MetaAudio: A Few-Shot Audio Classification Benchmark

MetaAudio: A Few-Shot Audio Classification Benchmark*

Calum Heggan[1][0000−0002−5694−1577], Sam Budgett[2][0000−0002−1137−7255], and Timothy Hospedales[3][0000−0003−4867−7486] and Mehrdad Yaghoobi[4][0000−0002−9847−8234]

University Of Edinburgh, Edinburgh, Scotland
s1529508@sms.ed.ac.uk

Abstract. The currently available benchmarks for few-shot learning (machine learning with few training examples) are limited in the domains and settings they cover, primarily focusing on image classification. This work looks to alleviate this reliance on image-based benchmarks by offering the first comprehensive, public and fully reproducible audio based alternative, covering a variety of sound domains and experimental settings. We compare the few-shot classification performance of a variety of techniques on seven unique audio datasets (spanning from environmental sounds to human-speech). Extending this, we carry out in-depth analyses of the joint training routine (where all datasets are used during training) and cross-dataset/domain adaptation, establishing the possibility of a generalised audio few-shot classification algorithm. Our experimentation shows gradient-based meta-learning methods such as MAML and Meta-Curvature consistently outperform both metric and baseline methods. We also demonstrate that the joint training routine helps overall generalisation for the environmental sound databases included, as well as being a somewhat-effective method of tackling the cross-dataset/domain setting.

1 Introduction

To date, the majority of the breakthroughs seen in machine learning have been in domains or settings where there was an abundance of labelled data, either real or simulated, for example in [12]. In contrast, the capability for humans to recognise and discriminate between types of classes/sensory inputs with few examples, e.g. in visual or acoustic settings, remains unmatched. The development of techniques that can perform such Few-shot Learning (FSL) tasks has seen significant interest within modern machine-learning literature, with particular focus on applying meta-learning (learning to learn). This approach allows machine learning to be applied to new tasks where classes are rare or labelled data is hard to produce or gather.
Most work looking at improving the generalisation performance of these types

* Supported by EPSRC, UDRC & Thales.

of algorithms is focused around the image domain with other data modalities and problem settings largely underrepresented. This potentially biases meta-learning algorithmic development towards images, therefore loosing potential performance on other types of data or task. Compounding this, the most commonly evaluated on datasets, e.g.Mini-ImageNet [16], as well as current benchmarks suffer from lack of real-world challenges.
Acoustic classification and event detection have seen a significant number of works in conventional fully supervised machine learning [6,11], with many public datasets having a standardised evaluation protocol that is adhered to by the community and allows for standardisation and fair comparison. This has however not extended to the few-shot equivalent, where the majority of the works that do exist make little attempt at preserving reproducibility, typically with respect to dataset management and lack of public source code [2,13]. This absence of standardisation poses significant issues when looking to compare novel and existing methods alike.
In this work, we look to alleviate this gap as well as tackle of some of the previous issues outlined by contributing the following: 1) Experimental evaluation of some of the most popular few-shot classifiers on a variety of audio datasets, spanning multiple sub-settings from environmental sounds to speech. 2) A fully reproducible few-shot audio classification benchmark with at least one published evaluation split per dataset along with custom data loading allowing for quick plug and play testing in future works. 3) A generalised prescription for dealing with variable length audio datasets in a few-shot setting. 4) Finally, in-depth analyses and evaluation of the joint training and cross-dataset/domain settings. We include all of our code in this GitHub repository.

2 Few-Shot Classification

2.1 Formulation

The general setting of few-shot learning involves training, validating and testing a model on disjoint sets of classes (e.g. classes of human non-speech sounds such as sneezing and coughing), $\mathcal{C}_{train} \notin \mathcal{C}_{val} \notin \mathcal{C}_{test}$. These sets of classes can be thought of as analogous to the training, validation and test data splits found in traditional machine learning, where splits have non-overlapping samples. In few-shot learning, splits are defined with the addition of non-overlapping classes. The goal of a few-shot classifier is to generalise to a set of $\mathcal{N}$ novel classes, given only a few-labelled examples from each class. These episodes (also referred to as tasks throughout) contain both a support set $\mathcal{S}$ which is used for training and a query set $\mathcal{Q}$ which is evaluated on.
Meta-Learning can either be trained with episodic training, where individual few-shot tasks are drawn from $\mathcal{C}_{train}$, or non-episodic training, where a simpler classifier is trained on all classes contained in $\mathcal{C}_{train}$ in order to learn an embedding in the second to last network layer.

