Task Arithmetic for Audio classification

August 5, 2024

Chiara Ballanti

Abstract

Task arithmetic has recently emerged as a cost-effective and scalable approach for editing pre-trained models directly in weight space. By leveraging algebraic operations on model parameters, this method allows for targeted modifications to adapt models to new tasks or domains without the need for extensive retraining. The effectiveness of this approach was thoroughly explored in vision and NLP tasks by (Ilharco et al., 2022), who demonstrated its efficacy across various applications. Inspired by these promising results, I propose to investigate the effectiveness of task arithmetic within the context of audio use cases. The code is available on GitHub.

1. Introduction

Foundation models have revolutionized machine learning with their remarkable ability to perform well across a wide range of tasks. However, their generalized nature often necessitates fine-tuning to reach optimal performance for specific applications. Traditionally, this fine-tuning process has been complex, time-consuming, costly, and resource-intensive. Task arithmetic has emerged as a promising solution, providing a more efficient approach by directly manipulating model parameters through algebraic operations. This project explores the potential of task arithmetic for audio classification problems.

2. Related work

A task vector (Ilharco et al., 2022) can be viewed as a set of weight adjustments specifically calibrated for a given task through fine-tuning, obtained by subtracting the task-specific weights from the original pre-trained weights. Different task vectors derived from the same pre-trained models can then be adjusted and combined through simple arithmetic operations such as addition and subtraction to achieve multi-task learning (Zhang et al., 2023) and task forgetting (Daheim et al., 2023).

Recently, task vectors have shown promise in vision (Hojel et al., 2024; Pham et al., 2024; Yang et al., 2023; Jin et al., 2024; Ortiz-Jimenez et al., 2024) and NLP (Huang et al., 2023; Bhardwaj et al., 2024), but their application in audio use cases remains relatively underexplored. Only a few papers examined the application of task vectors to ASR problems. For instance, (Ramesh et al., 2024) showed that task vectors are effective for multi-tasking and proposed novel approaches for zero-resource domain expansion and improving low-resource ASR by simulating high-resource data. In contrast, (Su et al., 2024) introduced *SYN2REAL* task vector for synthetic-to-real adaptation in ASR, focusing on using task vectors to mitigate the distributional shift between real and synthetic data. In another study, (Kalyan et al.) introduced a novel method to synthesize emotional speech by manipulating emotion vectors.

3. Method

Let $\theta_{pt} \in \mathbb{R}^d$ be the weights of a pre-trained model and $\theta_{ft} \in \mathbb{R}^d$ the corresponding weights after fine-tuning on a specific task. A **task vector** (TV) $\tau \in \mathbb{R}^d$ specifies a direction in the weight space of a pre-trained model, such that movement in that direction improves performance on the task. τ is a vector obtained by taking the element-wise difference between θ_{ft} and θ_{pt}: $\tau = \theta_{ft} - \theta_{pt}$.

Task vectors can be applied to any model parameters θ from the same architecture, via element-wise addition, with an optional scaling term λ, such that the resulting model has weights $\theta_{new} = \theta + \lambda\tau$.

Task vector operations provide a powerful and efficient method to adapt and refine models for various applications. **Forgetting via negation** involves negating a task vector to decrease performance on the target task, with little change in model behavior on control tasks. Then, $\tau_{new} = -\tau_{target}$. **Learning by addition** enhances performance across multiple tasks by adding task vectors together. Then, $\tau_{new} = \tau_{task_1} + \tau_{task_2}$. **Task analogy** involves using an analogy relationship of the form "*A* is to *B* as *C* is to *D*". By combining task vectors from three of the tasks, it is possible to improve the performance on the

Email: <ballanti.1844613@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

fourth one, even without training data from that task. Then, $\tau_{new} = \tau_{task_B} + \tau_{task_C} - \tau_{task_A}$.

4. Experiments setup

Datasets Four audio classification datasets were used. **ESC-50** (Piczak, 2015) consists of 2k environmental sound recordings across 50 classes. **UrbanSound8K** (Salamon et al., 2014) contains 9k urban sound recordings divided into 10 classes. **GTZAN** (Tzanetakis et al., 2001) includes 1k audio clips categorized into 10 music genres. The **Huan0806/gender_emotion_recognition** dataset (Huan0806) from *Hugging Face* features 12k audio speech divided into 12 classes based on emotion and gender of the speaker. It is a combination of **RAVDESS** (Livingstone & Russo, 2018), **Crema-D** (Cao et al., 2014), **TESS** (Pichora-Fuller & Dupuis, 2020), **Savee** (Jackson & Haq, 2014) and **EmoBD** (Burkhardt et al., 2005) datasets.

