KNOWLEDGE DISTILLATION BASED ON MONOCLASS TEACHERS FOR EDGE INFRASTRUCTURE

Cédric Maron[1,2], Virginie Fresse[1], Karynn Morand[2] and Freddy Havart[2]

[1]Laboratoire Hubert Curien, 18 rue Professeur Benoît Lauras Bâtiment F,
42000 Saint-Etienne, France
[2]SEGULA Technologie, 1 Rue des Combats du 24 Août 1944, 69200
Vénissieux, France

ABSTRACT

With the growing interest in neural network compression, several methods aiming to improve the networks accuracy have emerged. Data augmentation aims to enhance model robustness and generalization by increasing the diversity of the training dataset. Knowledge distillation, aims to transfer knowledge from a teacher network to a student network. Knowledge distillation is generally carried out using high-end GPUs because teacher network architectures are often too heavy to be implemented on the small resources present in the Edge. This paper proposes a new distillation method adapted to an edge computing infrastructure. By employing multiple monoclass teachers of small sizes, the proposed distillation method becomes applicable even within the constrained computing resources of the edge. The proposed method is evaluated with classical knowledge distillation based on bigger teacher network, using different data augmentation methods and using different amount of training data.

KEYWORDS

Neural network compression, knowledge distillation, edge infrastructure, data augmentation

1. INTRODUCTION

The accuracy of a neural network has been widely recognized to be influenced by its size and architectural complexity. Larger neural networks have the capacity to model intricate relationships between input data and desired outputs, thereby improving accuracy compared to smaller neural networks.

In the context of Edge Computing, computational resources are commonly multiple but constrained in terms of memory and computing power. Consequently, the choice of neural network architecture must be carefully considered based on the available resources within this infrastructure. Having to choose smaller neural network architecture typically lead to scarifying accuracy. Several technics can be used to improve accuracy of small networks and not suffer to much from the resources constraints of the Edge. Data augmentation consist in enhancing the

David C. Wyld et al. (Eds): CNSA, DMDB, NIAI, ICCIoT, AIAP -2024
pp. 113-128, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140107

model robustness and generalization by increasing the diversity of the training dataset. It typically consists in applying various transformation to a given dataset to generate new data.

One notable technique is knowledge distillation, which facilitates the transfer of knowledge from a teacher network to a student network, thereby enhancing the precision of the student network compared to training without distillation.

In a classical distillation context, the architecture of the student network is chosen according to material constraints such as computing power and available memory. The teaching network is chosen solely for the purpose of obtaining the best possible precision. In general, the higher the precision of the teacher network is, the more effective the distillation of knowledge will be on the student network, if the student network has an architectural capacity to imitate the teacher network [1].

The teacher network often requires too much computing power and memory to run on the resources present in the Edge. Thus, to carry out the distillation, it is generally necessary to use high-end GPUs which are typically found in Cloud Computing.

The proposed method in this article is based on the use of multiple monoclass teacher networks with architectures of small sizes (equivalent to the student network size). This allows to perform the entire distillation training in an Edge infrastructure composed of low/medium computational capacities.
The contribution lies in the field of computer vision using Convolutional Neural Networks (CNNs), images dataset and image data augmentation methods.

The proposed method has been evaluated using the three data augmentation, random crop, random horizontal flip and mixup. The experiments are based on the CIFAR10 dataset using different amount of training data and using a customized LeNet architecture. Different combinations of the three data augmentation methods have been evaluated during the distillation training using different number of training data. The best results are obtained when using only random crop and random horizontal flip techniques. When using the proposed method in combination with random crop and random horizontal flip, a consistent gain in accuracy is observed over a regular training using only data augmentation. The proposed method gives also overall better results than the regular distillation based on a large teacher with the same data augmentation methods.

2. RELATED WORK

This section presents data augmentation techniques used in the field of computer vision as well as the different knowledge distillation methods.

2.1. Data Augmentation

Data augmentation is a crucial element of computer vision tasks, aiming to enhance model robustness and generalization by increasing the diversity of the training dataset. Early computer vision research introduced foundational data augmentation methods such as noise, rotation, scaling, and flipping. These techniques have become the foundation for image data augmentation. Geometric transformations have been extended to include cropping, translation, and perspective transformations. Random cropping is used to increase the invariance of Convolutional Neural Networks (CNNs) to object location. The work of Wang et al. [2] explored the application of perspective transformations to augment training data for object detection. Takashi et al. [3]

proposed a method combining random cropping (RICAP) which randomly crops four images and patches them to create a new training image. Zhang et al. [4] proposed mixup a method that trains a neural network on convex combinations of pairs of examples and their labels.

Recent advancements have introduced advanced augmentation techniques leveraging deep learning. Cubuk et al. [5] proposed AutoAugment, a method that employs reinforcement learning to discover optimal augmentation policies for image classification tasks. Similarly, latter they introduced RandAugment[6], a simple yet effective method for augmenting data with random transformations, demonstrating state-of-the-art results in various tasks.

Generative adversarial networks (GANs) have gained prominence in data augmentation. For instance, Perez and Wang. [7] employed GANs to generate realistic samples for object detection, addressing the challenges of limited training data. GAN-based approaches have shown promise in creating synthetic data that complements real-world datasets.

The data augmentation field has evolved from basic transformations to advanced methods, leveraging deep learning and generative models. As we advance in our understanding of data augmentation, addressing biases and overfitting while improving computational efficiency remains a primary research focus.

