Exploring Automated Data Augmentation Approaches for Deep Learning: A Case Study of Individual Feral Cat Classification

Zihan Yang
The University of Melbourne
zihany1@student.unimelb.edu.au

James Bailey
The University of Melbourne
baileyj@unimelb.edu.au

Richard O. Sinnott
The University of Melbourne
rsinnott@unimelb.edu.au

Krista A. Ehinger
The University of Melbourne
kris.ehinger@unimelb.edu.au

Abstract

This paper evaluates the performance of several automated data augmentation (AutoDA) methods for image classification problems suited for scenarios with limited and potentially imbalanced data sets. We compare one-stage, two-stage and search-free methods. These are explored in the context of a case study to identify/count feral cats in rural Victoria. Our results show that a trade-off exists between accuracy and efficiency, with one-stage methods being faster but less accurate than two-stage methods. Search-free methods are fastest, but have limited improvement in the resultant classification accuracy.

Keywords: AutoDA, deep learning, computer vision, data augmentation, hyperparameter tuning

1. Introduction

Nowadays, data augmentation (DA) is regarded as an essential step for training effective deep learning systems. DA is used to increase the size and diversity of a dataset by artificially generating new data samples based on existing data. This is particularly useful in scenarios where the available training data is limited or imbalanced. In the computer vision domain, a variety of basic image transformation functions, such as rotation, scaling and colour adjustment, are commonly used for image data augmentation. DA is widely used in many state-of-the-art models, including data-driven models that typically rely on a large amount of labelled data, and few or zero-shot learning models that require data diversity.

However, despite the ubiquity and importance of image data augmentation in computer vision tasks, there has been less focus on establishing the most suitable DA strategies for specific tasks. One reason for the current research gap is that the search space for data augmentation policies can be very large, making it difficult to identify the best policy (Cubuk et al., 2019). The effectiveness of augmentation policies can be highly dependent on the specific task and dataset, making it challenging to generalize data augmentation policies across tasks (Krizhevsky et al., 2012). The selection of DA strategies still heavily depends on human expertise, which can be subjective and error-prone.

In recent years, there has been a growing interest in the field of automated machine learning (AutoML), which has stimulated the development of a number of innovative techniques designed to improve deep learning systems. One of these techniques is automated data augmentation (AutoDA), which has gained significant attention (Yang et al., 2023). Unlike traditional image augmentation that largely relies on the knowledge and experiences of human experts, AutoDA aims to automatically generate optimal DA policies based on potentially small amounts of input data to help improve the final performance of deep neural networks.

In this paper, we present a case study exploring existing state of the art AutoDA algorithms using a representative case study related to feral cats around Victoria. Individual feral cat identification is a particularly challenging task due to the insufficient

URI: https://hdl.handle.net/10125/106517
978-0-9981331-7-1
(CC BY-NC-ND 4.0)

and imbalanced data sets that are captured by motion cameras (Yang et al., 2021). The objective of this paper is to explore the effectiveness of AutoDA techniques in this context. The rest of the paper is structured as follows. In Section 2, we introduce related works. In Section 3 we introduce the data and associated methodology. In Section 4, we introduce the experiments and present the results. In Section 5 we discuss the results in more detail before concluding the work and identifying areas of future work in Section 6.

2. Related Works

Data augmentation (DA) is an essential technique to increase the size and diversity of a dataset, which is particularly useful when the available data is limited or imbalanced. The resulting augmented dataset can then be used to train more robust deep learning models that can generalize better to new data and new scenarios. Generally, larger and more diverse datasets achieve higher accuracy and better generalization performance (Sun et al., 2017). This is because a large training set provides a great number of data samples for models to learn from and ensures that models are exposed to a wider and more diverse range of data variations that may be encountered in the real world. Additionally, class imbalance problems can also be addressed by generating new data for minority classes. Thus, DA is typically critical in order to acquire sufficient and diverse datasets for deep learning.

As an oversampling method, the major objective of DA is to mitigate the negative effects of limited or imbalanced data by increasing the size of the training data set. The diversity of data is a crucial factor for model training. Merely duplicating existing data samples is unlikely to achieve a significant improvement in model performance (Deng et al., 2009). To achieve better generalization, it is necessary to introduce variations and complexities to the training data by using various image transformations. Examples of standard manipulations include rotations, translations, flips, and distortions.

One of the early examples of image DA can be traced back to the 1980s where it was applied to character recognition tasks (Rabiner and Juang, 1993). Geometric transformations were applied to handwritten characters to artificially create new examples. Another notable example of image DA can be found in object detection tasks, where researchers used affine transformations to simulate changes in illumination and viewpoint to enhance data diversity (LeCun et al., 1998). Since then, the concept of image data augmentation has been greatly extended and widely applied to various deep learning models, including AlexNet model, VGG, ResNet and Inception (Krizhevsky et al., 2012).

