Explainable Artificial Intelligence for the analysis of diffraction images

DESY Summer Student Programme, 2021

Barbara Klaudel
Gdańsk University of Technology

Supervisors:
David Pennicard, Shah Nawaz, Vahid Rahmani

September 8, 2021

Abstract

Contents

1 Introduction

Our notion of life is derived from our understanding of the molecular structure of the matter. The properties and functions of the matter are encoded in its molecular structure. The majority of protein structure models were obtained with X-ray diffraction. The determination of molecular structure of matter by X-ray diffraction is the subject of X-ray crystallography.

X-ray crystallography provides a fast and reliable means of acquiring the results. In fact, the frequency of incoming results exceeds the present rate of data saving. In the European X-Ray Free-Electron Laser (XFEL), up to 3500 images are generated every second. Linac Coherent Light Source (LCLS) and Cornell-SLAC Pixel Array (CSPAD) detectors produce approximately 2.5 terabytes of data in one hour. On the other hand, a typical value of percentage of recorded images containing crystal diffraction, referred as hit fraction, is approximately 5-10%. In some experiments, hit fractions of values even lower than 0.1% have been observed.

The high rate of incoming data impels incorporating a pipeline for filtering out images that did not capture diffraction. Recently, deep learning was proposed as means of achieving it [1], [2]. However, deep learning methods come with a limitation of being a black-box model and therefore, their predictions are non-transparent and obscured from humans. Deep learning models for diffraction images are usually trained on simulated data or data coming from detector other than the target detector[1], [2], [3], [4], [5]. Model trained on such data may fail when implemented in real experimental setting. Selecting diffraction images is an online processing problem, hence the filtered out images cannot be recovered. In such a case, a model needs to be carefully evaluated before potential implementation.

In the recent years, a field of explainable artificial intelligence (XAI) has emerged. XAI is a group of methods for humans to understand how artificial intelligence (AI) model makes a decision. The role of XAI is to provide the methods for creating explainable models while preserving high level performance [6].

In this report, XAI algorithms were used to explain the models from published research trained on publicly available datasets for selecting diffraction images.

2 Serial crystallography

Serial Crystallography determines molecular structure by X-ray diffraction and is considered primary means of molecular structure determination. Around 90% of samples stored at Protein Data Bank were obtained by this method. Crystallography produces highly-detailed models of molecular structures, usually approaching atomic level of detail. Molecular structure is a key element of structure-based drug discovery and structural biology.

2.1 X-ray diffraction principles

The idea behind serial crystallography stems from double slit experiment with light, which was illustrated in Figure 1. The experiment shows that illuminating a pair of slits produces a series of light and dark bands on the screen behind it. Overlapping series of waves results in constructive or destructive interference. Constructive interference increases its intensity, whereas destructive decreases it. Light bands correspond to two waves with constructive