The Table 1 present a relative comparison of the different data augmentation using their computational complexity and their expected accuracy gain. A low computational complexity is labeled when the operations applied are on the order of few matrix operations. A High computational complexity is labeled when a high number of matrix operations is required to apply a data augmentation method such as GAN based data augmentation methods that use neural networks. The GANs data augmentations in some cases shows a significant gain in accuracy compared to the use of regular data augmentation, Frid-Adar et al[8]. demonstrated this technique on a liver lesion classification task and achieved a significant improvement of 7% using synthetic augmentation over the classic augmentation. Therefore, the GANs based methods have been labeled as Medium/High expected accuracy gain and the traditional method such as Random crop and random horizontal flip have been label as low expected accuracy gain. The mixup method shows a 1-2% gain in accuracy[9] when the baseline accuracy is around 95-96%. This can be considered as medium gain. The AutoAugment[5] method shows a 3% gain in accuracy when the baseline is also around 93-94%. This can be considered as medium/high gain. Finally, the RandAugment[6] method shows a 1% gain in accuracy when the baseline accuracy is around 97% which can also be considered as Medium/High gain. In this paper, the data augmentation methods evaluated are random crop, random horizontal flip and mixup. These methods are chosen for their low computational complexity even if their expected effect on the accuracy is lower than more advanced data augmentation methods.

Table 1. Relative comparison of computational complexity vs expected accuracy gain for different data augmentation methods.

Data augmentation method	**Method computational complexity**	**Expected accuracy gain**
Random crop	Low	Low
Random horizontal flip	Low	Low
Mixup	Low	Medium
GANs based	High	Medium/high
AutoAugment	Medium	Medium/high
RandAugment	Medium	Medium/high

2.2. Distillation

Knowledge distillation is a technique that aims to transfer knowledge from a large, complex model (teacher) to a smaller, computationally efficient model (student). There are several knowledge distillation methods[10]. The distillation of knowledge based on the logits (un-normalized predictions) of outputs consists in training a student network with a restricted architecture to generate logits similar to a more complex teacher model and having a high accuracy. This method is often used in the context of image classification.

The distillation of knowledge based on feature maps generated by intermediate layers of a network consists in training a student network to extract feature maps similar to a teacher model. This method is often used in the context of image segmentation.

Gradient-based knowledge distillation is a method to improve the robustness of the two previous methods. The gradients generated during the training of a network make it possible to know which parts of the network are the most active. Therefore, it is possible to use this information to ensure that a student network can replicate the operation of the most active parts of the teacher network.

There are multiple distillation schemes, the three main ones are:

- Offline distillation consists of distilling knowledge from a pre-trained teacher network to a student network.
- Online distillation consists of jointly training a student network and a teacher network while carrying out the distillation in parallel.
- Self-distillation consists of carrying out distillation between the intermediate layers of the same network.

You et al. [11] proposed to use multiple teacher network to perform knowledge distillation. They showed that their method is capable of generating a well-performed student network.

The offline distillation methods presented in the literature focus mainly on using teacher networks of larger sizes than the student network as shown in Table 2. The proposed method uses teachers and student network architectures of similar sizes.

Table 2. Comparison of teacher/student parameters for different distillation techniques.

Method	**Teacher**	**Student**	**Ratio params T/S**
CTKD[12]	WRN-40-1 WRN-40-2	WRN-16-1 WRN-16-2	3.29 3.19
TOFD[13]	ResNet152 ResNet152	ResNeXt50-4 MobileNetV2	2.40 17.17
AdaIN[14]	ResNet26 WRN-40-2	ResNet8 WRN-16-2	4.63 3.19
FN[15]	ResNet110 ResNet56	ResNet56 ResNet20	2.0 3.15
Proposed method	LeNet (customized)	LeNet (customized)	1.0

2.3. Distillation and Data Augmentation

New studies are done to study the impact of data augmentation when using knowledge distillation. Das et al. [16] did an empirical analysis of the impact of data augmentation on knowledge distillation and proposed a class-discrimination metric to quantitatively measure the performances of different data augmentation methods. Wang et al. [17] try to respond to the question. What makes a “good” data augmentation in knowledge distillation? They suggest that a good DA scheme should reduce the covariance of the teacher-student cross-entropy. They presented a practical metric, the stddev of teacher’s mean probability (T. stddev).

2.4. Contribution

While knowledge distillation has been extensively explored with a single teacher, applying this approach with multiple monoclass teachers of equivalent sizes to the student is a novel direction. This paper aims to investigate the efficacy of offline-logits based distillation using multiple monoclass teachers. The choice of proposing an offline-logits based distillation is made because it allows easier implementation and less dependence to network architecture. The hypothesize is that this approach could lead to more accurate and robust student models for multiclass classification tasks while being applicable using low/medium computational resources. The analysis of the proposed method mixed with data augmentation methods provides good insight on the potential performance gains that can provide this new approach.

3. Proposed Method

In this part, different aspects of the proposed distillation method are presented starting from the motivation and contextualization to the implementation and finishing by the experimental objectives.

