Probabilistic Graphical Models
Are Generative Classifiers More Robust to Adversarial Attacks?

Manal Akhannouss, ENS Paris-Saclay
manal.akhannouss@eleves.enpc.fr

Paul Barbier, ENS Paris-Saclay
paul.barbier@eleves.enpc.fr

Alexandre Lutt, ENS Paris-Saclay
alexandre.lutt@eleves.enpc.fr

I. Introduction

II. Models

II.1. Discriminative versus generative models

Let us consider a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ of N samples, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $\boldsymbol{y}_i \in \mathcal{Y}$ is the corresponding label. A discriminative classification model (or discriminative classifier) aims to estimate the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$, *i.e.* the probability that the label $\boldsymbol{y}$ is associated to the feature vector $\boldsymbol{x}$. On the other hand, a generative classification model (or generative classifier) aims to estimate the joint probability $p(\boldsymbol{x}, \boldsymbol{y})$, *i.e.* the probability of observing the feature vector $\boldsymbol{x}$ and the label $\boldsymbol{y}$ at the same time. Both models can be used for classification purposes, *i.e.* can be used to predict the label $\boldsymbol{y}$ associated to a feature vector $\boldsymbol{x}$, but with very different interpretations. The discriminative classifier will directly estimate the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$, while the generative classifier will estimate the joint probability $p(\boldsymbol{x}, \boldsymbol{y})$ for each possible value of $\boldsymbol{y}$, and then use the Bayes rule to estimate the conditional probability $p(\boldsymbol{y}|x) = \dfrac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$.

One of the most common generative classifier is Naive Bayes. This simple model assumes a factorised distribution $p(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^{d} p(\boldsymbol{x}_i|\boldsymbol{y})$, which means that the features are independent given the label. This assumption is often far from being verified in practice for image datasets. For this reason, in the following, we will follow the path of [1] and use a latent-variable model $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ to design our generative classifier. Note that this model does not assume a factorised distribution for $p(\boldsymbol{x}|\boldsymbol{y})$, since in this case $p(\boldsymbol{x}|\boldsymbol{y}) = \dfrac{\int p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})d\boldsymbol{z}}{\int p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})d\boldsymbol{x}d\boldsymbol{y}}$. In order to fully define a latent-variable model, we need to explicitely chose a structure for $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. At this point, several choices are possible:

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p_{\mathcal{D}}(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) \tag{DFX}$$
$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) \tag{DFZ}$$
$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p_{\mathcal{D}}(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{z}) \tag{DBX}$$
$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) \tag{GFZ}$$
$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p_{\mathcal{D}}(\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) \tag{GFY}$$
$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \tag{GBZ}$$
$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p_{\mathcal{D}}(\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{z}) \tag{GBY}$$

In those acronyms, D stands for *discriminative*, G stands for *generative*, F stands for *fully-connected graph*, and the last letter indicates on which variable we assume a prior distribution (determined with $\mathcal{D}$ in the case of X and Y). In our case, we will focus on the DFZ and GFZ structures, in order to be able to compare the discriminative and generative approaches, but everything that we will see can easily be extended to the other structures.

II.2. Classifiers architecture

As mentionned above, we will consider two different classifiers. The first one will be a simple discriminative classifier (with DFZ structure), while the second one will be a generative classifier (with GFZ structure).

III. Adversarial attacks

In this section, we will present the different methods we used to create adversarial attacks on our models. These methods can be divided into two categories: white box attacks and black box attacks, depending on the knowledge of the attacker regarding the model. If the attacker has access to the model's parameters and architecture, we will talk about white box attacks. Otherwise, we will talk about black box attacks.

III.1. White box attacks

III.2. Black box attacks

IV. Experimental setup

IV.1. Attacks detection

IV.2. Models training

IV.3. Attacks benchmark

IV.4. Attacks detection

V. Results

V.1. Accuracy

V.2. Robustness to perturbations

V.3. Attacks detection

VI. Conclusion

VII. Appendix

References

[1] LI, Y., BRADSHAW, J., AND SHARMA, Y. Are generative classifiers more robust to adversarial attacks? 1