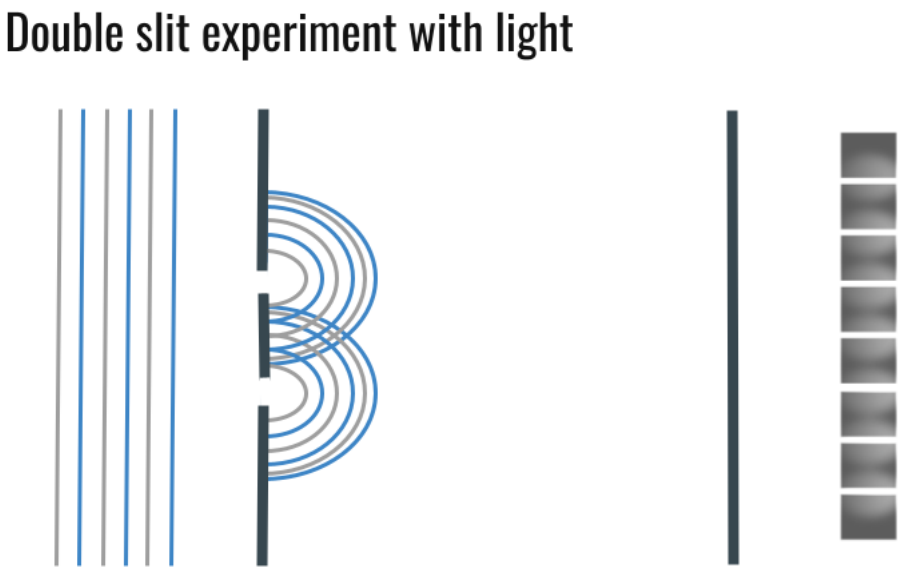

Figure 1: **Double slit experiment.** Light source illuminates a plate with 2 slits. The light passing through the slits is observed on a screen behind the plate. The light waves passing through the slits interfere with constructively or destructively, which corresponds to light and dark bands on the screen.

interference with resulting double amplitude. Dark bands correspond to destructive interference that canceled the pair of waves.

Double slit experiment is based on the principle that light has wavelength properties. The wavelength of X-rays is comparable to atomic size, which allows reaching near atomic level of detail of diffraction patterns.

One of the most famous examples of X-ray diffraction is the experiment of Rosalind Franklin, which shed light on the double helix structure of DNA.

Synchrotrons typically utilize Fraunhoffer diffraction, which requires the illuminated area of the sample to be minuscule compared to the distance to the source and to the detector. In serial crystallography, the protein sample is typically millimeters in size, whereas the distance from sample to the source and from sample to the detector can be of many meters. In Fraunhoffer diffraction, the incoming paths from the source to the sample and the outgoing paths from the sample to the screen are approximately parallel for a specific point on a screen. In crystal diffraction, this property results in strong constructive inference due to the regular arrangement of crystal planes. Natural crystals rarely self-assemble into regular arrangements, so the well-diffracting protein crystals need to be produced synthetically.

Figure 2 presents a diffraction image from protein crystal. The arrangements of spots in the diffraction image ascertains the arrangement of atoms or molecules inside a crystal. In the protein crystals, crystals consist of complex molecules. The crystal arrangement significantly boosts the X-ray diffraction signal compared to a single crystal molecule.

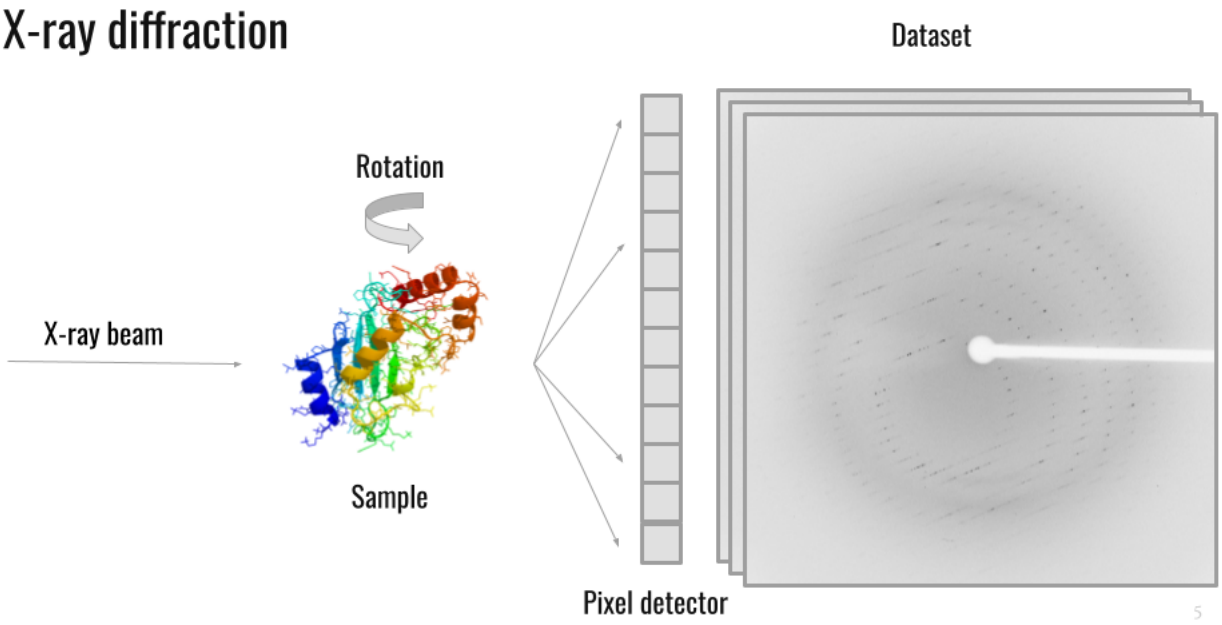

Figure 2: **X-ray diffraction.** Rotating crystalline structure hit by X-ray beam produces diffraction pattern on the screen of the detector.