3.1. Motivation and Contextualization

The proposed method is a multi-monoclass-teacher offline distillation method based only on the output logits. The typical multiclass distillation paradigm consists in doing the distillation training using high-ends GPUs, typically in the Cloud using a large pre-trained multiclass teacher and then transfer the trained student model in the edge. In this paradigm the Cloud is considered to be a distant server which has large computing and storage resources. The proposed distillation method presented in this paper consider a different paradigm. The paradigm considered consists in realising the distillation training directly in the Edge infrastructure by using multiple pre-trained monoclass teachers of sizes similar to the student network. In this paradigm the Cloud is mainly used as external storage space, therefore large computing resources can be available, but they are not needed. In both paradigms the edge computing resources are considered to be low to medium computational resources, typicaly not high end GPU such as Nvidia A100. Finally, in both paradigms the connection between Cloud and Edge infrastructure can be considered either stable or fluctuating. The
Figure 1 depict both paradigms.

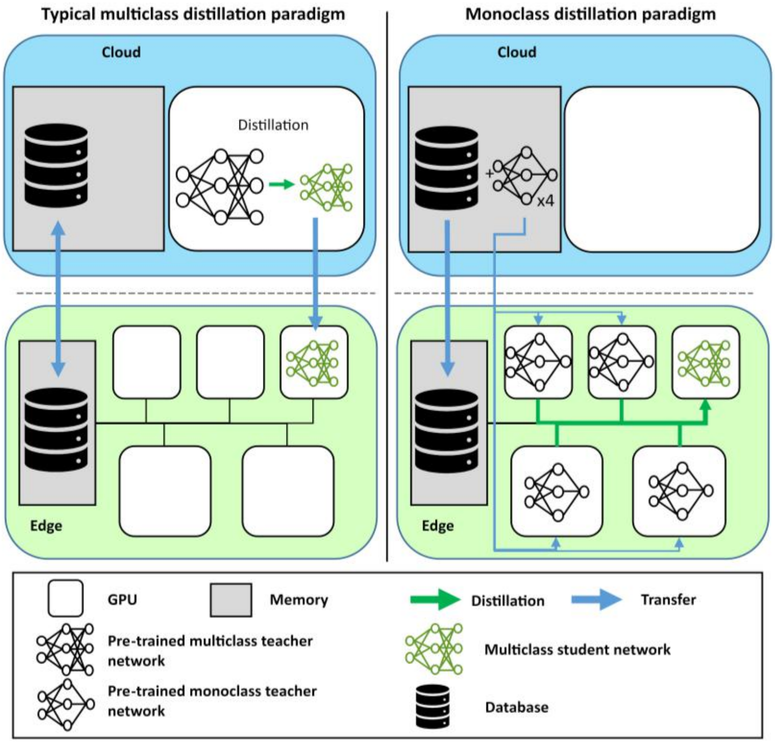
Typical multiclass distillation paradigm
Monoclass distillation paradigm
Cloud
Distillation
Cloud
x4
Edge
Edge
GPU
Memory
Distillation
Transfer
Pre-trained multiclass teacher network
Multiclass student network
Pre-trained monoclass teacher network
Database

Figure 1. Approach contextualization. On the left, typical multiclass distillation paradigm using a large multiclass teacher network to train a student network and then deploy the trained student in the edge. On the right, the proposed monoclass distillation paradigm that enable the distillation directly on the edge by using multiple monoclass teacher network of sizes equivalent to the student network.

3.2. Model Architecture

The architecture of the student network and the monoclass teacher networks are considered identical except for the output of the monoclass teacher networks, which has 2 outputs instead of n. The 2 classes output correspond either to the class to be predicted or to "other".

3.3. Training Process

The labels of the database need adapted for each monoclass teacher network to replace all the non-main class labels by “other”. The database used for teacher networks and student network training are considered the same. The student can also be trained only from a subset of the same dataset. The training of the teacher network can be realised either in the Cloud or in the Edge depending on personal needs. The multiple monoclass teachers are considered distributed on multiple resources to perform the training in parallel, therefore leveraging the global computational capacity present in the Edge. The student network is then trained using both the ground truth of the database and the aggregated outputs of the different monoclass teacher networks by following the aggregation method proposed.

3.4. Aggregation Method

The proposed aggregation method proposed in
Figure 2 to aggregates the main class logits of the n teacher networks to recreate a n logits vector corresponding to each class of the database. The aggregated vector is used to distill the knowledge acquired by the teacher networks by calculating the loss (cost function) of distillation

between the logits of the teachers and the logits of the student. The distillation loss used is MSE (Mean Squared Error) and the ground truth loss is Cross Entropy.