When it comes to applying DA in the field of computer vision, a variety of basic image transformations are available, such as scaling, translation, rotation, and more. However, selecting the most effective DA policies for a specific computer vision task or dataset can be challenging, as different tasks may necessitate application of distinct techniques due to the nature of the dataset. For instance, geometric transformations and colour adjustment are common DA strategies for general image classification tasks, such as CIFAR-10/100 and ImageNet where they are used to simulate different environmental scenarios. While for character or digit recognition datasets, rotation or flipping might not always be a reasonable choice as it may fail to preserve the original label of the data sample (Simard et al., 2003), e.g., a 9 can become a 6.

Hitherto, the selection of augmentation strategies often relies on intuition and prior knowledge of human experts Cubuk et al., 2019. However, human decisions can be influenced by biases and errors. Moreover, there is often no theoretical evidence to show that manually-decided DA policies can bring the best performance improvement for a given task and dataset. To address this issue, researchers have been exploring automated methods for finding the most effective DA policies. One such approach is automated data augmentation (AutoDA). AutoDA offers a technique that can automatically construct the best DA policies for a given task and dataset (Yang et al., 2023).

The first attempt to AutoDA was made by transformation adversarial networks for data augmentations (TANDA) (Ratner et al., 2017). This work inspired the design of later works such as AA in 2019 (Cubuk et al., 2019). Similar to TANDA, AA utilizes reinforcement learning to conduct the augmentation search. During the search, augmentation policies are sampled via a recurrent neural network (RNN) controller and these are used for model training. These policies are then fed into a simplified neural network for evaluation to select the best one so that the final classification model can yield the highest validation accuracy on the target task. The AA algorithm achieved 0.4% and 0.6% accuracy improvement on ImageNet and CIFAR-10 data respectively when compared to state-of-the-art at that time. More importantly, the policies acquired from ImageNet and CIFAR-10 could be transferred to effectively enhance model performance to various other datasets. Despite its success, AA often requires thousands of GPU hours to complete a given search, even under a reduced setting (Cubuk et al., 2019). Therefore, subsequent research has focused on improving the search efficiency of AutoDA models.

While AutoDA has been successfully applied to many standard image classification datasets, its effectiveness in real-world scenarios is not well defined. In this paper, we explore the use of various AutoDA algorithms on a customized dataset composed of images of feral cats with the goal of counting individual (unique) cats. Our aim is to investigate whether AutoDA can improve the performance of classifiction models on this type of dataset and if so, identify the most effective AutoDA algorithm for the given task. We compare the performance of models trained on the original dataset and on the augmented dataset generated by multiple AutoDA models. We then evaluate the impact of different DA policies on model accuracy. By conducting such a case study, we aim to provide insights into the applicability of AutoDA techniques to customized datasets and to demonstrate the potential of different AutoDA techniques for enhancing the generalization of deep learning models in challenging real-world applications.

3. Data and Methodology

3.1. Dataset

Feral cats are a menace to native species in Australia and ecologists need to know their abundance. The feral cat image data for individual cat identification was gathered from 938 trap camera sites located the Great Otway National Park and Otway Forest Park, in Victoria, Australia (38.42 °S, 142.24 °E). At each trap site, a sensing camera was installed with infrared flash and temperature-in-motion detector. This was triggered to capture five consecutive photographs when the camera detected the movement of nearby animals. The majority of cameras used were Reconyx Hyperfire HC600, while a small proportion consisted of PC900's HF2X's infrared camera.

Originally, manual data processing (labelling of unique cats) was done by at least two independent observers to ensure accuracy (Rees et al., 2019). This was based on comparison of the unique markings of feral cats. Each individual cat was assigned a unique identifier for later model training and identification. The data itself was often challenging to deal with in classifying the case as seen in the examples shown in Figure 1.

In order to ensure that the feral cats were the primary focus of the photographs, it was necessary to take into account the settings of the cameras. The position of all trap cameras were fixed during data collection. As a result of this positioning, in most images captured by the cameras, the feral cat was only a small part of the entire picture. This made it somewhat more difficult for the model to identify the cat in the image, as there may be other distinct objects or background elements that could potentially impact on the inference. That is, the presence of background elements in the image can be a source of confusion or noise for the model, making it harder for the model to isolate the cat compared to its surroundings. In addition, cameras typically use burst mode once triggered and take a series of photographs in a rapid succession. This feature results in a large number of images with similar backgrounds and of the same object (i.e. the same cat with minimal changes in its position in the image), which can further affect the model's ability to effectively learn and extract meaningful feature information. Figure 2 shows four photographs directly collected from the same trap camera in a short period of time. The captured images of feral cats tend to show the animal at a small scale, which makes a series of consecutive photographs appear highly similar due to the camera's fixed position and the cat's relatively stationary posture.