However, utilizing protein crystals comes with a limitation. The positions of the spots in protein crystal diffraction images only convey the information of how the molecules are arranged within a crystal and not what the molecule itself looks like. The information about the structure of a single molecule can be retrieved from the intensities of the spots. In view of this facts, the pattern can be considered less important than the precise intensities of the spots.

Diffraction pattern present in the image is a two dimensional slice of a three dimensional Fourier transform of the electron density of the crystal. Therefore, the crystal needs to be rotated to obtain the complete information about its Fourier transform. To retrieve information about the molecule structure from spot intensities, the relation between the produced 2D slice and complete 3D transform needs to be determined. X-ray diffraction experiment is conducted with a rotating crystal and the resulting images are created without the information about the phase that the crystal was in when X-ray beam hit it. Due to the fact that crystal phase angles are not directly accessible after the experiment, they need to be supplied by additional experiments or prior knowledge. Once the three dimensional dataset is created, it is possible to determine the structure of a molecule.

The detectors for crystal diffraction need to meet specific requirements. The detector needs to have pixel density large enough to capture full diffraction pattern, while being not to large to keep the spots at a distance from each other so that they do not overlap. Moreover, it needs to be sensitive enough not to capture only the large spots towards the center but also the spots towards the edge of the detector. Otherwise, the produced crystal structure will be only a rough estimation of the actual structure.

3 Explainable Artificial Intelligence

Deep Learning is being widely implemented in various areas of life, including medicine, recommendation systems and virtual assistance [7]. Despite the proliferating adoption of its applications, it has a potentially dangerous limitation, referred to as the black box problem. The black box problem is related to the fact that even model's creators do not know on what basis a model makes a prediction.

To resolve the black box problem, explainable artificial intelligence (XAI) field was created. The role of XAI is to provide a set of tools for providing human-readable explanation of the prediction process to create a model that can be trusted.

For the analysis of the models for diffraction images, two sets of methods were selected. The first group of method is based on the idea of visualization of parts of image that were used by the model to make a prediction. The aim of the second is evaluate what information about the input image is preserved by the specific layer of the model.

3.1 Class Activation Mapping methods

Class activation mapping (CAM)[8] is a set of methods that are the extensions of class activation mapping. CAM methods are class-specific, which means they are able to produce a separate visualizations for each of the possible classes. They can be used to explain the models that were trained on images with labels, not necessarily on images with semantic segmentation maps.

