplicative by the stride value.

Because of the projective field growth, even though there will be much computational redundancies in the early layers, towards the latter layers it will decrease or even have no redundancies. However, we empirically found that the projective field growth can be truncated up to a certain level without significantly affecting the accuracy. For a more intuitive understanding on why this would work consider the

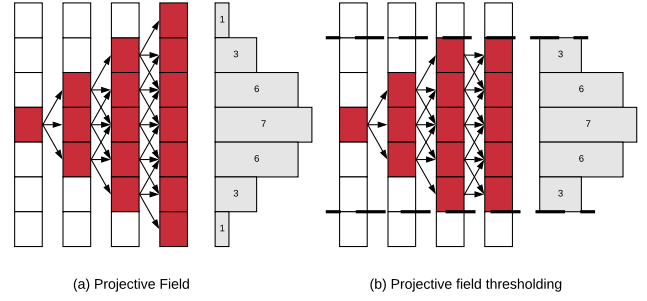


Figure 6: (a) One dimensional convolution demonstrating projective field growth (filter size = 2, stride = 1). (b) Projective field thresholding with $\tau = 5/7$. Histograms denote the number of unique change propagation paths.

simplified one dimensional convolution example shown in Figure 6 (a). In the example a single neuron is modified (marked in red) and a filter of size three is applied with a stride of one repeatedly four times. Since the filter size is three, each updated neuron will propagate the change to three neurons in the next output layer causing the projective field to grow linearly. At the end of the fourth layer, the histogram shows the number of unique paths that are available between each output neuron and the original updated neuron in the first layer. It can be seen that this distribution resembles a Gaussian where many of the paths are connected to the central region. The amount of actual change in the output layer is determined by both the number of unique paths and also the individual weights of the connections. But the actual change in the output will converge to a Gaussian in distribution over all possible weight values. For more details we refer the reader to [25], where a similar theoretical result has been proved for the receptive field³ of a deep CNN.

As most of the change will be concentrated on the center, we introduce the notion of a projective field threshold τ ($0 < \tau \leq 1$) which will be used to restrict the growth of the projective field. It essentially determines the maximum size of the projective field as a fraction of the size of the input. Figure 6 (b) demonstrates the application of projective field thresholding with a τ value of $5/7$. From the histogram generated for the projective field thresholding approach it can be seen that much of the final output change is maintained by this approach while some small changes are lost from the edges.

³Receptive field of a CNN neuron is the local region (including the depth) of the input volume which is connected to it.

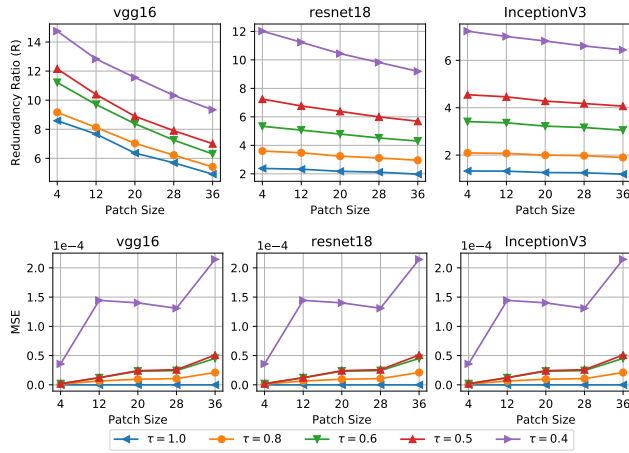


Figure 7: (a) Distribution of redundancy ratios for different occlusion patch sizes and projective field thresholding values. (b) Mean Square Error of the final convolution output with different projective field thresholding and occlusion patch sizes for a sample of Diabetic Retinopathy images.

4.3 Adaptive Drill-Down

4.4 System Tuning

5 EXPERIMENTAL EVALUATION

5.1 End-to-End Evaluation

5.2 Drill-Down Analysis

REFERENCES

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [6] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1790–1802, 2016.
- [7] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.
- [8] Mohammad Tariqul Islam, Md Abdul Aowal, Ahmed Tahseen Minhaz, and Khalid Ashraf. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*, 2017.
- [9] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016.
- [10] Farhad Arbabzadah, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Identifying individual facial expressions by deconstructing a neural network. In *German Conference on Pattern Recognition*, pages 344–354. Springer, 2016.
- [11] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. 2017.
- [12] Radiologists are often in short supply and overworked deep learning to the rescue. <https://government.diginomica.com/2017/12/20/radiologists-often-short-supply-overworked-deep-learning-rescue>. Accessed September 31, 2018.
- [13] Kyu-Hwan Jung, Hyunho Park, and Woohan Hwang. Deep learning for medical image analysis: Applications to computed tomography and magnetic resonance imaging. *Hanyang Medical Reviews*, 37(2):61–70, 2017.
- [14] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [15] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [17] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.
- [18] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.
- [19] Caffe model zoo. <https://github.com/BVLC/caffe/wiki/Model-Zoo>. Accessed September 31, 2018.
- [20] Models and examples built with tensorflow. <https://github.com/tensorflow/models>. Accessed September 31, 2018.
- [21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [22] Hung Le and Ali Borji. What are the receptive, effective receptive, and projective fields of neurons in convolutional neural networks? *arXiv preprint arXiv:1705.07049*, 2017.
- [23] Basic operations in a convolutional neural network - cse@iit delhi. <http://www.cse.iitd.ernet.in/~rijurekha/lectures/lecture-2.pptx>. Accessed September 31, 2018.
- [24] Saskia EJ de Vries, Stephen A Baccus, and Markus Meister. The projective field of a retinal amacrine cell. *Journal of Neuroscience*, 31(23):8595–8604, 2011.
- [25] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*, pages 4898–4906, 2016.