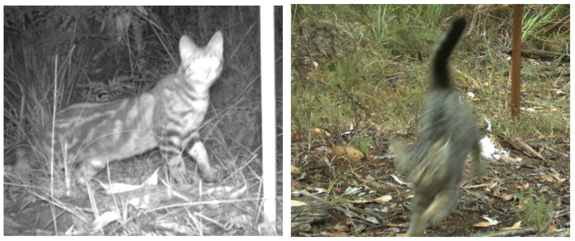

Figure 1. Examples of low quality images.

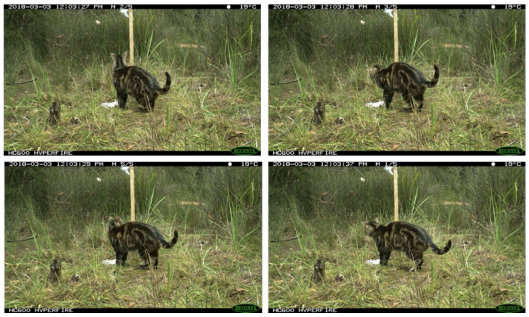

Figure 2. Example images with similar surroundings.

To reduce the potential impact of environmental surroundings and to better focus on the object of interest in each photograph, the raw data was pre-processed before applying the different AutoDA approaches. The observed object was centred in each photograph and cropped accordingly. All cropped images were then resized to a uniform size of 32 $\times$ 32 pixels using 0-padding to ensure consistency across the dataset. This

Table 1. Statistics of the Feral Cat Dataset.

Region	Class#	Image#
Annya	7	858
Cobboboonee	12	354
Hotspur	8	342
Mt Clay	9	462
Otways	87	10,644
Total	123	12,660

pre-processing step was necessary to ensure that the feral cats were the primary focus of the dataset and to improve the data quality.

Table 1 shows the statistics of the feral cat dataset used for testing the AutoDA models. The dataset consists of a total of $12,660$ images with a uniform resolution of 32×32 pixels (after pre-processing). There were 123 classes, i.e. individual / unique feral cats in the dataset, with each class containing different number of images. The class distribution was highly skewed as shown in Figure 3, with some classes having significantly more images than others. To address the class imbalance issue, we used stratified sampling to split the dataset into training and testing sets, ensuring that each class was represented proportionally in both sets. Specifically, we used a $5 : 1$ train-test split, resulting in $10,550$ images for training and $2,110$ images for testing.

3.2. Automated Data Augmentation

Image data augmentation has been shown to improve the performance of deep learning models, particularly when working with small datasets (Shorten and Khoshgoftaar, 2019). However, manual data augmentation can be time-consuming and labor-intensive. Automating this is desirable, but it may not always be possible to find the optimal augmentation strategy through trial and error (Cubuk et al., 2019), e.g., image rotations may have no effect on the model accuracy.

Automated Data Augmentation (AutoDA) is a technique that automates the process of finding the optimal data augmentation policies for a given dataset (Yang et al., 2023). In the context of image data augmentation, a DA policy refers to a collection of various image transformations, such as rotation, translation, flipping, cropping, zooming, contrast ratio manipulation, scaling and mosaicing. These operations are often applied randomly to each image in the training set during the training process, resulting in a larger and more diverse set of data. The primary objective of AutoDA is to maximise the effectiveness of the generated DA policies, in order to improve the performance of deep learning models trained with the augmented data. By automating the process of selecting the best DA policies, AutoDA eliminates the need for human expertise and reduces the trial-and-error process in finding the optimal DA policies. This results in significant savings in terms of time and resources required for manual design data augmentation. In this study, we apply a variety of state of the art AutoDA algorithms. The details of the AutoDA methods and their implementation are described in the following subsections.

3.2.1. Problem Formulation AutoDA involves finding an optimal data augmentation policy to enhance the performance of deep learning models. To achieve this goal, most research works formulate the generation of the optimal DA policy as a standard search problem (Hataya et al., 2020; Li et al., 2020). A typical AutoDA model comprises three key components: a search space, a search algorithm, and an evaluation function. The search space defines the set of possible data augmentation operations that can be applied to the input images. The search algorithm aims to identify the optimal set of operations that maximize the performance of the deep learning model. The evaluation function is responsible for evaluating the performance of the model using the augmented data.