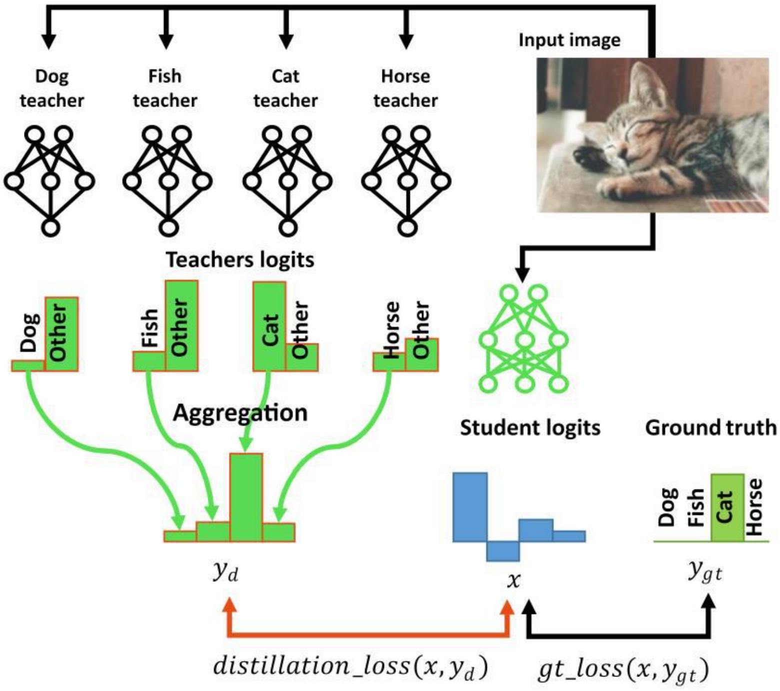

Figure 2. Main class logits aggregation using the monoclass teacher outputs followed by the calculation of the distillation loss in parallel with the calculation of the ground truth loss

3.5. Experimental Objectives

The objective of the experiments is to determine if the idea of combining multiple monoclass teacher networks can effectively improve the learning of a student network while having student and teacher network architectures of restricted and similar sizes.

4. Experimentations and Results

The proposed distillation method has been rigorously evaluated using a combination of data augmentation techniques, namely random crop, random horizontal flip, and mixup, with varying amount of training data. This evaluation process is essential to validate the effectiveness of the distillation approach in the context of enhanced training data with different numbers of data. By incorporating these augmentation methods into the evaluation, the model's performance can be thoroughly tested under a variety of conditions, including different image perspectives, orientations, and blended samples. This comprehensive evaluation not only ensures that the distillation process can effectively transfer knowledge from the teacher models to the student model but also verifies that the benefits of data augmentation, such as improved generalization and robustness, are retained throughout the training process. The results of this evaluation demonstrate the practical utility of the proposed distillation method in enhancing the performance of machine learning models.

First, the experimentation configuration is described. A baseline with data augmentation and without distillation is evaluated. The proposed distillation method is evaluated using different combination of data augmentation with various number of training data and compared to the baseline and a classical multiclass based distillation. Finally, the experiment results are analyzed.

4.1. Experiments Configuration

The training of the multiclass/monoclass teacher networks and the student network are carried out with a batch size of 96 images out of 100 epochs with the CIFAR10 database which includes images belonging to 10 different classes divided into 50,000 images of training and 10,000 test images.

During the experiment, the teacher networks are first trained with all the training and using the three data augmentation methods, random crop, random horizontal flip and mixup. The teachers are then distilled on the student network using different combination of the three data augmentation methods. The Adam optimizer is used with a learning rate of 0.001. The student trainings are performed using between 1 and 5000 images per class (which is the maximum number of training data). For each data point presented in the experiment, the trainings are performed three times to get an average of the max Top-1 test accuracy obtained for different number of images per class. The Top-1 accuracy is the standard measure of accuracy, requires the model's highest-probability prediction to precisely match the expected answer. The tests are performed on a computer with an i7-9700 CPU, 32GB of RAM and an Nvidia Quadro P5000 16GB graphics card.

The architectures used during the tests are LeNet type custom architectures. The architecture of the student network (51,880 parameters) and the single-class teacher networks (51,752 parameters) are identical except for the output of the monoclass teacher networks, which has 2 outputs instead of 10. The 2 classes output correspond either to the class to be predicted or to "other". The architecture of the multiclass network has the same depth as the student network but is however much wider (4,187,018 parameters). The Table 3 describe the multiclass teacher architecture. The Table 4 describe the student architecture.

Table 3. Multiclass teacher architecture

Layer	Feature Map	Size	Kernel Size
Input image	3	32x32	-
Convolution	32	32x32	3x3
ReLU	32	32x32	-
MaxPooling	32	16x16	2x2
Batchnorm	32	16x16	-
Convolution	128	16x16	3x3
ReLU	128	16x16	-
MaxPooling	128	8x8	2x2
Batchnorm	128	8x8	-
Fully connected	-	500	-
ReLU	-	500	-
Fully connected	-	100	-
ReLU	-	100	-
Fully connected	-	10	-

Table 4. Student architecture

Layer	Feature Map	Size	Kernel Size
Input image	3	32x32	-
Convolution	12	32x32	3x3
ReLU	12	32x32	-
MaxPooling	12	16x16	2x2
Batchnorm	12	16x16	-
Convolution	25	16x16	3x3
ReLU	25	16x16	-
MaxPooling	25	8x8	2x2
Batchnorm	25	8x8	-
Fully connected	-	30	-
ReLU	-	30	-
Fully connected	-	15	-
ReLU	-	15	-
Fully connected	-	10	-

4.2. Baseline without Distillation

First, the baseline of the student network trained using different data augmentation methods without distillation is evaluated by training it using all the different combination of data augmentation methods between random crop, random horizontal flip and mixup.