2.2 Meta-Learners & Other Approaches

Due to the number of meta-learning algorithms created and distributed in recent literature, discussed more in Section 3, we restrict our experimentation to a hand-selected and representative few. Specifically these are; Prototypical Networks [14], Model-Agnostic Meta-Learning [4], Meta-Curvature [10], SimpleShot [17] & Meta-Baseline [1]. This selection covers both metric and gradient-based meta-learning, as well as extensions to simpler baseline methods. We leave the specific details of the algorithms to the original papers and instead offer a very high level overview.

Prototypical Networks Part of the metric learning family of meta-learners, Prototypical Networks (ProtoNets) work by calculating class prototypes of the embedded support set, followed by the use of a distance function whose minimum can be used for classifying queries.

MAML & Meta-Curvature These gradient-based approaches [4,10] aim to learn a transferable initialisation for any model such that it can quickly adapt to a new task τ with only a few steps of gradient descent. At training time, the meta-objective is defined as query set performance after a few steps of gradient descent on the K support samples from the model's initial parameters. Meta-curvature expands on MAML by also learning a transform of the inner optimisation gradients such that the gradients themselves achieve better generalisation on new tasks. In this work, we experiment and report results with first order variants of these algorithms, as in initial experimentation comparing both variants we observed negligible or negative effects to performance.

SimpleShot & Meta-Baseline SimpleShot [17] and Meta-Baseline [1] are present in this work as baseline approaches, methods that aim at lowering computational cost and still achieving strong performance. Both methods train in a conventional way, outputting legits directly from a linear layer of size $|C_{train}|$, and validate/test using nearest centroid classification. To distinguish themselves, SimpleShot applies additional data transforms at test time, while Meta-Baseline performs episodic fine-tuning, similar to what is seen in ProtoNets but with cosine distance and logits scaling.

3 Related Work

Few-Shot Classification We consider only a small subset of available meta-learners in this work meaning many fall out of scope, this includes both extensions to learners used here and others that are more unique. MAML and Meta-Curvature most closely relate to other gradient-based meta-learning (GBML) schemes, designed around the idea of fast adaptation to new learning tasks using additional gradient descent steps. These include works such as Meta-SGD [8]. The included prototypical networks [14] relates mostly to other metric learners, which generally aim to learn a strong feature embedding space such that

support and queries can be compared. Included in this family are works such as Matching Networks [16]. All of these algorithms have been evaluated in the image domain to some degree and so are all relevant in that their performance in other domains, audio included, is largely unknown.

Few-Shot Acoustics Currently only a handful of studies exist that look at either few-shot audio classification or event detection. Of these, two are set in event detection [13,2] (classification of part of audio clip in time) with the other two focused on classification [3,18] (classification of entire audio clip), the focus of this work. Comparing these works, we see a variety of approaches taken toward dataset processing, split formulation and reproducibility. These variations, most importantly the dataset and its associated split used, make comparisons and ranking of the works impossible. Of these, one is distinct in that it provides both a fully reproducible code base and the dataset class-wise splits used for its experiments. The work's main contribution looks at fitting common metric based learners with an attention similarity module, attached to its purely convolutional backbone. This work is currently state-of-the-art for both the ESC-50 [11] dataset and its proprietary noise injected variant 'noiseESC-50'. As discussed more in Section 4, we use this work as a basis for some of the experiments carried out.

