Toward Generative Data Augmentation for Traffic Classification

Chao Wang
Huawei Technologies SASU, France
wang.chao3@huawei.com

Alessandro Finamore
Huawei Technologies SASU, France
alessandro.finamore@huawei.com

Pietro Michiardi
Eurecom, France
pietro.michiardi@eurecom.fr

Massimo Gallo
Huawei Technologies SASU, France
massimo.gallo@huawei.com

Dario Rossi
Huawei Technologies SASU, France
dario.rossi@huawei.com

ABSTRACT

Data Augmentation (DA)—augmenting training data with synthetic samples—is wildly adopted in Computer Vision (CV) to improve models performance. Conversely, DA has not been yet popularized in networking use cases, including Traffic Classification (TC). In this work, we present a preliminary study of 14 hand-crafted DAs applied on the MIRAGE19 dataset. Our results (*i*) show that DA can reap benefits previously unexplored in TC and (*ii*) foster a research agenda on the use of *generative models* to automate DA design.

1 INTRODUCTION

Network monitoring is at the core of networks operations with Traffic Classification (TC) being key for traffic management. Traditional Deep Packet Inspection (DPI) techniques, i.e., classifying traffic with rules related to packets content, is nowadays challenged by the growth in adoption of TLS/DNSSEC/HTTPS encryption. Despite the quest for alternative solutions to DPI already sparked three decades ago with the first Machine Learning (ML) models based on packet and flow features, a renewed thrust is fueled today by the rise of Deep Learning (DL), with abundant TC literature reusing/adapting Computer Vision (CV) training algorithms and model architectures [1].

In this work, we argue that *opportunities laying in the data itself are still underexplored*, based on two observations. First, CV and Natural Language Processing (NLP) adopt "cheap" *Data Augmentation (DA)* strategies (e.g., image rotation or synonym replacement) for improving models performance. Yet, almost no TC literature investigates DA. Second, *network traffic datasets are imbalanced in nature* due to app/service popularity skew, which calls for strategies to augment the minority classes. Again, the interplay between imbalance and model performance is typically ignored in TC literature.

In this paper, we propose a two-fold research agenda: (*a*) first, we study hand-crafted DA to assess its benefits and relationship with class imbalance (Sec. 2); then, (*b*) we charter a roadmap to pursue better augmentation strategies via *generative models*, i.e., learning DA in a data-driven fashion rather than adopting manual design (Sec. 3).

2 HAND-CRAFTED AUGMENTATIONS

We define hand-crafted DA as the family of transformations that can be described by simple mathematical formulations or algorithms, e.g., additive noise, random masking, and interpolation between samples (just to name a few). Such transformations are meant to be *directly* used in the **input space**, thus their design requires domain knowledge to control samples variety—too little produces simple duplicates; too much breaks class semantics and introduces undesired data shifts. At the same time, such DA aims at *indirectly* modifying DL classifiers decision boundaries in the **latent space**. Indeed, to be beneficial, DA should introduce additional training points that foster better clustering of the classes in the latent space. However, without explicitly knowledge of how input samples are projected in the latent space (as models are "black boxes"), domain knowledge hardly suffices for effectively designing these transformations—the use of DA is a trial and error process. Moreover, even if many studies investigated DA for CV and time series (e.g., in the medical field), reusing such methods is not trivial for TC as data suffers from two extra undesirable restrictions: *input samples are short*—traditionally, they are time series of the first N packets of a flow (e.g., the first 10 packet sizes and inter arrival times)[1] —and are *semantically weak*—interpreting packet time series is less obvious than interpreting electrocardiograms.

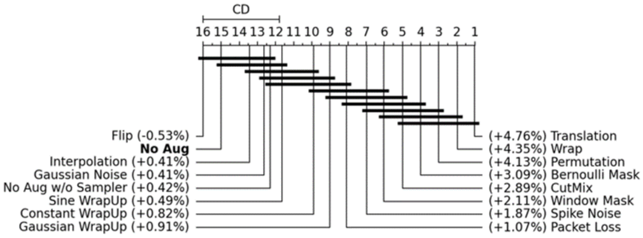
2.1 Our preliminary study

Inspired by time-series DA literature [2], we formulated a preliminary empirical study based on the MIRAGE19[2] dataset which gathers traffic from 20 Android apps. We benchmarked 14 augmentations applied on the first 20 packets size, direction and Inter Arrival Time (IAT) for each flow. Each augmentation is applied once to each batch of size B hence doubling its size to $2B$.[3] We compared against models trained without DA with batches of size $2B$. Given the imbalance, we use a

[1] Focusing on the first packets is required to enable early classification and effectively enforce traffic management policies.

[2] https://traffic.comics.unina.it/mirage/mirage-2019.html

[3] At the end of a training epoch a model has observed the original version of the dataset and a randomized version of all the training samples. Different epochs generate different randomized versions of the data.