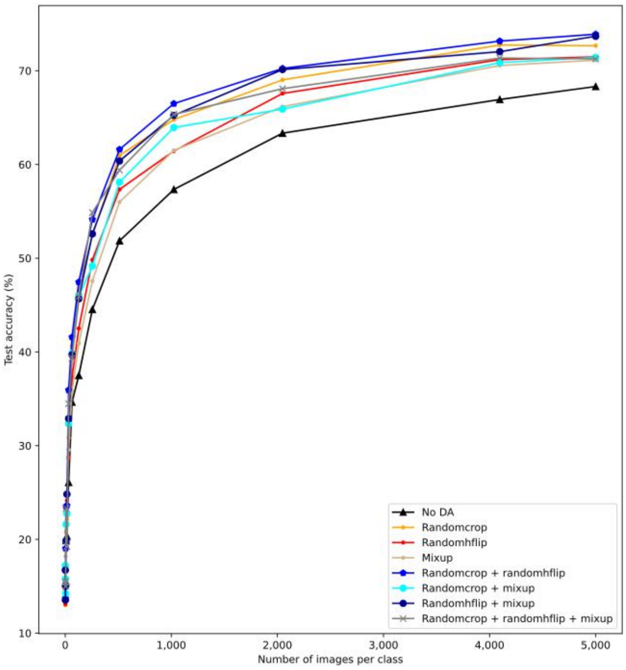

Figure 3. Max test accuracy obtained by the student network for different data augmentation combination and different number of images per class used during training without distillation

The

Figure 3 shows that the combination, random crop + random horizontal flip lead to the overall best accuracy for different number of training data. This baseline without distillation is used in the next evaluations.

4.3. Distillation without Data Augmentation

The proposed distillation method is evaluated without the use of data augmentation. The results are depicted in the

Figure 4. The distillation using monoclass teachers or one large multiclass teacher alone gives lower accuracy than a training using data augmentation.

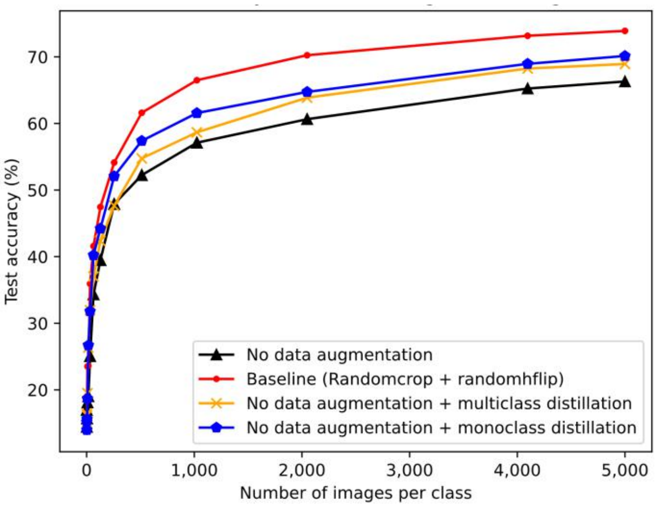

Figure 4. Max test accuracy obtained by the student network without data augmentation for the monoclass distillation and multiclass distillation.

4.4. Distillation Using Data Augmentation

The proposed distillation method is evaluated firstly using one data augmentation method at a time. The results for the random crop, random horizontal flip and mixup are respectively depicted in

Figure 5,

Figure 6 and

Figure 7.

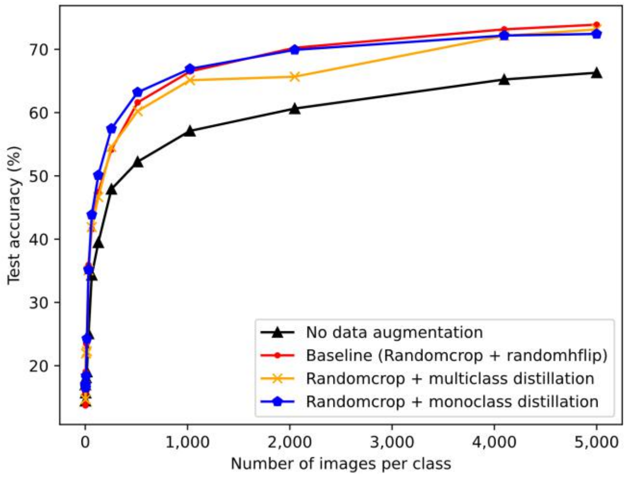

Figure 5. Max test accuracy obtained by the student network using the random crop data augmentation for the monoclass distillation and multiclass distillation.

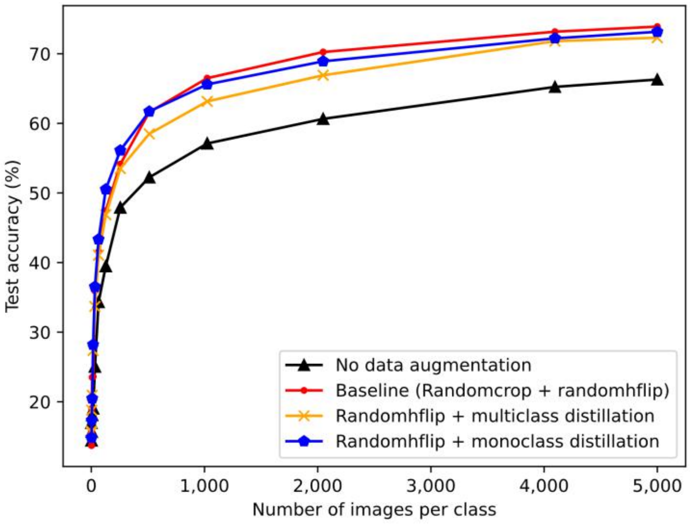

Figure 6. Max test accuracy obtained by the student network using the random horizontal flip data augmentation for the monoclass distillation and multiclass distillation.