Benchmarks Most relevant to this work are other few-shot and meta-learning benchmarks. Included in this are works such as Meta-Dataset [15] (an aggregation of 10 few-shot image based datasets) and MetaCC [7] (a modifiable set of channel coding tasks). Of the benchmarks currently available for few-shot classifier evaluation, none deal with acoustic classification. This is the primary area that this work aims to fill. Meta-Dataset is of particular relevance to this work as we aim to mimic both the depth and reproducibility achieved by the benchmark. Specifically, both the within and cross dataset evaluations as well as the public leaderboard are components which we find to be useful.

4 MetaAudio Setup

4.1 Setting & Data

As MetaAudio aims to be a diverse and reproducible benchmark, it covers a variety of experimental settings, algorithms and datasets. Throughout, we mainly consider 5-way 1-shot classification, with some additional analysis of the impact of k-shots and N-ways at test time.
We experiment with 7 total datasets, 5 of which we call our primary datasets which we split up for use in training and evaluation, and 2 held-out sets we use exclusively for testing. Of these 7, 3 are fixed length, and 4 are variable length. Additional details about the datasets, including size and specific setting, can be found in Table 1. Due to the massively variable sample size of the original

dataset and the issues that it presents with reproducibility, we primarily experiment with a pruned version of BirdClef 2020, where samples longer than 180s are removed along with classes with fewer than 50 samples.

Splits & Labels For every experiment setup, we apply a 7/1/2 train-validation-testing split ratio over all the classes belonging to an individual dataset. These ratios are chosen to be in line with the majority of machine learning and few-shot works. Any conventional sample based train/val/test splits are ignored, and the class splits are applied to all available data. Outside [3], from which we can obtain a reproducible split of ESC-50, we have no works with prior dataset splits to follow, and so we define our own. Most simply we assign random splits based on the available classes for a given set. However, we also define within-dataset domain-stratification and shift splits for sets that have additional internal structure and/or accompanying meta-data. Extensive experimentation with these more specific splits is not carried out in this work, instead favouring other experiments, however are included in our repository.
Labels for the datasets vary quite significantly with some having time precise, or strong, labels like BirdClef2020 with others having only whole clip-level, or weak, labels. In interest of consistency, for this work we drop the available strong labels for the datasets that have them and operate exclusively with weak labels. The tradeoff of this approach is that for datasets that have access to strong labels, we expect additional label noise to be present during training, possibly hurting final generalisation performance.

General Processing Pre-processing is kept minimal, with only the conversion of raw audio samples into spectrograms and some normalisation factor applied during loading. During this pipeline, we consider a fixed sample rate and spectrogram parameters over all datasets and contained samples. For normalisation, three techniques were considered; per sample, channel wise and global. Following initial experimentation, global, which uses average statistics across all examples, was used in all experiments due to performance and simplicity.

Table 1. High level details of all datasets considered in MetaAudio

Name	Setting	N^o Classes	N^o Samples	Format	Sample Length
ESC-50	Environmental	50	2,000	Fixed	5s
NSynth	Instrumentation	1006	305,978	Fixed	4s
FDSKaggle18	Mixed	41	11,073	Variable	0.3s - 30s
VoxCeleb1	Voice	1251	153,516	Variable	3s - 180s
BirdCLEF 2020	Bird Song	960	72,305	Variable	3s - 30m
BirdCLEF 2020 (Pruned)	Bird Song	715	63,364	Variable	3s - 180s
Watkins Marine Mammal Sound Database	Marine Mammals	32	1698	Variable	0.1 - 150s
SpeechCommandsV2	Spoken Word	35	105,829	Fixed	1s

