Explaining Local Discrepancies between Image Classification Models

Thibault Laugel
Xavier Renard
AXA, TRAIL
{thibault.laugel, xavier.renard}@axa.com

Marcin Detyniecki
AXA, TRAIL
Polish Academy of Science, Warsaw, Poland

Abstract

When optimizing for a predictive objective, multiple models often achieve the same performance, albeit with significantly different behaviors. Yet, understanding these differences is often overlooked, resulting in the arbitrary selection of one model over its competitors. To help practitioners deploy more ML pipelines more responsibly, we leverage recent works developed for tabular data and adapt it to the context of image classification. This extension addresses in particular the issue of generating model-agnostic explanations based on understandable features using variational autoencoders.

1. Introduction

When designing machine learning systems, ML practitioners rely on global performance metrics to select their best model among various candidates, often overshadowing their differences and the issues some of these can have.This model selection, in some aspects arbitrary, with its potential hazardous consequences has been the focus of several works, proposing for instance metrics to help practitioners quantify the degree of model variability at training time [4]. To further help dealing with the issue, DIG [5] was recently proposed as a hands-on tool to explain local differences between trained models for tabular data.

In this extended abstract, we propose an extension of DIG to the task of image classification with the goal of generating model-agnostic explanations of local differences in the classification behaviour of trained models that achieve similar predictive performance. The main challenge is to enable an efficient search for differences in classification behaviours and avoid an intractable search among the space of all images. To overcome this issue, we propose an approach based on learning features in an unsupervised manner using a VAE and adapting DIG to apply it to the learned representation.

2. Investigating Disagreements using DIG

Investigating differences in the local classification behaviours of trained classifiers is a poorly covered topic. A notable exception is DIG [5], a hands-on model agnostic tool that was recently developed for this purpose, and that we chose to leverage for image classification. In this section, we give a quick overview of the DIG algorithm, then discuss some of its shortcomings that we try address with our proposition.

2.1. An Overview of the DIG Algorithm

Given a set of models $\{f_k\}_{k \in K}$ trained on a dataset $X \subset \mathcal{X}$ and achieving similar predictive performance over a validation set, the objective of DIG is to approximate and explain *discrepancy areas* $\mathcal{D}$, i.e. regions of the input space where models have diverging predictions.

To discover these regions in a model-agnostic setting, DIG proposes a heuristic with a particular sampling of instances to efficiently explore the input space and detect regions where predictions change across models. This exploration is conducted along *counterfactual directions*, i.e. segments I delimited by training instances and their closest neighbors associated with a different predictions: $I = [x_i, x_j] s.t. x_i, x_j \in X$ and $\exists k : f_k(x_i) \neq f_k(x_j)$
These segments I are then iteratively refined through dichotomic search to delimit up to a desired level of precision the local discrepancy area, i.e. $I \cap \mathcal{D}$. In the end, the final explanation returned for a prediction $f(x)$ is a local interval, i.e. two instances close to x and belonging to I that delimits the borders of the discrepancy area.

2.2. Shortcomings Addressed

Designed for tabular data, DIG relies on sampling. Hence, applying it to the task of image classification raises two major issues, commonly encountered in high dimensional problems. First, the defined sampling along "counterfactual directions" (i.e. straight lines) is not appropriate in a high dimensional setting, resulting in non-realistic examples being generated. Second, a related issue is that the

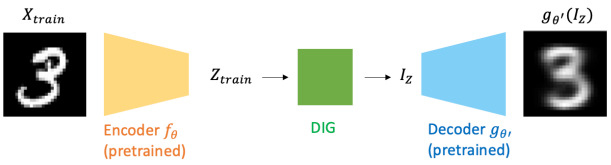

Figure 1. Proposed architecture for DIG-CV.

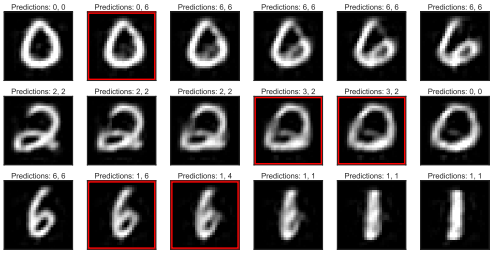

Figure 2. Results obtained for three instances of MNIST.

feature-by-feature explanation in the input space proposed by DIG is not actionable in the case of images (pixels are not naturally understandable).

These issues have been generally raised when trying to adapt model-agnostic explainability methods to image classification models. For instance, LIME [6] proposes to split the images into superpixels and use these as features, but this does not allow continuous transformations from one instance to another which is one benefit offered by DIG. Other methods have considered autoencoders [1], which we propose to use to extend DIG, as described in the next section.

3. Discrepancy in Computer Vision: DIG-CV

To circumvent these limitations, we propose to adapt DIG by adding a step of feature learning using variational autoencoders, more particularly a β-VAE [2]. This choice can be motivated by several reasons. First, the β-VAE loss function has been shown to favor the extraction of meaningful concepts. Second, transitioning from one image to the other by connecting in a straight line the representations learned by variational autoencoders in general has been already shown to generate meaningful images (see e.g. [3] for MNIST). The proposed architecture is represented in Figure 1. The exploration phase of the DIG algorithm is performed in the latent space Z, with the assessment of discrepancy being performed on the reconstructions of these generated instances: $f_i(g_{\theta'}(I))$.

4. Results and Discussion

In this section, we show the first results that were obtained by our method on two datasets, MNIST and FashionMNIST, on which we train two classifiers, a convolutional network and a SVM classifier.Although the accuracy difference between these models is relatively small (0.01% on MNIST and 0.03 on FMNIST), we measure that the models are disagreeing respectively over 2.23% and XX of the instances of the tests sets.

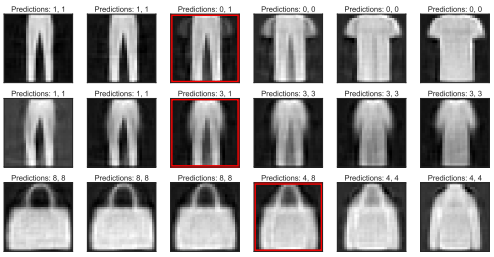

Figure 3. Results obtained for three instances of FashionMNIST.

To investigate these disagreements, we apply DIG-CV to both datasets and show some illustrative results in Figures 2 and 3. On the extreme left and right columns of each row are shown the reconstruction x' of two instances x from the training set. In-between are shown reconstructions $g_{\theta'}(I)$ of the instances sampled in the latent space Z by DIG. The instances whose reconstruction falls in a discrepancy area are highlighted using a red square around the image. Using these explanations, the practitioner can detect uncertain areas of the feature space, and depending on the case, perform remediating actions. Such actions include for instance labelling of more data in these uncertain areas, or asking for a human in the loop to perform the decision if models can not agree.

Pursuing this work, we aim in the future at conducting proper evaluation of the quality of the discrepancy area detection performed by DIG-CV, and usecases to illustrate how these explanations can be leveraged.

References

[1] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. Black box explanation by learning image exemplars in the latent feature space. In *Joint European Conf. on ML and KD in Databases*, pages 189–205, 2019. 2

[2] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016. 2

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[4] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proc. of the 37th Int. Conf. on ML*, volume 119, pages 6765–6774, 2020. 1

[5] Xavier Renard, Thibault Laugel, and Marcin Detyniecki. Understanding prediction discrepancies in machine learning classifiers. *preprint arXiv:2104.05467*, 2022. 1

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any

classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on KD and Data Mining*, pages 1135–1144, 2016. 2