EVALUATING BALANCED DOMAIN REGULARIZATIONS FOR MULTI-DOMAIN LEARNING IN BIRD CLASSIFICATION AUDIO TASKS

ALEXANDRE THUROW BENDER[1], ULISSES BRISOLARA CORRÊA[2], RICARDO MATSUMURA ARAUJO[3]

[1] *Universidade Federal de Pelotas - atbender@inf.ufpel.edu.br*
[2] *Universidade Federal de Pelotas - ulisses@inf.ufpel.edu.br*
[3] *Universidade Federal de Pelotas - ricardo@inf.ufpel.edu.br*

1. INTRODUCTION

Limited data quantity is already a well-established concern when training machine learning models since few examples to learn from may not be sufficient for a model to generalize well to new data (LECUN; BENGIO; HINTON, 2015). However, recent years have shifted the attention of researchers towards the importance of data quality in achieving high-performance models (SAMBASIVAN et al., 2021). Additionally, data will often be collected with specific devices (e.g. using the same camera for capturing images) or environmental conditions (e.g. recording audio clips indoors). Collections of data obtained or generated under similar conditions are referred to as domains or data sources.

Traditionally, the standard approach is to mix all training data without any particular concern for their pertaining domains. While doing this might be enough given sufficient data, significantly large datasets and the computational power to train models using them are not easily attainable. One of the reasons for this approach to be acceptable in these conditions is the high difference across examples and domains: if the data does not have a prominent domain, the model is pushed towards domain-agnostic representations. In other words, the domain-specific characteristics in data samples are diluted for not holding a common structure, and as such, they are discarded as noise.

As such, this work proposes injecting domain information by guaranteeing balanced representations of each domain in a batch, building upon the work of (BENDER, 2022). We investigate the effects of previously proposed approaches and expand them for further comparison

This research aims to understand the best way to learn from multi-domain datasets at once, dismissing the need to train multiple models for different situations. For this, we explore batch domain regularizations, which are usually overlooked in multi-domain learning.

2. METHODOLOGY

In order to evaluate the proposed multi-domain learning training methods, we need datasets that contain explicit domain characteristics. For this study, we select two bird call recording datasets, FF1010BIRD and WARBLRB10K.

Freefield (FF1010BIRD) (STOWELL; PLUMBLEY, 2013) and Warblr (WARBLRB10K) are both bird detection datasets, but they do not have any domain semantics attributed to example classes. For this reason, we use them together, each behaving as a domain. Despite both being bird presence detection datasets, they are very different. In fact, Freefield is a dataset of professional recordings of on-site observations of birds (collected from the FreeSound online database[1]). It is very

[1] https://freesound.org/

diverse in terms of location and environment. Expectedly, they use better recording equipment and usually there is not much background noise. Additionally, it has some label imbalance towards the negative class, supposedly because once the equipment is set on-site, it remains recording audio most of the time.

In contrast to FF1010BIRD, WARBLRB10K contains crowdsourced recordings of birds using the bird-watching smartphone app Warblr[2]. Its label imbalance is towards the positive class, as most users use their devices to record bird calls when in the presence of said birds. This dataset, however, has heavy background noise, including city sounds and even users imitating bird calls, allegedly to coax birds to answer. The recordings vary heavily in terms of audio quality, depending on the smartphone used.

The significant difference between the elected bird datasets is by design and desirable for this study, as domains too similar in nature would entail a difficult multi-domain analysis. The FF1010BIRD dataset contains only 25% bird presence, while the Warblr dataset contains 75% bird presence.

Stew. The more intuitive approach to using data from multiple sources at the same time is the Stew method (named due to the SpeechStew method (CHAN et al., 2021)). The method consists of simply mixing data from multi-domains together homogeneously, without any special processing or distinction. This method is already in use for various multi-domain tasks in areas such as speech recognition.

Balanced Domains. During training, the Stew approach is understood to have no balance of domains whatsoever. Essentially, the batches are expected to have more samples from the majority domains, since batches are randomly sampled from the mixed dataset. Thus the training step consists of sampling all obtainable domains, grouping the data into a single batch before presenting it to a model, and backpropagating the calculated loss from the batch.

Loss Sum. One way of penalizing the model whenever it underperforms on a training domain is achieved by balancing the domain representation. Another way to implement this is, instead of mixing domain samples into a single balanced batch, to calculate the Loss from each domain separately and sum the Loss across all domains before backpropagating it. Notably, this sort of approach was previously proposed by Bender (2022) and Tetteh et al. (2021) for image classification.

Random Sum. Admittedly, the Loss Sum method operates on a different scale than other regular methods, and this is due to the fact it sums up the loss of multiple domains. Previous studies in image classification have suggested an increase in F1-Score when training neural networks using the Loss Sum approach. However, it is unclear whether this is due to the separate loss calculation and sum operation or due to simply having a higher loss value. For this reason, we devise a counterfactual method that shares the same scale (higher loss) as the Loss Sum method but does not apply the loss function to domains properly separated, but does so in mini-batches containing examples sampled randomly from the entire dataset. Thus, we refer to this method as Random Sum.

