An Introduction to Unsupervised Domain Adaptation in Sound and Music Processing

Franca Bittner, Jakob Abeßer

Fraunhofer Institute for Digital Media Technology IDMT, 98693 Ilmenau, Germany, Email: franca.bittner@idmt.fraunhofer.de

Abstract

Common machine learning models require large amounts of training data with samples representing the intended application scenario. However, these models often do not generalize well to novel data distributions caused by variations of the expected conditions. Such a lack of robustness can lead to a significant decrease in the model performance. This issue is known as domain shift and can be caused in the case of audio data by deviations of microphone characteristics or acoustic environments between data from the source domain (training data) and target domain (test data). Unsupervised domain adaptation (UDA) aims to restore the model performance by transferring knowledge from labeled samples of the source domain to unlabeled samples of a related target domain. We first provide an overview over basics and general approaches of UDA. Then, we study UDA for two audio analysis tasks: sound event detection (SED) and automatic music transcription (AMT) of piano music. Our results show that domain shift caused by microphone mismatch has a greater impact on the model performance for SED than AMT. As a possible cause we suspect that while SED analyzes the full spectral envelope, AMT examines only the harmonic peaks whose positions are less affected by domain shift.

Introduction

Over the last years, the focus of research in audio processing shifted from traditional signal processing methods to Machine Learning (ML) techniques. While ML models have shown state-of-the-art results on various benchmark datasets, they require large amounts of training data. In addition, trained models often do not generalize well to unknown data distributions, which do not match the distribution of the training data. This problem can be mitigated by enhancing the available amount and variability of the training data using data augmentation techniques or by applying Domain Adaptation (DA).

A *domain* denotes a sample space, which is defined by common sampling conditions. The sample space of the training data is referred to as *source domain*, whereas the origin of application data is called *target domain*. If the distributions in both domains diverge from each other, the performance of ML models usually decreases. This issue is known as *domain shift*. In the field of audio processing, domain shift can be caused by mismatching microphones, for example. DA methods usually align the data distributions in both domains such that ML models trained on source domain data perform better on target domain data. Particularly in Unsupervised Domain Adaptation (UDA) algorithms, domain shift is reduced by using knowledge from both labeled samples of the source domain as well as unlabeled samples of the target domain. This approach is suitable for most real-world scenarios, where data can be easily recorded but its annotation is expensive or not feasible at all.

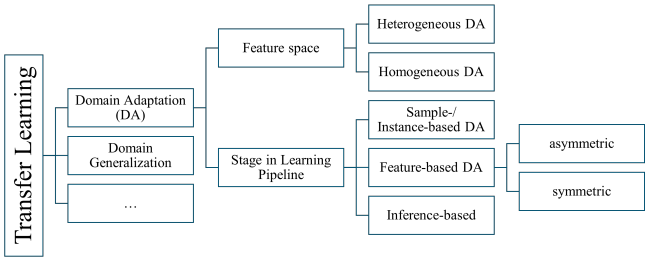

Figure 1: Taxonomy of Domain Adaptation (DA) methods based on [1, 2, 3, 4].

Domain Adaptation Taxonomy

In Figure 1, we illustrate a taxonomy of DA methods based on several state-of-the-art reports [1, 2, 3, 4]. DA can be considered as a type of Transfer Learning (TL). While *domain generalization* strives for a good performance across all domains, DA adapts models towards one particular target domain. In the taxonomy, two criteria are used to classify DA approaches—the feature space and the learning pipeline step they are applied in.

Criterion 1: Feature Space

Raw measurement data is usually converted into a suitable feature representation before being processed by ML models. The *feature space* denotes the common space for input features of the same type and dimensionality. The *feature type* can vary in style (e.g., sketch, photography, painting, etc.), modality (e.g., text, image, audio, etc.), and sensory range (e.g., ultrasonic range, human audible range, etc.). Furthermore, DA algorithms can align data distributions from different feature spaces (*heterogeneous* DA) or within the same feature spaces (*homogeneous* DA) [1, 3, 4]. Homogeneous DA has been the main focus of research. One example use-case of heterogeneous DA is sketch-to-photo retrieval [5], where a query sketch is used to find matching photos.

Criterion 2: Stage in Learning Pipeline

As a second criterion, DA algorithms can be categorized based on the stage of the learning pipeline in which they are applied. The simplified ML pipeline consists of three steps: data collection and dataset split, feature extraction, and model training. DA can be applied in any of these three stages.

Sample-based & Instance-based DA
In the data collection and sampling stage, DA can be implemented by re-weighting the influence of source domain samples according to their relevance in the target domain. Such DA methods are called *sample-based* [2] or *instance-based* [1, 3, 4] because data distributions are modified by operations on a sample level [6, 7].

Feature-based DA
DA can also be integrated into the feature extraction process. *Feature-based* DA [1, 2, 3] modifies the extracted features and transforms them to a common subspace [8, 9]. If both source domain and target domain features are changed, the process is referred to as *symmetric feature-based* DA. In contrast, *asymmetric* DA modifies features of only one of both domains [3].