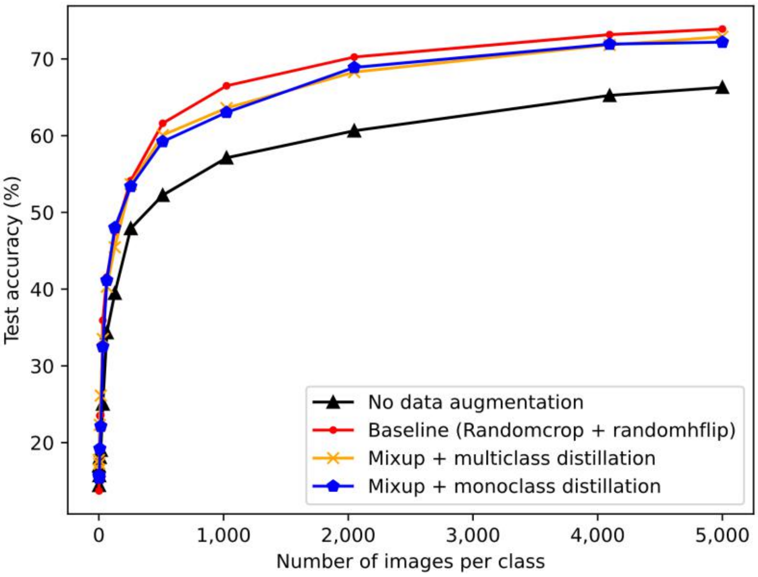

Figure 7. Max test accuracy obtained by the student network using the mixup data augmentation for the monoclass distillation and multiclass distillation.

From these three data augmentation methods the one that seems to give the best results when used alone is random crop then random horizontal flip and finally mixup.

Then the different combinations of data augmentations have been evaluated. The Figure 8 depict the results obtained using random crop + random horizontal flip and distillation. The Figure 9 depict the results obtained using random crop + mixup and distillation. The Figure 10 depict the results obtained using random horizontal flip + mixup and distillation. Finally, the Figure 11 depict the results obtained using random crop + random horizontal flip + mixup and distillation.

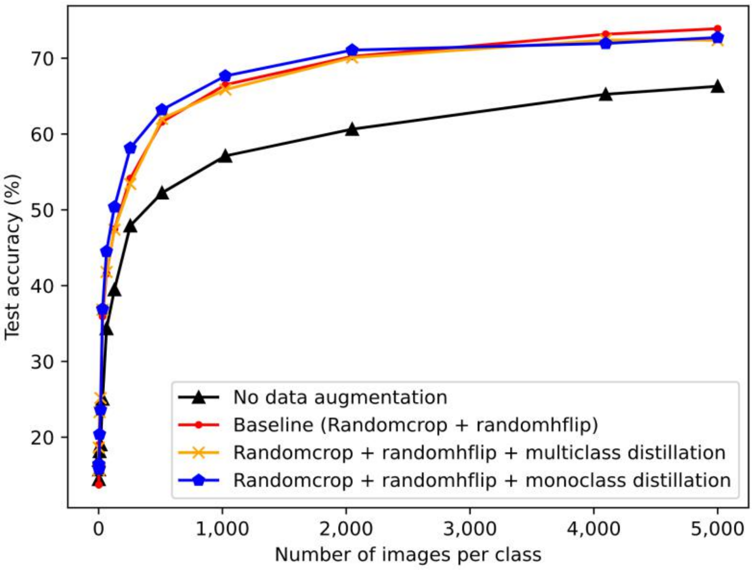

Figure 8. Max test accuracy obtained by the student network using the random crop + random horizontal flip data augmentations for the monoclass distillation and multiclass distillations.

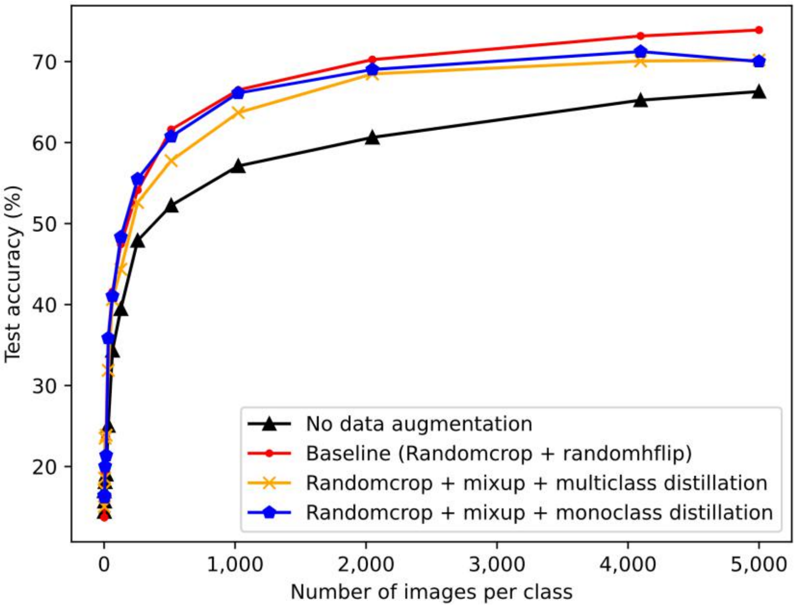

Figure 9. Max test accuracy obtained by the student network using the random crop + mixup data augmentations for the monoclass distillation and multiclass distillations.