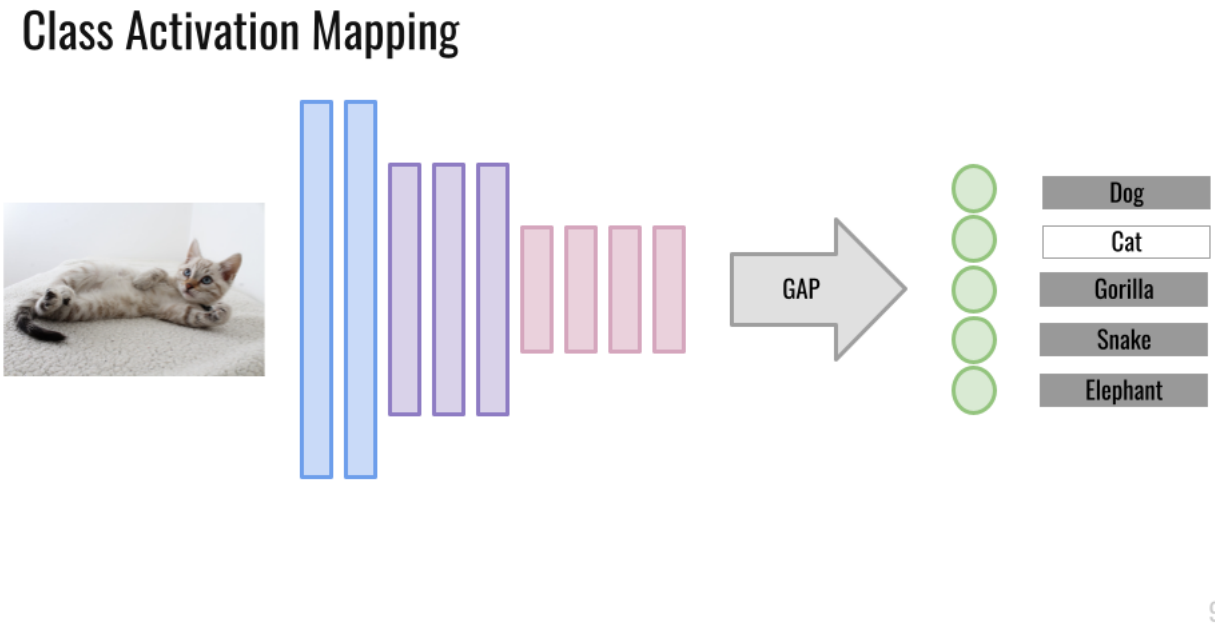

Figure 3: The structure of a neural network for image classification with a global average pooling operation required by class activation mapping.

3.1.1 Class Activation Mapping (CAM)

Class activation mapping (CAM) is the basis for the methods presented in following subsections.

Figure 3 illustrates a typical neural network for image classification. A typical neural network for image classification consists of two parts. The first part is a feature extractor composing of convolutional layers. Its role is to extract features from the image. The second part is build of fully-connected layers and is used to determine the likelihood of an image belonging to each of the classes based on the presence or absence of extracted features. To use the cam method, global average pooling operation needs to be inserted between the feature extractor and fully-connected layers. Global average pooling is based on the assumption that the last convolutional layer of the model extracts high level features and still preserves some spatial information. Global average pooling calculates the average value for each of the feature maps of the last layer of the feature extractor. The average value can be interpreted as success rate of extracting the feature of the filter. The feature maps can be colored according to the values yielded with gradient average pooling to form a heatmap. Juxtaposing the heatmap with the input image shows the importance of of image regions in making a prediction. A procedure for obtaining CAM images for a class with the highest score in the output layer is listed below.

1. Make a prediction on the image with the model.
2. Find class with the highest score,
3. Get the output of the final convolutional layer.
4. Apply global average pooling on the retrieved feature maps.
5. Color the feature maps according to their global average pooling score.
6. Reshape and project maps on the original image.

3.1.2 Gradient-based Class Activation Mapping (Grad-CAM)

Gradient-based Class Activation Mapping (Grad-CAM)[9] is an extension of CAM. Grad-CAM does not require global average pooling and can use any layer of the model, mot necessarily the last convolutional layer. The only condition is that the layer needs to be differentiable to make it possible to calculate the gradient. The importance of feature maps is assessed based on the alpha values. Alpha values are calculated based on the gradients.

1. Compute the gradient of the score for y^c (the output of the class before softmax) with the respect to the feature map activations A^k of a convolutional layer.
2. Average the gradients to calculate alpha values
3. Calculate the final Grad-CAM heatmap.
4. Reshape and project maps on the original image.

