Kaggle Project (compte-rendu)

Deep learning models are widely used for image classification. The goal of this project is to compare state-of-the-art models on a subset of ImageNet dataset. The link of the project is : `https://www.kaggle.com/competitions/modia-ml-2024/overview`

Notice : You should upload a report, together with your code to Moodle, for the validation of the course. Do not copy-paste other people's report or code.

Due : 29 Juin 2024.

1 Learn how to use Pytorch

Refer to the website `https://pytorch.org/tutorials/beginner/basics/intro.html`.

2 Train state-of-the-art CNN models

The goal is to achieve a good classification accuracy on the test dataset, using Pytorch.

- Implement one or two models from the following list :
  - LeNet Model [1]
  - AlexNet Model [2]
  - ResNet Model [3]
- In your report, you should give a precise definition of the model which you use, e.g. the number of layers and the type of each layer in CNN.

- Pre-process your images in black and white

- Pre-process your images so that they have the same input size to your model, e.g. use data augmentation.

- Train your model using mini-batch SGD. Specify the optimization method which you use and report the total training time. To reduce the training time, you may use a GPU card.

- Perform your parameter turning on a validation set to avoid over-fitting. Summarize your results in table/figure.

3 Machine learning and AI safety

3.1 Bias introduction in ImageNet

We try to understand how an image classification model (such as the ones trained in section 2) can be biased towards a specific outcome. To make things easier we consider the binary classification problem (2 classes).

We propose to implement the following setup :

3.1.1 Biased Dataset Construction

We decide to concatenate to each of our images x from ImageNet a new set of variables ϵ. As we shall specify, the second variable ϵ is a noise-like image, which is not directly computed from x. Yet, it introduces some bias as we choose the value of ϵ to be strongly correlated to the label (0 or 1) of the images.

More precisely, we assume $(\widetilde{x}, y) = ([x, \epsilon], y)$ is a random sample of the modified dataset. We shall specify the value of ϵ using y. Let $p_0 \in [0, 1]$, $p_1 \in [0, 1]$ be two probabilities. Given $(\widetilde{x}, y)$, the bias variable S is defined as

$$S \sim Bernoulli(p_k), \quad \text{if } y = k.$$

Then we choose ϵ according to S as below :

— $\epsilon = 0$ if $S = 0$

— $\epsilon \sim \mathcal{N}(0, I)$ if $S = 1$

In the extreme case where $p_0 = 0$ and $p_1 = 1$, one can ignore the original image x and only use ϵ to predict y. In that case, our predictions are never using the image x and we are only relying on the noise-like ϵ that we introduced to the dataset.

Advice : in Pytorch, we can build a biased dataset by simply adding the biased variables $\epsilon^{\{i\}}$ as a dedicated channel to the original images $x^{\{i\}}$ for each image index i.

Question 1 : Build a biased dataset from ImageNet using the approach described in 3.1.1

3.1.2 Model Bias Evaluation

We compare two different settings :

— $model_1$ is trained on the unaltered ImageNet dataset

— $model_2$ is trained on the biased version of ImageNet (where we append the biased variables ϵ to the images in a separate channel)

We finally compare the actual bias in both $model_1$ and $model_2$.

Assume $\widehat{y}$ is the prediction of a model based on $\widetilde{x}$. We shall separate the dataset set into 2 groups, one with $S = 0$, the other with $S = 1$. The bias of this model can be computed from a ratio of these two groups, using the DI metric [4] defined as :

$$DI = \frac{P(\widehat{y} = 1|S = 0)}{P(\widehat{y} = 1|S = 1)}$$

In practice, a model is considered unbiased as long as the DI metric is close to 1.

Question 2 : Study experimentally if $model_2$ is biased or not on the test data. Compare you results to what you have with $model_1$.

Question 3 : Compare the accuracy scores on your test data for both $model_1$ and $model_2$ for the binary classification task.

Question 4 : What can you conclude from the two previous questions ?

3.2 Bias Study in the Literature

Bonus Question : This article gives an overview of current AI system requirement in EU [4]. Read it carefully and write a short essay (about one page) through your critical thinking on "Est-ce qu'on peut faire confiance à mon modèle ?" from the perspective of model bias.

You can also use other articles in the literature. More broadly, you may refer to the reports of CNIL [5].

Références

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6) :84–90, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] `https://hal.science/hal-03253111`.

[5] `https://www.cnil.fr/fr/intelligence-artificielle/guide/conformite-des-systemes-dia-les-autres-guides-outils-et-bonnes-pratiques`.