Inference-based DA
The last stage of the ML pipeline is the model training. DA can be integrated into the process of model parameter optimization in various ways (*inference-based* DA [2]). Three approaches we want to highlight are Domain-Adversarial Neural Networks (DANNs) [7, 11, 12], encoder-decoder networks [13, 14], and loss functions with the objective to minimize domain shift [15].

UDA for Sound and Music Analysis

In this section, we introduce two exemplary use-cases of applying UDA for sound and music analysis, present experiments to evaluate its effectiveness, and compare the results to highlight the main differences between the two domains.

Sound Event Detection

As a first use-case for Sound Event Detection (SED), we investigate an industrial sound analysis task. The goal is to identify one out of three materials used for coating small metal balls based on audio recordings of these balls rolling down a metal slide. The dataset IDMT-ISA-METAL-BALLS [16] includes audio recordings of sliding metal balls in different recording setups. The angle of the slide and microphone positions were changed in four variation datasets, which are provided along with a training and test dataset. Therefore, we expect the microphone mismatch to cause a noticeable domain shift. The examined DA methods are feature-based and consist of various normalization techniques as discussed in [18]. We expand the experiments presented in [18] by adding results without normalization as well as individual performance scores for each variation dataset.

Automatic Music Transcription

In a second use-case, an Automatic Music Transcription (AMT) algorithm is trained to transcribe piano recordings and extract the fundamental frequency values of all musical note events. In [19], we evaluated a Multi- Pitch Estimation (MPE) model for polyphonic piano transcription, which was developed for integration into mobile devices. The cross-domain dataset IDMT-PIANO-MM [17] is used to evaluate the performance across several mobile devices. Within the scope of this work, the assessed DA methods are limited to the normalization methods used in [18]. Further results based on other DA methods are discussed in [19].

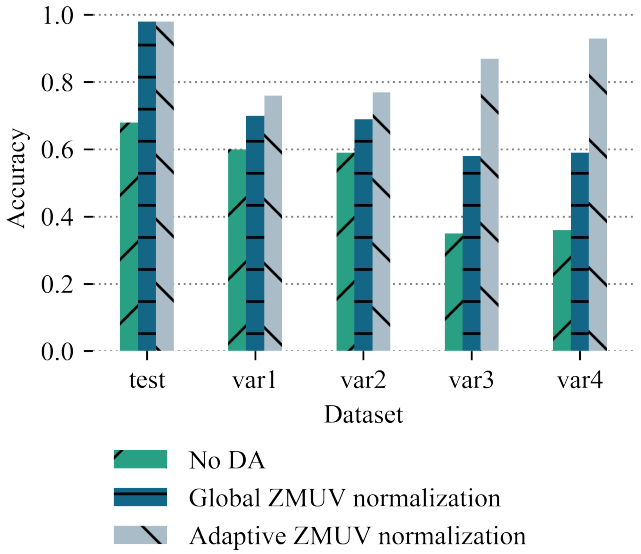

Figure 2: Accuracy scores for the Sound Event Detection (SED) model from [18] across the test (source domain) and variation datasets (target domain) of the IDMT-ISA-METAL-BALLS dataset [16].

Comparison of Results

The results of the SED task are depicted in Figure 2 and are based on the experiments in [18]. Without normalization, the accuracy score of the SED model is in the low range between 0.35 and 0.68 across all test sets, but in particular worse for variation dataset 3 and 4. The best accuracy score of 0.68 is achieved on the test set (source domain). These results confirm that in particular for the variation sets, a strong domain shift exists, which causes the model performance to decrease. We first perform a *global normalization* to zero-mean and unit variance (ZMUV) in the source domain, which results in an increase in accuracy of up to 0.98. When applying the source domain statistics to normalize each of the variation datasets, the accuracy similarly increases to values of at least 0.58 for the "var3" set or higher values for the other sets. However, if this normalization is applied individually for each dataset using only its own statistics (*adaptive normalization*), we observe the best performance across all variation datasets. For the test dataset, both global and adaptive normalization lead to similar results. As a conclusion, adaptive ZMUV normalization allows to reduce the impact of domain shift and the model performance can be mostly restored across all variation datasets.

Figure 3 shows the results of the AMT task from [19]. In contrast to the SED use-case, the model performance remains at a similar level across all domains (i.e., recording devices) and DA methods. In contrast to the SED task, neither global nor adaptive ZMUV normalization improve the average performance notably.