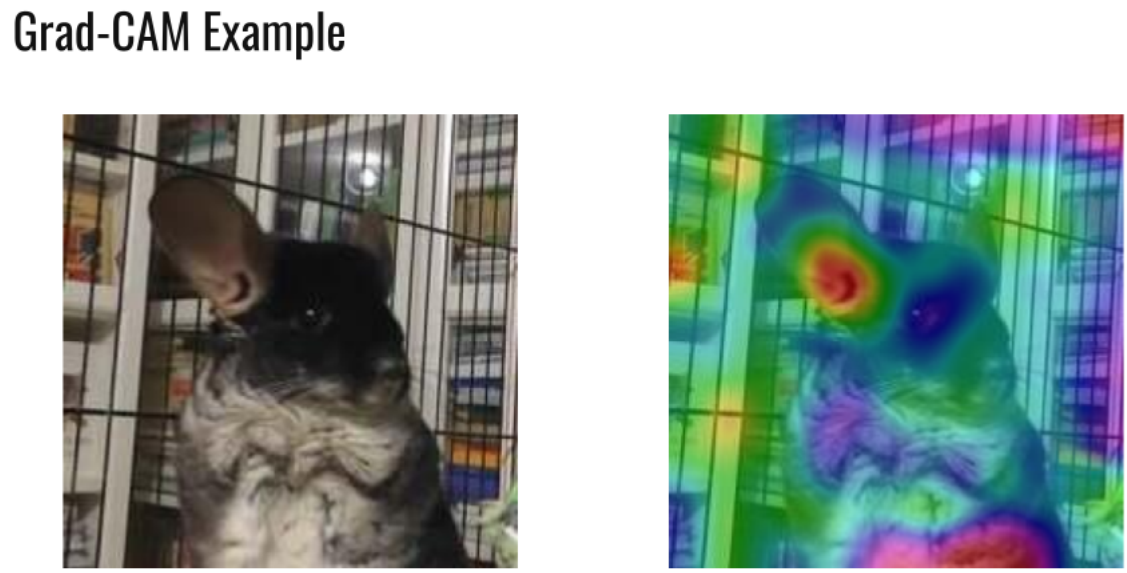

Figure 4: **The example of Grad-CAM.** Grad-CAM generates a heatmap with colors corresponding to the values of gradients associated with these regions.

3.1.3 Guided Gradient-based Class Activation Mapping (Guided Grad-CAM)

Guided Grad-CAM[9] is a combination of guided backpropagation and Grad-CAM methods. The difference between guided Grad-CAM and Grad-CAM is that with Grad-CAM we omit negative gradient values. It can be interpreted as taking into consideration only the features that a neuron detected and omit features that the features that a neuron did not detect.

3.2 Inverting Visual Representations

Inverting visual representations [10] is a method used to check if a specific layer retains the ability to reproduce the input image based on its output. The feature extracting part of the network is called an endcoder because it encodes the input image into a simplified representation. Inverted visual representations are created by attaching a decoder with deconvolutional upsampling layers to the encoder of the network. The first convolutional layers usually extract basic features, such as vertical edges, so typical inverted visual representations of the first convolutional layer recreate the input image with great level of detail. With the deeper layers, the inverted visual representation should get more abstract, however, it should preserve the key information that makes the image belong to the specific class. If the key information is missing, a model's quality can be questioned.