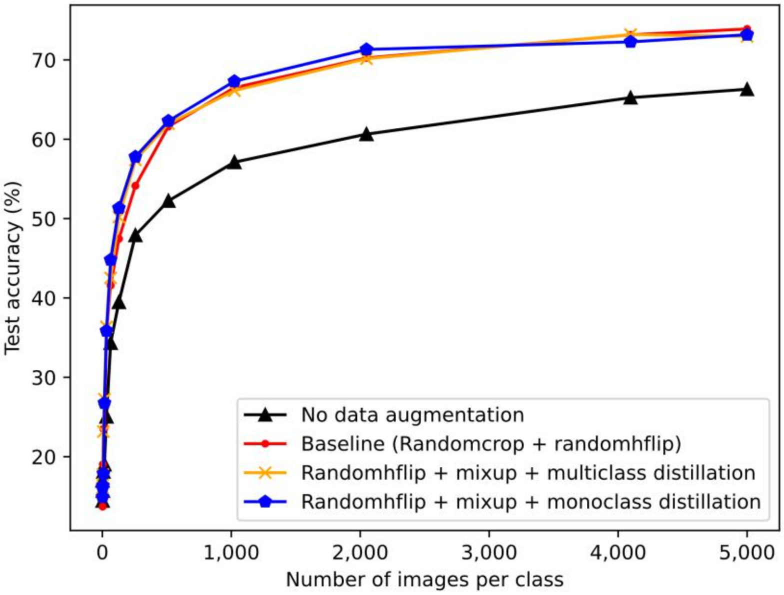

Figure 10. Max test accuracy obtained by the student network using the random horizontal flip + mixup data augmentations for the monoclass distillation and multiclass distillations.

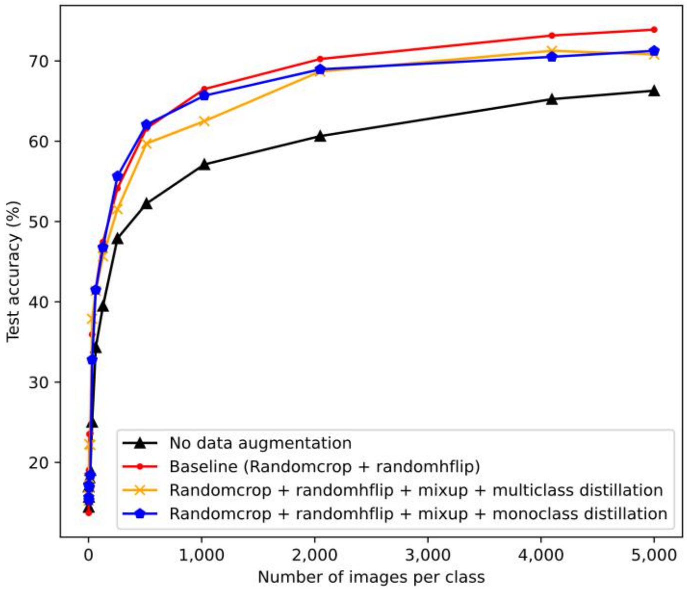

Figure 11. Max test accuracy obtained by the student network using the random crop + random horizontal flip + mixup data augmentations for the monoclass distillation and multiclass distillations.

The results show that the use of random crop and random horizontal flip during the training is the combination that gives the overall best results (blue curve over red curve). The use of other combinations of data augmentation lead to overall worst result than the baseline (blue curve under red curve). This means that the use these combinations of data augmentation associated with distillation does not bring any gain compared to a training performed only with data augmentation.

4.5. Results Analysis

The choice of the data augmentation methods used during the training can impact significatively the max test accuracy obtained during the training. The combination of random crop and random horizontal flip is the one that lead to the best results. The use of the others combination of data augmentation seems to decrease the accuracy gain compared to training performed only with data augmentation.

The accuracy gain provided by the proposed method is higher when a lower number of data is used for training. This could be explained by the fact that higher number of data gives more information to the student network which reduce the benefits of the added knowledge given by the distillation. The proposed method seems to provide a more consistent accuracy gain compared to regular knowledge distillation based on multiclass teacher.

The computational complexity of the different networks used in the experiments can be expressed by the number of multiplications and additions (multi-adds) required to perform the inference as well as its memory usage. The student network inference requires 1071631 multi-adds and 233472 bytes of memory. One monoclass teacher inference requires 1071503 multi-adds and 232960 byte of memory. The multiclass teacher inference requires 14502298 multi-adds and 17165824 bytes of memory. The monoclass teachers require 13.53 times less multi-adds operations than the multiclass teacher and 73.69 times less memory during inference.

5. CONCLUSIONS

This article present an offline-logits based distillation method that uses multiple monoclass teachers of sizes equivalent to the student network. The key advantage of the proposed method is its ability to perform offline knowledge distillation on multiple small computational resources which suit the Edge context very well. It eliminates the need for high-performance external computing resources to train or infer on a large multiclass teacher. By employing the proposed method, edge infrastructures can benefit from enhanced network precision without relying on resource-intensive cloud-based computations.