4.2 Sampling Strategies

Throughout MetaAudio, we utilise a variety of sampling strategies for experimentation. The basis of these is our fixed length approach. The steps used for this can be summarised into: 1) Sample a set of N-way classes $\mathcal{C}_{\mathcal{N}}$ from the necessary split of dataset $\mathcal{D}$ and 2) For each class $\mathcal{C}_{\mathcal{N}}$, sample both support and query examples, for support the number will be k-shot.
We extend this fixed length strategy in order to build a method for dealing with variable length sets. Due to how varied sample length is within some of the considered datasets, we opt for fixed length representation to avoid the need for specific neural architectures which can naturally deal with variable length or incredibly powerful hardware. Specifically, we choose to split our variable length samples up into $\mathcal{L}$ length sub-clips. This along with the later conversion of the sub-clips to individual spectrograms is done entirely offline, a decision made to avoid bottlenecking during training.
Combining a variety of datasets in a joint training and/or evaluation routine has already seen some focus in the image space [15]. We mimic this and expand upon it for the considered acoustic datasets and task. Sampling tasks from the available datasets in this setting can be done in a few distinct ways, none of which are an immediately better choice with respect to downstream generalisation capability. We consider this to be an additional area of investigation. Specifically, we propose two variants of sampling, one which allows task construction between datasets, meaning that the $\mathcal{N}$ classes sampled could belong to different sets, and one which does not. We refer to these techniques as 'Free Dataset Sampling' and 'Within Dataset Sampling' respectively.
During these sampling strategies, we largely ignore the class sample imbalance seen in the majority of the dataset we experiment with, we do this for a few reasons. The first of these is that recent works, such as [9], suggest it is less detrimental than in conventional learning. The second is that, these imbalances allow algorithms to differentiate themselves with respect to how they handle the more difficult setting. One area in which we do experiment with alleviating the effect of this imbalance is in the re-weighting of the loss functions used in the conventional learning parts of the Meta-baseline and SimpleShot algorithms. To create this dataset custom loss for these scenarios, we employ inverse-frequency class weighting, where the class-wise contribution to the loss function is the inverse of the number of samples present in that class.

5 Experiments

5.1 Details

Experimental results presented are collected similarly to adjacent few-shot works, where reported classification accuracies along with their 95% confidence intervals are the conclusion of just one end-to-end training and evaluation procedure, where 10,000 tasks drawn from the test set have been considered. For all experiments we use Adam, with no early stopping and a fixed learning rate. Tuning

of algorithms was kept to a minimum however was still performed. Due to this minimal approach, we expect it to be fairly trivial to obtain a specific result marginally better than those presented, however it is important to note that this does not undermine the comparison and experimental settings investigated in this work.

Motivated by the increasing performance gap between the commonly used CNNs and other neural architectures currently present in conventional acoustic learning, as seen in works like [6,5], we briefly investigated the role of the base learner in the few-shot acoustic setting. Due to space restriction, we don't include the explicit setup or results here, however we note that our best performing model using MAML and ProtoNets on ESC-50 was a lightweight hybrid CRNN. Due to this and it's relatively low computational cost compared to more heavily parametrised models, we opt to the model throughout.

In the majority of the results presented for variable length datasets, the value of $\mathcal{L}$ is set to 5 seconds. We chose this value based on external experiments (not included due to space limitations) where, for Kaggle18, $\mathcal{L} = 5s$ performed best on average when compared against 1 and 10-seconds. Setting a universal value of $\mathcal{L}$ also allows us to more comfortably facilitate joint training and cross-dataset evaluation without the need for massive padding.

5.2 Baseline Splits

The main contribution of this work looks at benchmarking fixed splits of datasets within a variety of few-shot learning algorithms. From Table 2 a), we can identify a few interesting behaviours. However, first we note that the ESC-50 ProtoNet using the CRNN backbone performs at least as well as the CNN in the same algorithm used in [3] with the same split.

Out of the two fixed length sets, ESC-50 appears to be the harder problem, with classification accuracies much lower than in NSynth. This is somewhat expected given the problem setting of NSynth where the discriminations are being made between classes and samples that are both cleaner in origin and more related to one another, compared to the more varied and noisy classes belonging to ESC-50. This idea is backed up by a few observations, firstly that our metric learning algorithms (specifically ProtoNets which has no test-time adaptation and assumes similar tasks between training and testing) perform significantly better than MAML which has adaptation capabilities. How separated the performances of MAML and Meta-Curvature are also support this idea, as the main difference between the two is Meta-Curvature's ability to learn local gradient curvatures, which performs best under the assumption of more similar tasks. The variable length datasets appear to represent harder settings in general, with significant drops off in average performance. Specifically, we see very low classification accuracy for Kaggle18, a behaviour likely due to larger amounts of label noise. Over all datasets, we see that the GBML methods performs very well, with Meta-Curvature taking SOTA in 4 out of 5 cases, and MAML in the 5th. We propose that this is due to the aforementioned adaption mechanism, making it particularly useful for settings which have classes of higher intra and inter-variance. In

