Explainable deep learning-based classification on small datasets

Members: Zong Fan, Xiaobai Li

1. Project description and goals

Background: Deep learning (DL)-based techniques have achieved a giant success in various image classification challenges. However, unlike traditional machine learning methods such as decision tree and logistic regression which can easily interpret the importance of each understandable feature, it's difficult for DL techniques to explicitly explain the meaning of the extracted feature. In some degree, DL works as a black box that we cannot analyze how the model predicts the result based on the given data. Understanding which region of a figure that makes the network predict is important for further improving fine-grained classification whose inter-class variance is very small. What's more, interpretability is critical in medical applications such as tumor diagnosis and disease grading, where security is extremely addressed.

A saliency map is a heatmap corresponding to a figure that highlights the potential region of interest (ROI) where humans probably focus on. It's a useful tool in computer vision that makes the classification explainable. Many studies have been proposed to predict the saliency map in DL-based image classification tasks. For example, Karen et.al used weights of intermediate layers to visualize the saliency map which indicates the importance of each pixel to the class [1]. It's a widely-used method but is an unlearning method. Ramprasaath et al. proposed gradient-weighted class activation mapping (Grad-CAM) which uses the gradient of target concept to visualize the saliency map [2].

Goals: In order to predict a more accurate saliency map, location annotations such as bounding box and segmentation could be employed if they are provided in the datasets. So in this study, we'd like to investigate the use of bounding box or segmentation for achieving better object saliency map in classification tasks. Here are the general outlines for this project.

Experiment outline:

- **Dataset search**: Find relatively large datasets with bounding box or segmentation annotation. Also, find some small datasets without such annotations.
- **Network architecture design:** To make the classification network predict saliency map while predicting class labels under the supervision of location annotation, two branches are desired in this framework - the classification branch and the saliency map prediction branch.
- **Training:** Train the network on the selected dataset.
- **Evaluation**: Try different architectures in the two branches and evaluate their performance on the testing datasets, in order to find the optimal design of the classification framework. Visualize the saliency map to check whether the highlights match the focus regions of the object.
- **Comparison study:** Use classic classification networks to train the network on the datasets without the use of location annotations. It's used to evaluate whether the model trained with object location information could achieve higher performance than the normal classification method. Also, replicate existing methods to get saliency maps that are trained on our selected dataset. Compare their effects with ours.

Refrence:

[1] Simonyan, Karen et al. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." *CoRR* abs/1312.6034 (2014): n. pag.

[2] Selvaraju, Ramprasaath R. et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *International Journal of Computer Vision* 128 (2019): 336-359.

2. Member roles

Zong Fan mainly gets involved in the network design, training, and evaluation.

Xiaobai Li mainly focuses on virtualizing saliency maps and evaluation.

3. Resource

1) Dataset candidates:

1. Caltech-UCSD Birds 200 dataset

Link: http://www.vision.caltech.edu/visipedia/CUB-200.html

The dataset has 200 kinds of birds with both bounding box and segmentation annotations classifying different attributes.

Bird image and annotation samples:

2. Stanford Cars dataset

Link: https://ai.stanford.edu/~jkrause/cars/car_dataset.html

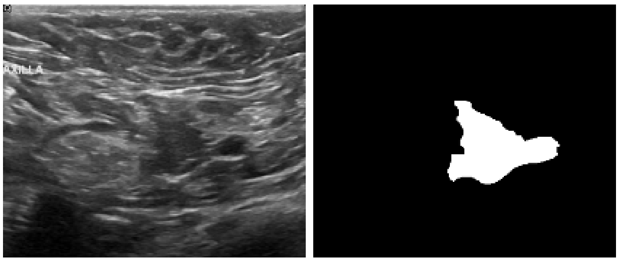
It contains 16,185 images of 196 classes of cars. Some of them have bounding box information.

Car image and annotation samples:

3. Breast Tumor Segmentation (BUSI) dataset

Link: https://scholar.cu.edu.eg/?q=afahmy/pages/dataset

It contains ultrasound images with breast tumor segmentation and classification annotation.



2) Computing source:

This project involves deep neural network training which is relatively computationally intensive. The potential training platform includes the following choice.

1. campus cluster with GPU node;
2. personal computer with single NVIDIA GPU card.

3) Implementation platform

Mainly use PyTorch and official Torchvision package for coding.

Code repositories as comparison include:

Grad-CAM: https://github.com/jacobgil/pytorch-grad-cam

and potential other repositories in the future experiment.

4. Reservations

One of the biggest problems in this study is that the location annotated datasets are relatively small scale and not widely available in many applications. This is because location annotation especially segmentation annotation needs large efforts and expertise knowledge to label. Some applications even have difficulties in obtaining large amounts of images. Small-scale datasets without any location labeling are quite common situations.

To solve this problem, one potential way is to find relatively large datasets with similar classes that have location labeling information. Then train the network capturing the saliency map on this dataset. After training, fine-tune it on the target dataset and check whether the saliency works.

Another concern is that the performance of using the location annotation information cannot achieve better performance compared to weight or weight gradient-based methods without the explicit need for location annotation.

5. Relation to your background

Zong Fan is a BIOE student whose major research interests lie in medical imaging reconstruction and synthesis. He is familiar with deep neural network implementations and useful computer vision packages such as PyTorch, TensorFlow, OpenCV, numpy, etc. This project has little relationship with his major researches, but the point of interpretability of classification interests him a lot, since confidence in deep learning models is very critical in medical applications.

Xiaobai Li is a ECE student whose major research focus is on machine learning applications on Healthcare or Autonomous vehicles. He is not familiar with Computer Vision but understands the logic of deep neural networks. This course is his first course related to Computer Vision, and this project is not really related to his research focus but the concept of CV is very interesting and helpful for his research in the future.