Through testing on the CIFAR10 database, the method presented outperforms regular knowledge distillation based on multiclass teacher. The method also gives more consistent accuracy gain when considering different number of training data and different data augmentation methods.

Moving forward, the future perspectives revolve around reducing the database requirements for training monoclass networks. This is crucial to accommodate the limited storage capacities typically found in Edge infrastructure.

ACKNOWLEDGEMENTS

This work is funded by the Association Nationale de la Recherche et de la Technologie (ANRT) under the industrial agreement (CIFRE contract) number 2022/0190.

REFERENCES

[1] J. H. Cho and B. Hariharan, "On the Efficacy of Knowledge Distillation," 2019, pp. 4794–4802. Accessed: Jun. 22, 2023. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Cho_On_the_Efficacy_of_Knowledge_Distillation_ICCV_2019_paper.html

[2] "Perspective Transformation Data Augmentation for Object Detection." https://ieeexplore.ieee.org/abstract/document/8943416/ (accessed Nov. 09, 2023).

[3] "Data Augmentation Using Random Image Cropping and Patching for Deep CNNs." https://ieeexplore.ieee.org/abstract/document/8795523/ (accessed Nov. 09, 2023).

[4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," ArXiv171009412 Cs Stat, Apr. 2018, Accessed: Nov. 09, 2023. [Online]. Available: http://arxiv.org/abs/1710.09412

[5] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Policies from Data," ArXiv180509501 Cs Stat, Apr. 2019, Accessed: Nov. 09, 2023. [Online]. Available: http://arxiv.org/abs/1805.09501

[6] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical Automated Data Augmentation With a Reduced Search Space," 2020, pp. 702–703. Accessed: Nov. 20, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w40/Cubuk_Randaugment_Practical_Automated_Data_Augmentation_With_a_Reduced_Search_Space_CVPRW_2020_paper.html

[7] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," ArXiv Prepr. ArXiv171204621, 2017.

[8] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification.” 2018.

[9] Z. Liu et al., “AutoMix: Unveiling the Power of Mixup for Stronger Classifiers.” 2022.

[10] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.

[11] S. You, C. Xu, C. Xu, and D. Tao, “Learning from Multiple Teacher Networks,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, Aug. 2017, pp. 1285–1294. doi: 10.1145/3097983.3098135.

[12] H. Zhao, X. Sun, J. Dong, C. Chen, and Z. Dong, “Highlight Every Step: Knowledge Distillation via Collaborative Teaching,” ArXiv190709643 Cs, Jul. 2019, Accessed: Jul. 12, 2023. [Online]. Available: http://arxiv.org/abs/1907.09643

[13] L. Zhang, Y. Shi, Z. Shi, K. Ma, and C. Bao, “Task-Oriented Feature Distillation,” in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 14759–14771. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/a96b65a721e561e1e3de768ac819ffbb-Paper.pdf

[14] J. Yang, B. Martinez, A. Bulat, and G. Tzimiropoulos, “Knowledge distillation via adaptive instance normalization,” ArXiv200304289 Cs, Mar. 2020, Accessed: Jul. 12, 2023. [Online]. Available: http://arxiv.org/abs/2003.04289

[15] K. Xu, L. Rui, Y. Li, and L. Gu, “Feature Normalized Knowledge Distillation for Image Classification,” in Computer Vision – ECCV 2020, Cham, 2020, pp. 664–680. doi: 10.1007/978-3-030-58595-2_40.

[16] D. Das, H. Massa, A. Kulkarni, and T. Rekatsinas, “An Empirical Analysis of the Impact of Data Augmentation on Knowledge Distillation,” ArXiv200603810 Cs, Jun. 2020, Accessed: Oct. 31, 2023. [Online]. Available: http://arxiv.org/abs/2006.03810

[17] H. Wang, S. Lohit, M. N. Jones, and Y. Fu, “What Makes a ‘Good’ Data Augmentation in Knowledge Distillation - A Statistical Perspective,” Adv. Neural Inf. Process. Syst., vol. 35, pp. 13456–13469, Dec. 2022.

AUTHORS

Cédric Maron graduated with an electrical engineering degree in 2021 from Polytech Clermont-Ferrand, France. Since 2022 he is a Ph.D. student in Computer science working at the Hubert-Curien Laboratory in Saint-Etienne, France in collaboration with Segula Technologies. His research interests are artificial intelligence and computer vision.

Virginie Fresse is an associate professor in the Jean Monnet University, in Saint Etienne, France. She got her PhD degree in Electrical Engineering in INSA Rennes in 2001 and got a post-doctorate position in the University of Strathclyde, in Glasgow from 2001-2003. Her reserearch projects are on embedding image processing algorithms on embedded systems containing FPGA, DSP or embedded CPU devices put on cloud and edge infrastructure. The integration of CNN models for videos and images is also an actual research project with an industrial partner.

Karynn Morand is currently a Research and Development Manager in SEGULA Technologies, Lyon. She graduated with a mechanical and electrical engineering degree in 2001 from ECAM Lyon, France. Her area of work includes artificial intelligence, computer vision and materials.

Freddy Havart graduated with an Energy and Propulsion engineering degree in 1992 from INSA of Rouen, France. Since 2018, he works for the Research and Development department of Segula Technologies. He started out working on mechanical engineering and numerical simulation projects. Now, his area of work also includes artificial intelligence, computer vision and embedded systems.