Table 2. Headline and main benchmark 5-way 1-shot classification results. Table a) contains the baseline results, where models are trained for each dataset individually and then evaluated with that datasets test split. Tables b) and c) contain results from the joint training scenario, where we train meta-learners over all datasets simultaneously and then evaluate on individual test splits.B) and c) differ in that in b) we only allow tasks to be samples using classes from one of the datasets, whereas in c) we allow cross-dataset task creation.

a) Baseline Within Dataset Results

	Dataset	FO-MAML	FO-Meta-Curvature	ProtoNets	SimpleShot CL2N	Meta-Baseline
	ESC-50	74.66 ± 0.42	**76.17 ± 0.41**	68.83 ± 0.38	68.82 ± 0.39	71.72 ± 0.38
	NSynth	93.85 ± 0.24	**96.47 ± 0.19**	95.23 ± 0.19	90.04 ± 0.27	90.74 ± 0.25
	Kaggle18	**43.45 ± 0.46**	43.18 ± 0.45	39.44 ± 0.44	42.03 ± 0.42	40.27 ± 0.44
	VoxCeleb1	60.89 ± 0.45	**63.85 ± 0.44**	59.64 ± 0.44	48.50 ± 0.42	55.54 ± 0.42
	BirdClef (Pruned)	56.26 ± 0.45	**61.34 ± 0.46**	56.11 ± 0.46	57.66 ± 0.43	57.28 ± 0.41
	Avg Algorithm Rank	2.4	1.2	3.8	4.0	3.6

b) Joint Training (Within Dataset Sampling)

Trained	ESC-50	68.68 ± 0.45	**72.43 ± 0.44**	61.49 ± 0.41	59.31 ± 0.40	62.79 ± 0.40
	NSynth	81.54 ± 0.39	82.22 ± 0.38	78.63 ± 0.36	**89.66 ± 0.41**	85.17 ± 0.31
	Kaggle18	39.51 ± 0.44	**41.22 ± 0.45**	36.22 ± 0.40	37.80 ± 0.40	34.04 ± 0.40
	VoxCeleb1	**51.41 ± 0.43**	51.37 ± 0.44	50.74 ± 0.41	40.14 ± 0.41	39.18 ±0.39
	BirdClef (Pruned)	**47.69 ± 0.45**	47.39 ± 0.46	46.49 ± 0.43	35.69 ± 0.40	37.40 ± 0.40
Cross	Watkins	57.75 ± 0.47	**57.76 ± 0.47**	49.16 ± 0.43	52.73 ± 0.43	52.09 ± 0.43
	SpeechCommands V1	25.09 ± 0.40	**26.33 ± 0.41**	24.31 ± 0.36	24.99 ± 0.35	24.18 ± 0.36
	Avg Algorithm Rank	2.0	1.6	4.0	3.4	4.0

c) Joint Training (Free Dataset Sampling)

Trained	ESC-50	**76.24 ± 0.42**	75.72 ± 0.42	68.63 ± 0.39	59.04 ± 0.41	61.53 ± 0.40
	NSynth	77.71 ± 0.41	83.51 ± 0.37	79.06 ± 0.36	**90.02 ± 0.27**	85.04 ± 0.31
	Kaggle18	44.85 ± 0.45	**45.46 ± 0.45**	41.76 ± 0.41	38.12 ± 0.40	35.90 ± 0.38
	VoxCeleb1	39.52 ± 0.42	39.83 ± 0.43	40.74 ± 0.39	**42.66 ± 0.41**	36.63 ± 0.38
	BirdClef (Pruned)	**46.76 ± 0.45**	46.41 ± 0.46	44.70 ± 0.42	37.96 ± 0.40	32.29 ± 0.38
Cross	Watkins	**60.27 ± 0.47**	58.19 ± 0.47	48.56 ± 0.42	54.34 ± 0.43	53.23 ± 0.43
	SpeechCommands V1	**27.29 ± 0.42**	26.56 ± 0.42	24.30 ± 0.35	24.74 ± 0.35	23.88 ± 0.35
	Avg Algorithm Rank	2.1	2.1	3.4	3.0	4.3

comparison, our metric and baseline algorithms underperform, suggesting that they likely trade off performance for speed (specifically at inference time), except in cases of simple and similar tasks. Out of these, Meta-Baseline performs most competitively, possibly suggesting that combining some traditional learning followed by episodic learning is a more favourable approach.