Loss Mean. The Loss Mean method is another counterfactual method, complementing the Random Sum. It is also missing from previous studies with similar batch regularization proposals. In this approach, we perform the same procedure described for the Loss Sum process, but we divide it by the total of domains present in the task before backpropagating the loss. By doing this, we force the loss scale

[2]https://www.warblr.co.uk/

back to being comparable to other regular methods.

2.1 COMPARISON METRICS

The baseline for our comparison is the Stew training method, as it is commonly used and *de facto* standard in the literature. The performance of models in an experiment for each method (Stew, Balanced Domains, Loss Sum, Random Sum, and Loss Mean) is calculated using the average F1-Score across domains. The F1-Score was chosen because it summarizes the learning objective: learning all domains at the same time while generalizing the classes. It does so by calculating the harmonic average between precision and recall. The recall metric denotes how many positive samples were identified from to the actual positive example amount (i.e. if we understand our capacity of identifying positive samples as a fishing net, it is the ratio of how many fish we manage to catch from the total fish present in the pond). Whereas the precision metric depicts how many positive instances were correctly classified (i.e. using the same previous analogy, it is the ratio of how many of the things caught in our net are actually fish).

2.2 EXPERIMENTAL SETUP

Each audio clip is processed to handle trailing silence at the beginning and the end. Audio clips are then converted from waveform to mel-spectrograms. This representation visually represents the signal amplitude across different frequencies over time. Ultimately, spectrograms can be understood as the application of Fourier transforms on overlapping windowed segments of the signal. The mel scale is a unit of pitch to approximate the human perceived frequencies. The use of mel-spectrograms is common in audio processing because humans do not perceive frequencies on a linear scale.

The experiments use the ResNet-18, pre-trained with imageNet and fine-tuned using the bird datasets. We use 10 epochs with a batch size of 256 and an 80/20 split for train/test. Each experiment is repeatedly performed with 30 repetitions with different seeds.

3. RESULTS

The experiment uses the entirety of the bird detection datasets (which includes FF1010BIRD and WARBLRB10K as domains). The results can be viewed in Table 1.

Tabela 1: Experiment 1 — Bird Detection, Original Dataset Size, Micro F1-Score

Domain	Stew	Balanced	Loss Sum	Random Sum	Loss Mean
WARBLRB10K	0.623	0.694	0.803	0.618	0.799
FF1010BIRD	0.631	0.574	0.779	0.627	0.779
Average	0.627	0.634	0.791	0.622	0.789

When looking at the average score of each method in Table 1, we notice the best result is from using the Loss Sum approach. Additionally, the Random Sum results are far worse than Loss Sum, despite being in the same loss scale. This is

evidence against the argument stating that Loss Sum is better because of its higher loss scale. In fact, even the baseline Stew approach performed better than Random Sum. Furthermore, the Loss Mean method performs similarly to Loss Sum, despite not operating in the higher loss scale. This is yet another argument against the higher loss scale being responsible for the Loss Sum improved performance. Notably, the Balanced method also improved the average performance when compared to the Stew baseline. However, there is a tradeoff where the war domain increased in performance at the cost of lower performance in the FF1010BIRD domain.

4. CONCLUSION

This work presented an evaluation of multi-domain learning training approaches to regulating domain presence in batches when addressing audio classification tasks. When handling domains with different class distributions, Balanced domains, and Loss Sum seemed to mitigate model domain favoritism. Loss Sum consistently presented competitive results in most experiments, improving baseline results in most scenarios.

The results suggest that using explicit domain information by presenting them separately in individual batches for each domain potentially benefits the learning when training models in multi-domain tasks. Overall, multi-domain learning techniques using individual domain loss calculation, such as Loss Sum, provide an interesting strategy when dealing with multiple domains.

Future studies could also look at how different domain sizes can impact the learning of the model using the proposed methods. Additionally, it may be interesting to explore different techniques of combining the loss from different domains. Finally, exploring the interaction of different loss functions with the Loss Sum approach.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

5. REFERÊNCIAS

BENDER, I. B. **Evaluating Machine Learning Methodologies for Multi-Domain Learning in Image Classification**. 51 p. Dissertação (Master's Thesis (Computer Science)) — Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2022.

CHAN, W.; PARK, D.; LEE, C.; ZHANG, Y.; LE, Q.; NOROUZI, M. Speechstew: Simply mix all available speech recognition data to train one large neural network. **arXiv preprint arXiv:2104.02133**, 2021.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

SAMBASIVAN, N.; KAPANIA, S.; HIGHFILL, H.; AKRONG, D.; PARITOSH, P.; AROYO, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In: **proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**. Okohama, Japan: ACM, 2021. p. 1–15.

STOWELL, D.; PLUMBLEY, M. D. An open dataset for research on audio field recording archives: freefield1010. **arXiv preprint arXiv:1309.5275**, 2013.

TETTEH, E.; VIVIANO, J. D.; KRUEGE, D.; BENGIO, Y.; COHEN, J. P. Multi-domain balanced sampling improves out-of-distribution generalization of chest x-ray pathology prediction models. **Medical Imaging meets NeurIPS**, 2021.