Conclusion

Based on our findings, we hypothesize that the effectiveness of UDA is task-dependent. For the task of SED, the sound characteristics, which are important for the clas-

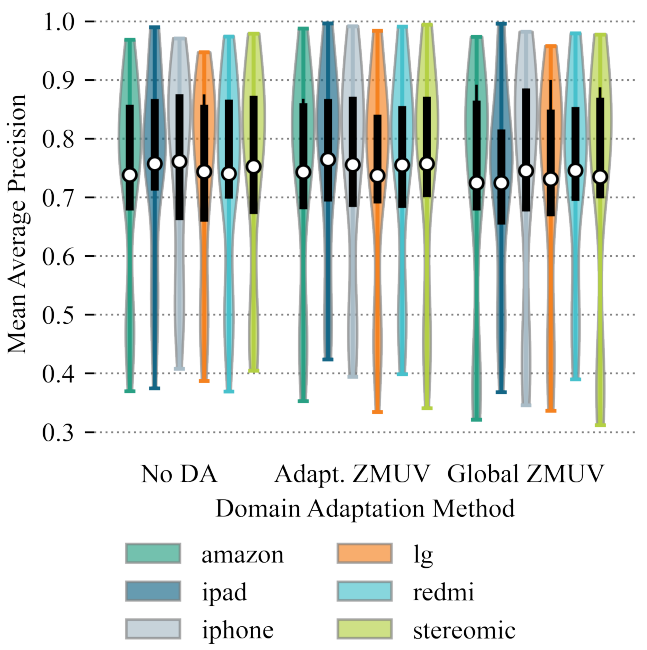

Figure 3: Violin plots of the mean average precision score for the Multi- Pitch Estimation (MPE) model of [19] across six different recording devices from the IDMT-PIANO-MM dataset [17]. Black lines indicate the first and third quartiles, white circles indicate the mean performance score.

sification task, depend on the spectral energy distribution within a larger frequency range. Here, microphone mismatch can have a strong influence of this distribution and the model decisions. For AMT, mostly the frequency location of spectral peaks are important for the pitch estimation task. A microphone mismatch alters the peak magnitudes but not their frequency locations, which might explains the higher robustness of MPE methods against domain shift.

Acknowledgments

This study was supported by the German Research Foundation (AB 675/2-2).

References

[1] Wang, M. & Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing 312 (2018), 135–153

[2] Kouw, W. M. & Loog, M.: A Review of Domain Adaptation without Target Labels. IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2021), 766–785

[3] Zhuang, F. et al.: A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE 109 (2021), 43–76

[4] Pan, S. J. & Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22 (2010), 1345–1359

[5] Yang, F. et al.: Instance-Level Heterogeneous Domain Adaptation for Limited-Labeled Sketch-to-Photo Retrieval. IEEE Transactions on Multimedia 23 (2021), 2347–2360

[6] Xiao, N. & Zhang, L.: Dynamic Weighted Learning for Unsupervised Domain Adaptation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), 15237–15246

[7] Zhang, J., Ding, Z., Li, W. & Ogunbona, P.: Importance Weighted Adversarial Nets for Partial Domain Adaptation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018), 8156–8164

[8] Mezza, A. I., Habets, E. A. P., Müller, M. & Sarti, A.: Unsupervised Domain Adaptation for Acoustic Scene Classification Using Band-Wise Statistics Matching. Procedings of the 28th European Signal Processing Conference (EUSIPCO) (2021), 11–15

[9] Mezza, A. I., Habets, E. A. P., Müller, M. & Sarti, A.: Feature Projection-Based Unsupervised Domain Adaptation for Acoustic Scene Classification. Proceedings of the IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP) (2020), 1–6

[10] Sun, B., Feng, J. & Saenko, K.: Return of Frustratingly Easy Domain Adaptation. Proceedings of the AAAI Conference on Artificial Intelligence 30 (2016), 2058–2065

[11] Ganin, Y. et al.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research (JMLR) 17 (2016), 2096–2030

[12] Gu, X., Sun, J. & Xu, Z.: Unsupervised and Semi-supervised Robust Spherical Space Domain Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022), 1-17

[13] Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D. & Li, W.: Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation. European Conference on Computer Vision (ECCV) (2016), 597–613

[14] Deng, J., Zhang, Z., Eyben, F. & Schuller, B.: Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition. IEEE Signal Processing Letters 21 (2014), 1068–1072

[15] Sun, B. & Saenko, K.: Deep CORAL: Correlation Alignment for Deep Domain Adaptation. Computer Vision – ECCV 2016 Workshops vol. 9915. Springer, Cham, 2016, 443–450

[16] Grollmisch, S., Abeßer, J., Liebetrau, J. & Lukashevich, H.: Sounding Industry: Challenges and Datasets for Industrial Sound Analysis. 27th European Signal Processing Conference (EUSIPCO) (2019)

[17] Abeßer, J., Bittner, F., Richter, M., Gonzalez, M. & Lukashevich, H.: A Benchmark Dataset to Study Microphone Mismatch Conditions for Piano Multipitch Estimation on Mobile Devices. Proceedings of the Digital Music Research Network One-day Workshop (DMRN+16) (2021), 22

[18] Johnson, D. S. & Grollmisch, S.: Techniques Improving the Robustness of Deep Learning Models for Industrial Sound Analysis. Proceedings of the 28th European Signal Processing Conference (EUSIPCO) (2021), 81–85

[19] Bittner, F., Gonzalez, M., Richter, M., Lukashevich, H. & Abeßer, J.: Multi-pitch Estimation meets Microphone Mismatch: Applicability of Domain Adaptation. Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR) (2022), 477–484