5.3 Joint Training & Cross Dataset

Our investigation with joint training is two-fold. We first consider the joint training to individual testing regime, i.e. training using all datasets and then using the model to test on each dataset's test split separately. As well as this, we also look to evaluate the cross-dataset performance, by applying the trained models directly on some held-out datasets. Chosen datasets for held-out testing are detailed earlier in Table 1. For the Watkins Mammal Sound Database, we process

with a $\mathcal{L}$ value of 5 seconds, chosen as it closely resembles the expected value of the dataset's sample length distribution. Results for these sampling techniques can be found in Table 2 b) and Table 2 c) respectively.
First, we consider the joint training regime as a whole (so including both sampling regimes) and contrast against the more holistically trained and evaluated within dataset experiments. For both ESC-50 and Kaggle18 we obtain new SOTA results with MAML and Meta-Curvature respectively, both from the free dataset sampling routine. For all other datasets for which we have baseline results (Table 2 a)), we see a degradation of performance. This difference varies in magnitude between datasets and sampling routines. One possible explanation of this behaviour (with the possible exception of NSynth) is that these sets are more domain specific than the likes of ESC-50, meaning that the features needed to successful discriminate between their classes are likely not easily learned by a joint training routine. Some evidence supporting this is that for VoxCeleb and BirdClef, 3/4 of the best results over both sampling routines are in gradient-based methods, which have some opportunity to adapt at test time. Moving to the other fold of our joint training interest, we query how the sampling routines directly compare to one another, as well as how they perform on the held-out cross-dataset tasks. For our main datasets, we observe 3/5 of the top results were obtained using the free sampling method, with the 2 outliers belonging to VoxCeleb and BirdClef - further evidence that their tasks require significantly different model parametrisation, as the within dataset task sampling would allow more opportunity to learn these more specialised features. For the cross-dataset tasks, we also see the strongest performance coming from the free sampling routine, where it outperforms it's within dataset counterpart by $\sim$2% in both held-out sets. As for the absolute performances obtained on the held-out sets, we see that our joint training transfers somewhat-effectively, with the model in one case attaining a respectable 50-60% and another obtaining accuracies just above random.

5.4 External Data & Pretraining

A full training and evaluation pipeline for a specific dataset can be incredibly expensive and arguably deviates from the goal of an inclusive training policy for meta-learning. In this vein, we both frame and experiment with the pretraining to simple classifier setting. We primarily aim for this area of the benchmark to be more varied than the other scenarios covered, where the use of any relevant and fair external data is permitted.
For our experiments, results presented in Table 3, we employ a pretraining step with a proprietary subset of AudioSet (trained model included in code repo) followed with a variety of simple classifiers for testing.
Except for the cases of ESC-50 and Kaggle18, we see very low transfer performance using AudioSet pretraining with no fine-tuning, with performance on NSynth halving compared to its lowest scoring individually trained and evaluated counterpart. This result likely follows due to the classes contained within our subset leaning more toward environmental sounds than musical instruments,

Table 3. 5-way 1-shot performance on our main datasets using a pretrained AudioSet model and a variety of simple linear classifying methods. We compare these to the results for SimpleShot using dataset specific training and evaluation.