More formally, given a dataset $\mathcal{D}$ and a search space of possible augmentation policies Θ, the goal of the AutoDA search problem is to find the optimal data augmentation policy θ^* that maximizes the performance of the deep learning model on the target task. This can be formulated as an optimization problem:

$$\theta^* = \arg\max_{\theta \in \Theta} \mathbb{E}_{\boldsymbol{x},y\sim\mathcal{D}}[\mathcal{L}(f_{\boldsymbol{\omega}}(\boldsymbol{x}'), y)] \tag{1}$$

where $\boldsymbol{x}$ and y denote an input and its label drawn from $\mathcal{D}$, $\boldsymbol{x}'$ is the augmented version of $\boldsymbol{x}$ using the DA policy θ, $f_{\boldsymbol{\omega}}$ is the deep learning model parameterized by $\boldsymbol{\omega}$, and $\mathcal{L}$ is the loss function. The augmentation policy θ consists of a set of transformation functions $T = t_1, t_2, ..., t_n$ with corresponding probabilities $p_1, p_2, ..., p_n$, and intensity parameter α and can be given as:

$$\theta = (t_i, p_i, \alpha_i)_{i=1}^n, \tag{2}$$

where t_i is a transformation function with probability p_i and intensity α_i. The search problem can be solved using a search algorithm that samples different sequences of data augmentation policies from the search space Θ, and an evaluation function that evaluates their performance on the validation set. It

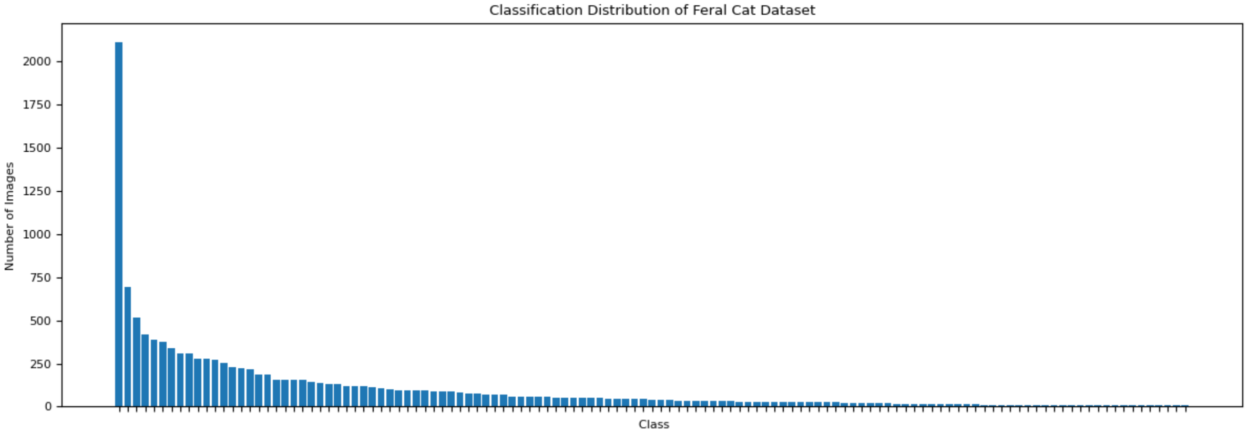

Figure 3. Class Distribution of the Feral Cat Dataset.

updates the next candidate policy based on the obtained reward signal.

3.2.2. Existing AutoDA Approaches A standard AutoDA pipeline involves two stages: a generation stage and an application stage (Yang et al., 2023). In the generation stage, an optimal data augmentation policy is generated for a given dataset by a search algorithm that samples candidate strategies from a defined search space. The efficacy of the searched policies is then evaluated using an evaluation function. In the application stage, the learned policy is applied by augmenting the target dataset using the obtained DA policy to increase both the data quantity and variety. The classification model is then trained on the transformed training set.

Based on the application of these two stages, existing works can be divided into two major categories: one-stage and two-stage approaches. Two-stage approaches involve separate generation and application stages. The optimal augmentation policy is generated in the first stage and then applied to the training set to train the model in the second stage. In contrast, one-stage approaches combine generation and application through the use of gradient approximation methods, allowing simultaneous optimization of the augmentation policy and classification model.

This study aims to evaluate the performance of various AutoDA methods. Specifically, we conduct tests on 4 two-stage approaches and 2 one-stage approaches to determine their respective strengths and weaknesses. The two-stage approaches we tested included TANDA, faster AutoAugment (faster AA) (Hataya et al., 2020), RandAugment (RA) (Cubuk et al., 2020), UniformAugment (UA) (LingChen et al., 2020). TANDA, being the first proposed method in the automated DA field, has inspired the development of subsequent AutoDA approaches. Comparing other methods with TANDA highlights recent advancements in this field. However, since most two-stage approaches are extremely resource-intensive, we use Faster AA due to its balance between performance and efficiency (Cubuk et al., 2019; Hataya et al., 2020). In addition, we include two search-free AutoDA methods to highlight the contrast between these approaches and classical two-stage methods that rely on optimization algorithms. The one-stage approaches we test included differentiable automatic data augmentation (DADA) (Li et al., 2020) and automated dataset optimization (AutoDO) (Gudovskiy et al., 2021). Using both two-stage and one-stage AutoDA approaches, we compare the model performance on feral cat data in terms of accuracy, efficiency, and generalization ability. We also provide insight into which approach may be more suitable for different applications.