	Aug. Name	Alter Feat.			Description
		size	*dir*	*iat*	
	No Aug	○	○	○	Reference baseline
	No Aug w/o Sampler	○	○	○	–
Amplitude	Gaussian Noise	◐	○	◐	Gaussian noise added to all values
	Spike Noise	◐	○	◐	Absolute of Gaussian noise added to at most 3 non zero values
	Gaussian WrapUp	◐	○	◐	Gaussian signal multiplied to all values
	Sine WrapUp	◐	○	◐	Sinusoidal signal multiplied to all values
	Constant WrapUp	○	○	●	Single uniformly sampled value multiplied to all IAT values only
Mask	Bernoulli Mask	●	●	●	Zeroing values with predefined probability
	Window Mask	●	●	●	Zeroing a segment of size 2 at random starting point
Order	Interpolation	●	●	●	Add avg. values in-between each pair, and select 20 consec. values from random start
	Flip	●	●	●	Swap values left to right
	Packet Loss	●	●	●	Discard packets arrived in an interval $\triangle t$ and add zero padding (if needed)
	Translation	●	●	●	Select a random starting point and shift values left or right
	Wrap	●	●	●	Interpolate or discard packets according to a randomized indicator
	Permutation	●	●	●	Split the time series into segments of random sizes and shuffle them
	CutMix	●	●	●	Create new samples by patching segments from 2 different samples

Segment: a consecutive group of timestamps. A feature is altered with probability $p = 1$ ●, $p = 0.5$ ◐, or $p = 0$ ○

Figure 1: Critical distance chart on `MIRAGE19` (No Aug refers to weighted F1 score = 75.21%. Difference of augmentation's metric w.r.t. baseline in brackets).

class-weighted sampler which creates batches by selecting minority class samples more frequently.

Figure 1 reports the Critical Difference (CD) chart of the weighted-F1 score, across 30 seeds, for each augmentation. To create such summary, first the performance of a given seed are ranked (the smaller, the better); then the ranks across seeds are analyzed via a Friedman test with Nemenyi post-hoc test which assesses their overlap and connects with an horizontal line augmentations that are not statistically different. First of all, notice how using the weighted samplers alone slightly hurts performance (No Aug −0.4% w.r.t. No Aug w/o Sampler). This is because the accuracy for majority classes reduces when sampling more frequently minority classes—showing more often the same minority samples does not help the learning. Conversely, combining this sampling with DA yields sizable improvements (up to +4.76% w.r.t. No Aug). In fact, despite the higher attention toward smaller classes, we also observe relative improvement in performance for larger classes (results not reported for lack of space). This hints that the samples variety added by some DAs indeed helps models to learn better data representations.

3 GENERATIVE DATA AUGMENTATION

Studies like the one in Fig. 1 are empirical and costly.[4] More importantly, implications on performance of DA methods are hard to predict. However, we argue that these campaigns are instrumental to charter the road toward *automatically learning augmentations*. In particular, we identify three stages.

Latent space geometry. First of all, it is reasonable to assume that the augmentations performance gaps is rooted in the geometry of the latent space. In fact, good DAs encourage the learning of more general and robust features, resulting in better classes separation. This raises questions such as *"Where would be more effective to project synthetic points? What level of samples variety is most effective for training?"* which we will address by using clustering metrics and latent space geometry analysis—we aim to uncover how augmentations can help to "regularize" the latent space.

Generative models. Second, better DA should be viable via generative models such as Generative Adversarial Networks (GAN) and Diffusion Models (DM). These techniques approximate the input data distribution, generate diverse samples, and can be guided by conditional mechanisms to steer their projection in the latent space of a classifier. The high realism and diversity of the obtained synthetic samples have motivated their use for augmenting training datasets for classification tasks [3].

However, generative models are usually trained separately from the final downstream task and with datasets having a large variety of samples. Unfortunately, state of the art datasets in TC offer only a modest variety compared to CV datasets.[5] Linking back to the previous stage, we envision a first exploration based on *conditioning* the generative models on the latent space properties learned via hand-crafted DA. Then, we will target the more challenging scenario of training *unconditionally* using datasets enlarged with hand-crafted DA and verify if effective regularizations are automatically learned.

Training pipeline. Last, we expect generative models based on pre-training to be sub-optimal in TC due to lower variety in the data. To link generative models to classification needs we will consider also an end-to-end training pipeline where both classifier and generative model are learned jointly. This calls for *self-supervision* mechanisms which have already been reported as useful is recent TC studies [4].

4 CONCLUSION

In this paper we presented a preliminary study supporting the use of DA and outlined a research agenda for adopting generative models in TC. We also raised awareness toward a more careful investigation of dataset imbalance. We believe that tackling the highlighted challenges will bring meaningful insights to TC.

[4] The search space expands when "stacking" DAs as done in CV.

[5] CESNET-TLS22 has 191 services across 141 million flows; LAION5B has 5 billion image-text pairs.

REFERENCES

[1] Ola Salman, Imad H. Elhajj, Ayman Kayssi, and Ali Chehab. A review on machine learning–based approaches for internet traffic classification. In *Annals of Telecommunications*, 2020.

[2] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In *IJCAI*, 2021.

[3] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2023.

[4] Md. Shamim Towhid and Nashid Shahriar. Encrypted network traffic classification using self-supervised learning. In *NetSoft*, 2022.