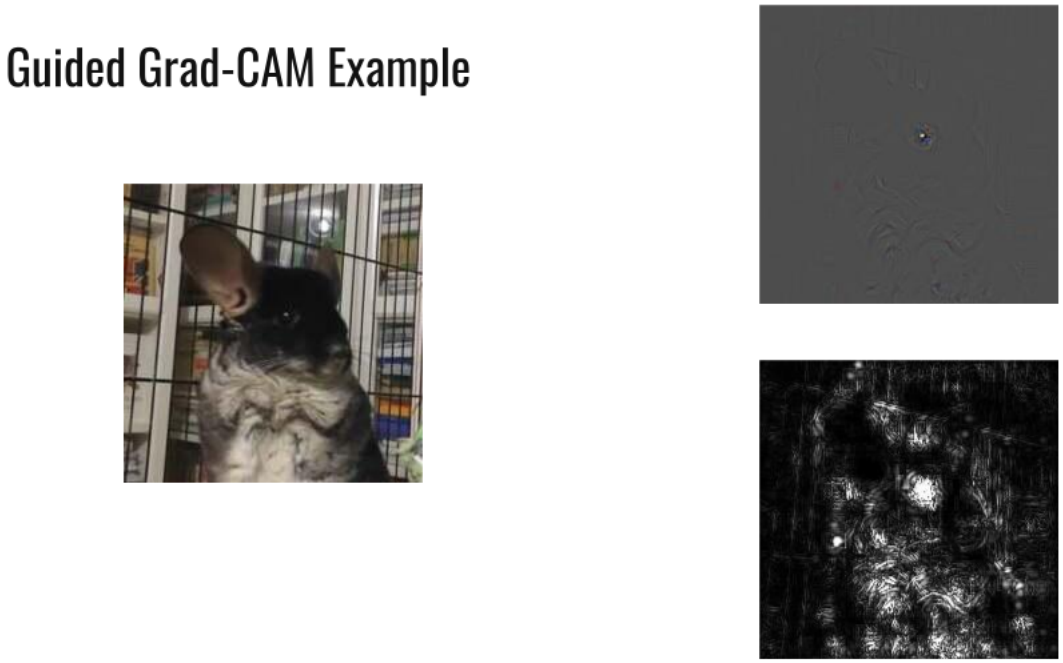

Figure 5: **The example of Guided Grad-CAM.** Guided Grad-CAM highlights the regions with large positive values of gradients.

4 Experiments and Results

In this paper, an AlexNet [11] model trained on the DiffraNet dataset [1] was analyzed with XAI methods. The reason behind selecting AlexNet was justified by the fact that diffraction patterns have simpler structure than natural images.

4.1 Dataset description

DiffraNet dataset is a publicly available dataset with 3 subsets.

- diffranet synthetic: contains images simulated with nanoBragg simulator belonging to 1 of the 5 classes: (blank, no-crystal, weak, good and strong). The dataset is divided into 3 subsets: train, val and test. Test set was used for tests.

- diffranet real raw: contains real raw images belonging to 1 of the 2 classes: (no diffraction, diffraction). The dataset is divided into 2 subsets: val and test. Test set was used for tests.

- Diffranet real preprocessed: contains real preprocessed images belonging to 1 of the 2 classes: (no diffraction, diffraction). Preprocessing was done by downsampling, cropping and removing beamstop shadow. The dataset is divided into 2 subsets: val and test. Test set was used for tests.

4.2 Tested models

The experiments were conducted with 3 variations of the model. The models are listed below.

- Model 1: AlexNet pretrained on ImageNet. Only fully connected layers were trained on diffranet synthetic dataset (train subset used for training, val subset for validation).
- Model 2: AlexNet pretrained on ImageNet. All layers were trained on diffranet synthetic dataset (train subset used for training, val subset for validation).
- Model 3: AlexNet pretrained on ImageNet. Then all layers were trained on a diffranet synthetic dataset (train subset used for training, val subset for validation). Then only fully-connected layers were trained on diffranet real preprocessed dataset (val subset used for training, test subset for validation).

5 Conclusions and Future Work

Deep learning models should be carefully analyzed before application in real world environment. The paper presented the model with AlexNet architecture. XAI methods shown that even shallow architecture, such as AlexNet, is too deep for the sparse structure of diffraction images. The original DiffraNet paper used ResNet50 architecture [**?**], which is significantly deeper than AlexNet. Implementing a model for selection of images with diffraction patterns without testing it with XAI methods, could have potentially resulted in excessive usage of resource without real need for it.

References

[1] A. Souza, L. B. Oliveira, S. Hollatz, M. Feldman, K. Olukotun, J. M. Holton, A. E. Cohen, and L. Nardi, "Deepfreak: learning crystallography diffraction patterns with automated machine learning," *arXiv preprint arXiv:1904.11834*, 2019.

[2] A. Czyzewski, F. Krawiec, D. Brzezinski, P. J. Porebski, and W. Minor, "Detecting anomalies in x-ray diffraction images using convolutional neural networks," *Expert Systems with Applications*, vol. 174, p. 114740, 2021.

[3] Z. Ding, E. Pascal, and M. De Graef, "Indexing of electron back-scatter diffraction patterns using a convolutional neural network," *Acta Materialia*, vol. 199, pp. 370–382, 2020.

[4] R. Liu, A. Agrawal, W.-k. Liao, A. Choudhary, and M. De Graef, "Materials discovery: Understanding polycrystals from large-scale electron patterns," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2261–2269, IEEE, 2016.

[5] J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh, and K.-S. Sohn, "A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[6] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52138–52160, 2018.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[10] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4829–4837, 2016.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.