3.3. Baseline

We include two baseline models to compare the performance of the AutoDA methods. The first baseline model was a model trained on the original dataset without any data augmentation, which we refer to as the *no augmentation baseline*. This baseline model helps to determine how much the AutoDA methods improve the accuracy of the model compared to just training on the raw (non-augmented) data. The second baseline model used was a model trained on a randomly augmented dataset, which we refer to as the *random augmentation baseline*. Random augmentation involves applying a random set of DA transformations to the training data, which can improve performance to some extent, but is not tailored to the specific dataset or task at hand. It allows to measure the effectiveness of data augmentation with some level of data manipulation, but without any optimization or intelligence in the policy selection.

By comparing the performance of AutoDA methods against these baseline models, we quantify the performance improvement that AutoDA can achieve and hence the strengths and weaknesses of each approach. In addition, we also compare the additional computational cost and time required for each AutoDA method against the two baselines, as this is an important consideration in many real-world applications. Overall, our experiments provide a comprehensive evaluation of the performance of different AutoDA methods, and highlight the potential trade-off of using AutoDA between model accuracy and training efficiency.

3.4. Evaluation Metrics

The evaluation metrics used in this study were selected to provide a comprehensive and informative evaluation of the performance of the different AutoDA methods tested. These include:

Accuracy: The accuracy metric is a widely used and standard metric for measuring the performance of deep learning models. In our study, we measure the accuracy of the models on a test dataset that was independent from the training data. The test dataset was carefully chosen to ensure that it represented a diverse range of examples and was not biased towards any particular class or feature of the data. By measuring accuracy, we were able to determine how well the different AutoDA methods improved the performance of the baseline models and compare their relative performance.

Efficiency: Efficiency is a crucial factor in evaluating the performance of different AutoDA methods. In addition to measuring accuracy, we also measured the computational cost and time required for each AutoDA method. This included different aspects such as the maximum memory usage and the total GPU hours. Memory usage was measured as the maximum amount of memory used during policy training, while the GPU hours reflected the total time used for policy generation and end model training.

By considering these evaluation metrics, we were able to gain insight into the performance of the different AutoDA methods. The combination of accuracy and efficiency metrics also allowed us to make informed decisions when selecting specific AutoDA methods for different applications.

4. Experiments and Results

In this section, we present the experimental evaluation for the different AutoDA approaches including TANDA, Faster AA, RA, UA, DADA and AutoDO using both the CIFAR-10 dataset and feral cat dataset, and compare results with two baseline models. Our goal was to empirically evaluate the performance of these methods based on their accuracy and efficiency. We initially conducted experiments using 6 AutoDA methods on the CIFAR-10 data (Section 4.2.1). We compared our results with the results reported in their original papers. Our findings indicate that the tested AutoDA methods were largely reproducible, as there was not much difference between our results and those published in the literature. Next, we investigated the generalization ability of AutoDA methods across datasets using feral cat data (Section 4.2.2). The dataset used in these experiments was significantly different from the ones used in their original studies, since the feral cat data comprised a skewed class distribution and many low quality images. Our results showed that some AutoDA models can improve the final accuracy of classification models even for new tasks and datasets, while some approaches, e.g. search-free models, could potentially result in a performance drop without prior knowledge of the target domain. In addition to accuracy, we also assessed the computational efficiency of the different AutoDA methods. Overall, our experiments provide a comprehensive evaluation of different AutoDA methods and highlight the potential trade-off of using AutoDA between model accuracy and training efficiency.

4.1. Experimental Setup

Most AutoDA algorithms follow the same problem formulation as AA (Cubuk et al., 2019). Specifically, each policy is comprised of five sub-policies, where each sub-policy includes two augmentation operations. Although the search space in different AutoDA models may vary slightly due to different image processing libraries used (e.g. Python's Pillow and PyTorch), we ensured that the same basic image processing operations were included, namely: ShearX, ShearY, TranslateX, TranslateY, Rotate, AutoContrast, Invert, Equalize, Solarize, Posterize, Contrast, Color, Brightness, Sharpness, Cutout, and Sample Pairing.