Dataset	SimpleShot (UN)	SimpleShot (CL2N)	SVM	SimpleShot (CL2N) from Table 2 a)
ESC-50	72.00 ± 0.37	**72.19 ± 0.37**	71.98 ± 0.35	68.82 ± 0.39
NSynth	45.65 ± 0.43	44.47 ± 0.43	45.27 ± 0.43	**90.04 ± 0.27**
Kaggle18	36.53 ± 0.40	38.60 ± 0.41	37.43 ± 0.40	**42.03 ± 0.42**
VoxCeleb1	26.26 ± 0.36	26.58 ± 0.36	25.79 ± 0.35	**48.50 ± 0.42**
BirdClef (Pruned)	31.28 ± 0.38	32.31 ± 0.39	31.22 ± 0.38	**57.66 ± 0.43**
Avg Rank	2.8	2.2	3.4	1.6

speech or bird song. We also see a clear benefit here of using the centred and L2-normalised feature extractions for SimpleShot, corroborating the claims made in the initial work [17].

5.5 N-Way k-Shot Analysis

Although we only trained and evaluated on the task of 5-way 1-shot, we are interested in the effect of larger shots and wider ways on algorithm performance. To bridge this gap, we experiment with these components at test time, using our already trained 5-way 1-shot models. We consider all of our primary datasets and algorithms, covering values of N from 5-30, and k from 1-30. Varying N-ways and k-shots are treated separately and not stacked. This, for example, means that the 15-way 15-shot setting is never considered, but all of 5-way 1, 5, 10, ..., 30-shot are. We do this in order to avoid the compounding computational complexity of the problem. For algorithms which have a fixed size head output (i.e. GBML methods) we exclude the varying N-ways and focus on the k-Shot analysis. The only other exception to note is that both ESC-50 and Kaggle18 have only 10 and 7 classes belonging to their test sets respectively, and so analysis further than 10/5-way is impossible. We include a sample of these result plots in Figure 1. Varying the numbers of shots used, we observe a clear trend of GBML methods

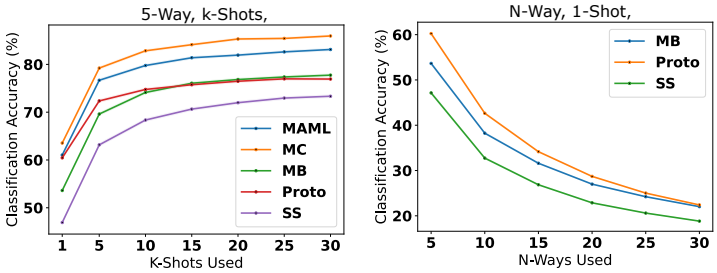

Fig. 1. N-Way k-shot analysis plot for the VoxCeleb1 dataset.K-shots (left) and N-ways (right).

outperforming baseline and metric learning approaches. This appears especially

true for large k-shots, where the rise in performance also occurs faster. For both fixed length sets, we see some additional distinction between gradient based methods and the others, where methods without adaptation not only stagnate heavily after 5-shot, but also start to decrease in performance. Up to 30-shot, we do not observe this same behaviour in variable length sets, however it is possible that this is simply due to the complexity of the problems relative to the fixed length sets, and that by increasing k-shot further we would see similar trends. Of the three non-gradient-based methods, which algorithm performs best over the variety of k-shots appears to be dataset specific, with each outperforming in at least one dataset. Although we are more limited in varying the number of ways we test over, we still observe some interesting trends. All of our tested algorithms show a non-linear decay in performance, with results at 30-way still reaching $\sim$20-25% for our VoxCeleb and BirdClef sets (approx 7$\times$ random). For speed of drop-off, we see a similar story as we saw in increasing k-shot, with all algorithms showing best performance in at least one set.

6 Conclusion

In this work, we presented MetaAudio, a new large-scale and diverse few-shot acoustic classification benchmark. We experimented with a variety of algorithms and datasets, covering a variety of sound domains and experimental settings. For both our baseline fixed and variable length settings, we showed that algorithms with adaptation capability performed better than those without. This behaviour extended throughout most of our experimentation, only countered occasionally for our simplest dataset NSynth. Although not SOTA, we did observe generally strong performance from our baseline methods, with them remaining competitive over most experiments. This lower performance did however come at the benefit of being significantly faster at inference time. When performing joint training, we showed that on average the free dataset sampling outperformed the within sampling routine.Although this was the case on average, there were some interesting nuances of when each of the routines performed well. For datasets that were significantly different from one another or from the training set as a whole, we saw benefit in using within dataset sampling, possibly explained by the need for more specific or fine-grained features to solve the task at test time. For cross-dataset evaluation, we observed the opposite, where using free sampling resulted in the best overall performance. Through our k-shot test time analysis, we also find evidence of gradient-based methods being able to use additional shots more effectively, consistently rising in performance, while other methods stagnate or decline.

