How Do Neural Networks See New Things?

Zero-Shot Learning Meets Explainable AI

Hoki Fung, Guillaume Delepoulle, Ng Jian Lai, Anand Sundaram

School of Computing, National University of Singapore

Background

In recent years, deep neural network models have achieved outstanding, human-level performance in image classification tasks [1]. However, supervised learning methods rely heavily on the availability of labelled data and the image classes that the models can recognize are limited to those they were trained on, which make these models less useful in real-life scenarios.

To overcome this limitation, **Zero-Shot Learning** (ZSL) [2], a problem setting in machine learning where we want the model to classify objects from classes that it has not seen during the training phase, has been proposed. State-of-the-art ZSL models can often achieve a top-1 accuracy at around 60% for images in unseen classes [3].

Objective

The objective of our project is to bring the concept of **Explainable AI** (XAI), an emerging field in artificial intelligence that sets out to address how black box decisions of AI systems are made, to zero-shot learning. In short, we aim to understand how ZSL models classify images from classes they have not encountered before.

In this project, we visually examine the features of unseen images that lead to correct and incorrect classifications using a technique called **Gradient-weighted Class Activation Mapping** (Grad-CAM) [4].

Data

We curated a subset of images and annotations from the Large-scale Attribute Dataset [5], and selected 15 classes for training, and 5 zero-shot (unseen) classes for testing.

Methods: Zero-Shot Learning & Grad-CAM

1. Zero-Shot Learning

We adopted an embedding-based (as opposed to generative) approach. Our model had access to labelled images from “seen” classes and auxiliary information for both “seen” and “unseen” classes during the training phase (inductive as opposed to transductive) , and the images to be recognized at test time belong only to unseen classes (conventional as opposed to generalized).

Auxiliary Information – Class Embedding

This information consists of semantic descriptions or attributes or word embeddings of the image classes. They are real-valued vector representations of the class labels that act as a bridge between the ”seen” and “unseen” classes.

Model – Feature Extraction & Projection Network

We first used a pre-trained convolutional model, VGG16, to extract image features. We then followed up with a small fully-connected projection networks that maps the N dimensional visual space input to a D dimensional output In the semantic space.

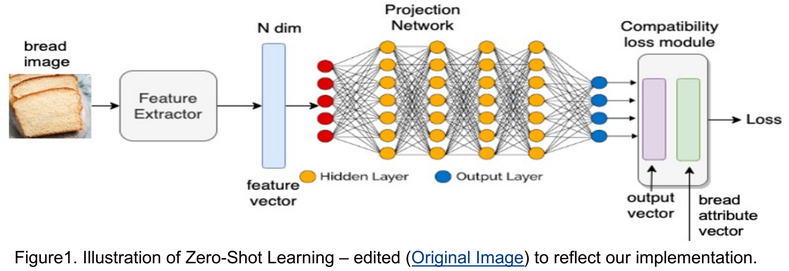


Figure1. Illustration of Zero-Shot Learning – edited (Original Image) to reflect our implementation.

2. Grad-CAM

Grad-CAM is a form of post-hoc attention. It uses the gradients of the target classes flowing into the final convolutional layer of a trained neural network to produce a heatmap highlighting the important regions used in the prediction. The visualization provides visual insights into the model decisions and help explain model performance.

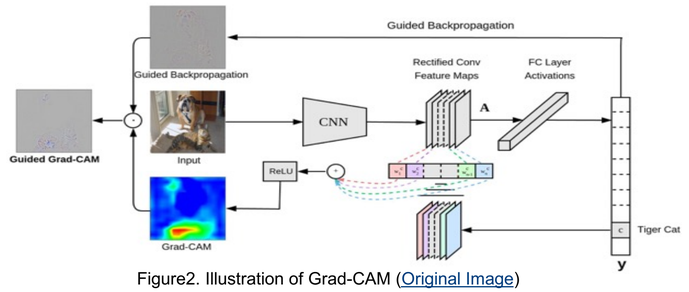


Figure2. Illustration of Grad-CAM (Original Image)

Procedures

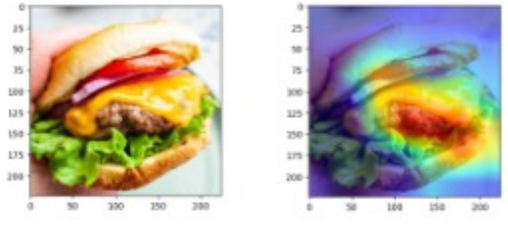
Step 1: Back Propagation

Step 2: Global Average Pooling

Step 3: ReLU(Weighted combination of feature maps)

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \longrightarrow L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Results



--- Top 5 Predictions ---

- Sandwich (seen)
- Bread (seen)
- Chicken (seen)
- Food (unseen)
- Child (seen)

Conclusions

We used Grad-CAM to visualize how a zero-shot learning model “sees” new classes of images. Our next step would be to examine the difference between a good prediction and a bad prediction. We would also like to see if using a better model (e.g. TF-VAEGAN) would give us different results.