Consistent with previous studies, we utilized Wide-ResNet-28-10 as the base model for both CIFAR-10 and feral cat datasets except TANDA due to implementation limitations. To align the experiments with the design of TANDA algorithm, we utilized an all-convolutional CNN with four convolutional layers and leaky ReLU activations as the discriminator, while the generator was implemented using ResNet hence ensuring that it was aligned with the original work (He et al., 2016; Ratner et al., 2017). The implementation of each AutoDA method was adapted from its original source code with minor adjustments made for deployment purposes. In order to ensure a consistent and fair comparison, we adhered to the default experimental settings used in the original works for the CIFAR-10 data. For feral cat data, we reduced the batch size to 32 to accommodate the limited data size and further improve the final model performance. During the policy application phase, all classification models were trained with 200 epochs before being evaluated on the testing set. To consistently assess the efficiency of each method, we conducted all experiments except TANDA on the NVIDIA GeForce 1080Ti GPU to eliminate any potential variability that might arise from different hardware configurations.

4.2. Results

4.2.1. CIFAR-10 In Table 2, we compare the results obtained from our experiments on the different AutoDA methods and the results reported in the original papers. One additional and important aspect of the work is to assess the reproducibility of the methods. We reported average accuracy values over three separate runs based on the same experimental settings. The results presented in the table indicate that our implementations achieved comparable accuracy to the original works with only a slight performance drop. It is worth noting that some of the results in the original papers were reported with different evaluation metrics based on their (individual) experimental settings, which makes it challenging to make an exact comparison. However, all AutoDA approaches investigated here were subjected to identical (collective) experimental conditions, and we employed the same evaluation metrics to mitigate any potential impacts from external factors. Overall, our results confirmed the effectiveness and robustness of the AutoDA methods and provide evidence of their practical applicability.

4.2.2. Feral Cat Data Table 3 presents the experimental results of various AutoDA methods on the feral cat data and their comparisons with the two baseline methods. The Top1 accuracy of the final classification model trained on the augmented data was measured. The reported accuracy values were obtained by taking the highest accuracy achieved out of the three runs. Additionally, the total GPU time and maximum memory usage were recorded for efficiency comparisons. The total GPU time included the time used for both policy generation and model training.

In terms of accuracy, the results show that the best performance was achieved by Faster AA, which obtained a Top1 accuracy of 0.5422. It should be emphasised that there were 123 unique classes (feral cats) with challenging data, e.g., cats with many similar markings, hence accurate classification is especially challenging compared to other image classification scenarios. The random augmentation baseline method had a relatively low accuracy of 0.2453, while the no augmentation baseline performed even worse with a Top1 accuracy of 0.2603. DADA also achieved good accuracy at 0.2953 and was the second-best performing method. The RA and UA methods achieved Top1 accuracies of 0.2332 and 0.2299 respectively. The AutoDO method achieved a Top1 accuracy of 0.2246, slightly worse than both baselines.

We observed that TANDA had the Top1 accuracy between 0.2043 to 0.3632. However, it is worth noting that TANDA has a large variation in performance due to the multiple models trained in one run. The reported values represent the worst and best accuracy achieved out of three runs. Moreover, we did not measure the total GPU time for TANDA as it was originally designed to be executed on a CPU instead of a GPU.

In terms of efficiency, AutoDO was the most time-efficient method with a total GPU time of 3.4238 hours, followed by RA with 8.399 hours. However, both methods had relatively low accuracy values of 0.2488 and 0.2332 respectively. UA had the highest maximum memory usage of $3,059$MiB, while DADA used the most GPU memory with $4,275$MiB.

5. Discussion

Our results demonstrate the effectiveness and potential of different AutoDA methods for improving model performance and reducing the need for manual hyperparameter tuning. Specifically, we compared the performance of TANDA, Faster AA, RA, UA, DADA and AutoDO. Two baselines (no augmentation and random augmentation), were used to evaluate the effectiveness of the AutoDA methods. The results indicated that some AutoDA methods tested were able to improve the performance of the final classifier, as compared to both baselines. This indicates that selecting the appropriate AutoDA method based on the given task can effectively improve image classification models. Additionally, it can address the limitations of traditional manual data augmentation methods.

Faster AA achieved the highest accuracy on both datasets, followed by TANDA. Both methods are based on two-stage AutoDA approaches, in which the generation of augmentation policies requires a resource-demanding search phase based on the original data. Therefore, Faster AA used the most total GPU time to complete one run. Due to the difference in deployment, we were unable to compare the efficiency

Table 2. Accuracy of Results on CIFAR-10 using different AutoDA Methods

AutoDA method	Accuracy (ours)	Accuracy (original)	Deviation
TANDA	0.732 ± 0.046	0.815	0.0826
Faster AA	0.952 ± 0.002	0.963	0.011
RA	0.966 ± 0.001	0.973	0.007
UA	0.973 ± 0.003	0.973	~ 0
DADA	0.937 ± 0.002	0.945	0.008
AutoDO	0.946 ± 0.005	0.951	0.005

Table 3. Summary table of performance and efficiency results using different AutoDA methods on feral cat data.