Acknowledgement This work is supported by the Engineering and Physical Sciences Research Council of the UK (EPSRC) Grant number EP/S000631/1 and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing.

References

1. Chen, Y., Wang, X., Liu, Z., Xu, H., Darrell, T.: A new meta-baseline for few-shot learning. CoRR **abs/2003.04390** (2020), https://arxiv.org/abs/2003.04390
2. Cheng, K.H., Chou, S.Y., Yang, Y.H.: Multi-label few-shot learning for sound event recognition. In: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP). pp. 1–5 (2019). https://doi.org/10.1109/MMSP.2019.8901732
3. Chou, S., Cheng, K., Jang, J.R., Yang, Y.: Learning to match transient sound events using attentional similarity for few-shot sound recognition. CoRR **abs/1812.01269** (2018), http://arxiv.org/abs/1812.01269
4. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. CoRR **abs/1703.03400** (2017), http://arxiv.org/abs/1703.03400
5. Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X.: FSD50K: an open dataset of human-labeled sound events. CoRR **abs/2010.00475** (2020), https://arxiv.org/abs/2010.00475
6. Gong, Y., Chung, Y., Glass, J.R.: AST: audio spectrogram transformer. CoRR **abs/2104.01778** (2021), https://arxiv.org/abs/2104.01778
7. Li, R., Bohdal, O., Mishra, R.K., Kim, H., Li, D., Lane, N.D., Hospedales, T.M.: A channel coding benchmark for meta-learning. CoRR **abs/2107.07579** (2021), https://arxiv.org/abs/2107.07579
8. Li, Z., Zhou, F., Chen, F., Li, H.: Meta-sgd: Learning to learn quickly for few shot learning. CoRR **abs/1707.09835** (2017), http://arxiv.org/abs/1707.09835
9. Ochal, M., Patacchiola, M., Storkey, A.J., Vazquez, J., Wang, S.: Few-shot learning with class imbalance. CoRR **abs/2101.02523** (2021), https://arxiv.org/abs/2101.02523
10. Park, E., Oliva, J.B.: Meta-curvature. CoRR **abs/1902.03356** (2019), http://arxiv.org/abs/1902.03356
11. Piczak, K.: Esc: Dataset for environmental sound classification. pp. 1015–1018 (10 2015). https://doi.org/10.1145/2733373.2806390
12. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al.: Improved protein structure prediction using potentials from deep learning. Nature **577**(7792), 706–710 (2020). https://doi.org/10.1038/s41586-019-1923-7
13. Shi, B., Sun, M., Puvvada, K.C., Kao, C., Matsoukas, S., Wang, C.: Few-shot acoustic event detection via meta-learning. CoRR **abs/2002.09143** (2020), https://arxiv.org/abs/2002.09143
14. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. CoRR **abs/1703.05175** (2017), http://arxiv.org/abs/1703.05175
15. Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P., Larochelle, H.: Meta-dataset: A dataset of datasets for learning to learn from few examples. CoRR **abs/1903.03096** (2019), http://arxiv.org/abs/1903.03096
16. Vinyals, O., Blundell, C., Lillicrap, T.P., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. CoRR **abs/1606.04080** (2016), http://arxiv.org/abs/1606.04080
17. Wang, Y., Chao, W., Weinberger, K.Q., van der Maaten, L.: Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. CoRR **abs/1911.04623** (2019), http://arxiv.org/abs/1911.04623
18. Wolters, P., Careaga, C., Hutchinson, B., Phillips, L.: A study of few-shot audio classification (2020)