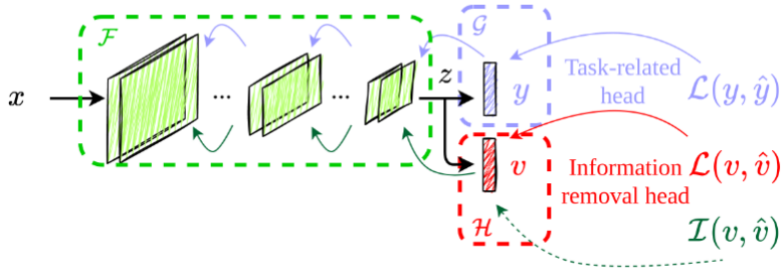
Unsupervised online debiasing in deep neural networks

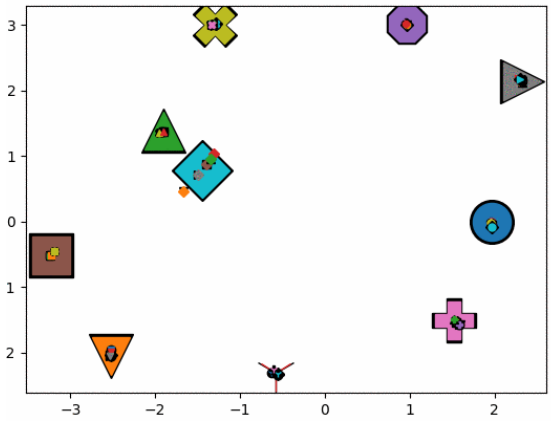
Overview

Currently, a significantly large portion of problems is being solved through the deployment of deep learning models, considered by most the “universal problem solving tool” [1]. For instance, these are being deployed in high-stakes applications, ranging from candidate job hiring to facial recognition systems. It is a known problem that deep models, when trained, take advantage of “spurious” correlations from their training data, which lead to significant performance variance across sub-populations, sometimes across sensitive attributes like race and gender. This causes what is known, in the literature, as bias of the deep model.
Learning these spurious correlations has several effects, of which the most evident one is poor performance on under-represented dataset sub-populations and out-of-distribution test data [2]. Finding a solution to the problem of biases in the deep models is currently a topic of broad interest from the community [3]. Many works have tried to address this problem, ranging from addressing the problem to transformers [4] to unsupervised scenarios [5]. A large part of approaches intrinsically propose a re-weighting over the biased information, why not to entirely remove it?



In a recent work [6], a method towards information removal has been proposed, in which irrelevant information related to some target task, which can naturally be propagated through the model, is really eliminated. This is accomplished by training a secondary head, attached to the main model, whose aim is to estimate how much information penetrates to some “bottleneck” of the deep model. This information is then erased maximizing the confusion over the private information. Surprisingly, when testing such an approach in well-known debiasing benchmarks, there are some cases in which the performance on the original task, instead of deteriorating (as we subtract some information), actually improves, as the removed information was biasing the model itself.
Leveraging on this observation, it is possible to design a debiasing strategy where the information is not necessarily erased as many strategies claim, but simply weighted. In most real-life cases, biased sources are unknown; hence, it is impossible to address the problem of debiasing in a supervised scenario. This poses a question: how can we recognize a bias?

Let us plot the features in the bottleneck of the model (figure on the right) - the big symbols are the centroids for the target class prediction. The tiny remaining symbols are the misclassified samples, where their color indicates the ground-truth class and the symbol indicates on the contrary the presence of a specific biased feature which leads to their misclassification. Also the “big symbols” have a dominant bias component aligned, which acts as “attractor” for the misclassified samples.
The goal of the proposed project is, besides studying and replicating state-of-the-art debiasing approaches, the implementation of an unsupervised online approach for automatic bias correction.



Objectives at a glance

- Study the most recent state-of-the-art approaches around debiasing methods in CNNs.
- Understand the underlying learning mechanisms involved in common debiasing strategies.
- Replicate/reproduce state-of-the-art results, with partial or total re-implementation of the approaches.
- Development of an unsupervised online debiasing strategy for image classification.

Tutor

Enzo Tartaglione, MdC for the Multimedia equipe enzo.tartaglione@telecom-paris.fr

References

[1] Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. Applied and Computational Harmonic Analysis, 43(2):233–268, 2017.
[2] Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In British Machine Vision Conference (BMVC), volume 1, page 3, 2021.
[3] P. Henriksen, Kerstin Hammernik, Daniel Rueckert, and Alessio Lomuscio. Bias field robustness verification of large neural image classifiers. In BMVC, 2021
[4] Sruthi Sudhakar, Viraj Prabhu, Arvindkumar Krishnakumar, and Judy Hoffman. Mitigating bias in visual transformers via targeted alignment. In BMVC, 2021
[5] Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In British Machine Vision Conference (BMVC), volume 1, page 3, 2021
[6] E. Tartaglione. Information removal at the bottleneck in deep neural networks. 2022