AutoDA method	Top1 accuracy	Total GPU time (hrs)	Max memory usage
No-aug (baseline)	0.2603 ± 0.017	-	-
Random-aug (baseline)	0.2453 ± 0.034	-	-
TANDA	0.2828 ± 0.114	-	-
Faster AA	0.5422 ± 0.021	39.525	1057MiB
RA	0.2332 ± 0.055	8.399	993MiB
UA	0.2299 ± 0.057	7.012	3059MiB
DADA	0.5090 ± 0.028	29.179	4275MiB
AutoDO	0.2488 ± 0.024	3.424	2097MiB

of TANDA with other methods. However, it took more than 48 hours to complete the training and testing of the CNN in TANDA. Our findings are consistent with previous studies that have shown the effectiveness of two-stage approaches for optimizing deep learning models (Hataya et al., 2020; Ratner et al., 2017; Yang et al., 2023). However, it is important to note that these methods are computationally expensive and require a significant amount of computational resources.

In our experiments, DADA achieved comparable accuracy levels to some of the TANDA models while requiring less time to complete. However, as a one-stage AutoDA method, DADA only showed modest improvement in accuracy when compared to no augmentation baseline. On the other hand, AutoDO was the least effective method among all tested models except two search-free approaches, achieving an accuracy of only 0.2488 on the feral cat data. This performance loss can be attributed to the gradient approximation used in one-stage AutoDA algorithms. In order to accelerate the overall training time, one-stage methods estimate the gradient of augmentation parameters and simultaneously generate and apply the DA policy. This approach could lead to sub-optimal DA policies because the estimation process may be inaccurate during the parameter tuning, especially when compared to two-stage approaches.

We also observed that not all AutoDA methods resulted in performance improvement. Indeed RA and UA had a negative impact on the model performance. These methods fall under the category of search-free AutoDA as they bypass the search phase of DA policies. Instead, RA and UA directly sample DA policies from a predetermined range via random sampling and uniform sampling respectively. The augmentation policies are selected from a set of image transformation functions that are empirically chosen based on prior knowledge of the given task. As a result, the final performance of search-free methods largely depends on the decision of the policy range. Although search-free approaches greatly enhance the efficiency of generating DA policies, they can lead to a loss in performance as the augmentation policies are not necessarily optimised for the target data. For the feral cat task, the policy range might not include the optimal DA operations due to the lack of knowledge about the data, which could impede the effectiveness of the methods.

Overall, our results suggest that the performance of AutoDA methods is highly dependent on the specific algorithm and the search strategy used. Two-stage approaches have shown promising results, but their computational cost can be significant. While DADA showed promising results in terms of time efficiency and accuracy, one-stage AutoDA methods, in general, may not be as effective as two-stage methods due to the potential for sub-optimal DA policy generation. Search-free methods are recognized for their efficiency and require the least amount of computational resources, however, their effectiveness is limited and can even have a negative impact on model performance. Future research could focus on developing improved gradient estimation techniques for one-stage methods or refining the policy range of search-free approaches.

Figure 4 indicates the trade-off between accuracy and efficiency in AutoDA methods. In general, one-stage methods such as DADA and AutoDO, demonstrate much faster performance in terms of computational resources and time required for training when compared to two-stage methods. However, one-stage methods often achieve limited improvements in accuracy when compared to two-stage methods.

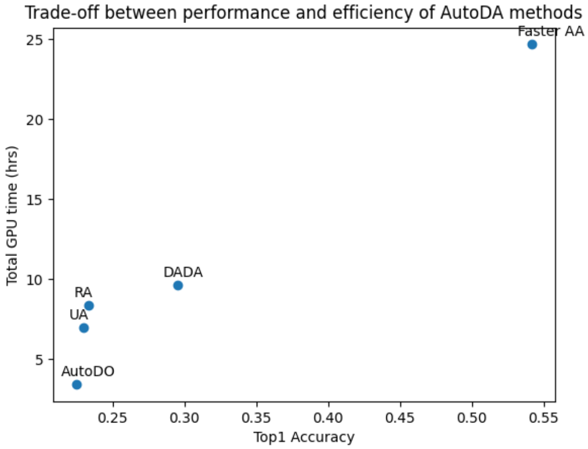

Figure 4. Trade-off between performance and efficiency of AutoDA methods on feral cat classification task. The plot displays the relationship between total GPU time used for policy generation model training (y-axis) and top 1 accuracy of the final classification model after training (x-axis).

In contrast to one-stage methods, two-stage AutoDA methods tend to produce improved results but require more computational resources as shown in Figure 4. TANDA and Faster AA are two examples of two-stage methods that use gradient-based optimization to search for the best augmentation policies. Although these methods require more GPU hours than other methods, they achieve the highest accuracy. This is likely due to the fact that they can take advantage of the information from the proxy task during the search phase to optimize the augmentation policies, albeit the accuracy improvement achieved by two-stage methods comes at the expense of longer training times and higher GPU usage.

Furthermore, search-free methods, such as RA and UA, were the most efficient as they did not require any policy search, with total time equaling the network training time. However, the performance of search-free approaches largely depends on the selection of the augmentation policy range, which can lead to performance degradation if the range is not properly chosen.

The choice between one-stage and two-stage approaches as well as search-free and search-based methods should be carefully considered depending on the specific use case and the available resources. For instance, search-based methods can offer higher accuracy but require more computational resources, while search-free methods are more efficient but may not always result in optimal performance.

Depending on the type of data being used, the augmentation policies selected by AutoDA techniques can vary greatly. For instance, when working with CIFAR-10 data, AutoDA models predominantly choose color-based transformations, such as Brightness, AutoContrast, Color, and Solarize, which is in line with prior research (Cubuk et al., 2019). However, geometric transformations like ShearX and ShearY are rarely used in successful policies for CIFAR-10 data.

On the other hand, when dealing with feral cat classification task, geometric transformations such as ShearY, ShearX, and TranslateY are commonly employed in the final policies generated by AutoDA models. Furthermore, we observed that Cutout is extensively utilized to augment the raw data. This is possibly due to the tendency of feral animals to hide behind environmental objects, which often results in occlusion in the training set. As such, it is crucial to prepare the model to accurately classify the cat even when it is obscured by environmental objects. In contrast, color-based transformations like Solarize and Equalize are less frequently selected by AutoDA models. This may be attributed to the fact that the dataset mainly comprises images captured at night with an infrared camera, resulting in black and white images that limit the effectiveness of color adjustments.

6. Conclusions and Future Work

In this study, we evaluated the performance of several AutoDA methods for image classification tasks. Our results demonstrated a trade-off between accuracy and efficiency of AutoDA methods, with one-stage methods being faster but achieving modest improvements in accuracy, while two-stage methods produce more consistent and higher accuracy results but require more computational resources. We also found that search-free methods were the most efficient but that an inappropriate selection of policy range could lead to performance loss.

This work study has several limitations that should be addressed in future research. Further studies are needed to evaluate the performance of AutoDA methods on a wider range of datasets and deep learning models. Such studies can help determine the generalizability and scalability of these methods and their applicability for various real-world applications. The AutoDA methods evaluated in this study are all computationally expensive, making them challenging to use in many real-world scenarios. Future research might explore more efficient AutoDA methods that can be used on large-scale datasets and with more complex models. This can help overcome the computational limitations of current AutoDA methods and enable their broader adoption in various domains.

AutoDA methods also have the potential to improve the efficiency and accuracy of deep learning models, but there is still much to explore and improve in this field. As deep learning continues to advance and become more widely adopted, the development and optimization of AutoDA techniques will become increasingly important.

References

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 113–123.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Gudovskiy, D., Rigazio, L., Ishizaka, S., Kozuka, K., & Tsukizawa, S. (2021). Autodo: Robust autoaugment for biased data with label noise via scalable probabilistic implicit differentiation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16601–16610.

Hataya, R., Zdenek, J., Yoshizoe, K., & Nakayama, H. (2020). Faster autoaugment: Learning augmentation strategies using backpropagation. *European Conference on Computer Vision*, 1–16.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, Y., Hu, G., Wang, Y., Hospedales, T., Robertson, N. M., & Yang, Y. (2020). Dada: Differentiable automatic data augmentation. *arXiv preprint arXiv:2003.03780*.

LingChen, T. C., Khonsari, A., Lashkari, A., Nazari, M. R., Sambee, J. S., & Nascimento, M. A. (2020). Uniformaugment: A search-free probabilistic data augmentation approach. *arXiv preprint arXiv:2003.14348*.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc.

Ratner, A. J., Ehrenberg, H. R., Hussain, Z., Dunnmon, J., & Ré, C. (2017). Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, *30*, 3239.

Rees, M., Pascoe, J., Wintle, B., Le Pla, M., Birnbaum, E., & Hradsky, B. (2019). Unexpectedly high densities of feral cats in a rugged temperate forest. *Biological Conservation*, *239*, 108287.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1–48.

Simard, P. Y., Steinkraus, D., Platt, J. C., et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. *Icdar*, *3*(2003).

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE international conference on computer vision*, 843–852.

Yang, Z., Sinnott, R., Ke, Q., & Bailey, J. (2021). Individual feral cat identification through deep learning. *2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies (BDCAT'21)*, 101–110.

Yang, Z., Sinnott, R. O., Bailey, J., & Ke, Q. (2023). A survey of automated data augmentation algorithms for deep learning-based image classification tasks. *Knowledge and Information Systems*, 1–57.