Model Experiments were conducted using **CLAP** (Wu et al., 2023; Elizalde et al., 2023), an open-vocabulary audio classifier. This open-ended model allows fine-tuning on downstream tasks without adding new parameters. Furthermore, the weights of CLAP's text encoder were frozen.

Analogy fine-tuning Once the target class for the analogy is determined, three different models are fine-tuned, one for each of the remaining categories. Each model is fine-tuned and evaluated on the 50 classes of the *ESC-50* dataset, along with a new class called "something". This approach simulates the learning of a new class that the pre-trained classifier has never encountered before.

5. Experimental results

Table 1. Forgetting audio classification tasks via negation. Performance (accuracy) comparison.

Task	Pre-trained		Negated TVs	
	control	target	control (↑)	target (↓)
GTZAN	80.50%	32.00%	76.50%	10.00%
US8k	80.50%	71.76%	70.50%	50.76%

Forgetting via negation *ESC-50* was used as the control task because it is a dataset on which the pre-trained *CLAP* model performs quite well. Experiments were conducted on both the *GTZAN* and *UrbanSound8k* datasets as target tasks (Table 1). The results on *GTZAN* are promising. The accuracy on the target task decreased significantly to 10%, while the accuracy on the control task experienced only a minor reduction of 3%, remaining largely stable. In contrast, the results on *UrbanSound8k* are considerably worse. Although the control and target datasets differ in scope and focus, there is some overlap in the classes related to common urban and animal sounds. Decreases in the accuracy of the target dataset lead to notable, even if not dramatic, reductions in the accuracy of the control dataset.

Learning by addition Experiments were conducted on both *GTZAN* and *UrbanSound8k* as target tasks (Table 2). Due to suboptimal performance with a single scaling coefficient, two different scaling coefficients were used, one for each task vector. Scaling each task vector by its coefficient improved the performance of the resulting model. To handle the difference in difficulty levels of the tasks, the accuracy of each task has been normalized with respect to the accuracy of the model fine-tuned on that specific task.

Table 2. Build multi-task audio classifier via addition. Performance (accuracy) comparison.

Task	Pre-trained	Fine-tuned	Added TV	Normalized added TV
GTZAN	32.00%	90.00%	83.00%	92.22%
US8k	71.76%	95.04%	91.22%	95.62%

Task analogy Experiments were conducted on both the *ESC-50* and *gender_emotion_recognition* datasets (Table 3). By defining the analogy "female_happy(A) : male_happy(B) = female_sad(C) : male_sad(D)", a new classifier was built. This classifier performs well on a new category (e.g. "male_sad") leveraging data from three related classes that form an analogy relationship (e.g., "female_happy", "man_happy", and "female_sad"). The results are impressive, with the accuracy on the new task surpassing the individual accuracies of the other three tasks in the analogy. Moreover, the final model outperforms the pre-trained one by 7% on *ESC50*, which was used as a support dataset for fine-tuning.

Table 3. Learning via analogy. Performance (acc) comparison.

Pre-trained	Fine-tuned			Analogy TV
D	A	B	C	D
35.29%	93.63%	94.12%	92.65%	95.10%

6. Conclusions and future work

This work explored task arithmetic for audio classification tasks for the first time, confirming this approach as an effective and efficient way to edit models. This method enables precise adjustment of model performance on one or more specific tasks, either enhancing or reducing performance as needed, while minimizing computational costs and eliminating retraining time. Future work will involve introducing new parameters to explore the architecture-agnostic potential of this approach and expand the method's application to other unexplored audio domains.

References

Bhardwaj, R., Anh, D. D., and Poria, S. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. *arXiv preprint arXiv:2402.11746*, 2024.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. A database of german emotional speech. In *Interspeech*, volume 5, pp. 1517–1520, 2005.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Daheim, N., Dziri, N., Sachan, M., Gurevych, I., and Ponti, E. M. Elastic weight removal for faithful and abstractive dialogue generation. *arXiv preprint arXiv:2303.17574*, 2023.

Elizalde, B., Deshmukh, S., and Wang, H. Natural language supervision for general-purpose audio representations, 2023. URL https://arxiv.org/abs/2309.05767.

Hojel, A., Bai, Y., Darrell, T., Globerson, A., and Bar, A. Finding visual task vectors. *arXiv preprint arXiv:2404.05729*, 2024.

Huan0806. Gender emotion recognition dataset. https://huggingface.co/datasets/Huan0806/gender_emotion_recognition.

Huang, S.-C., Li, P.-Z., Hsu, Y.-C., Chen, K.-M., Lin, Y. T., Hsiao, S.-K., Tsai, R. T.-H., and Lee, H.-y. Chat vector: A simple approach to equip llms with new language chat capabilities. *arXiv preprint arXiv:2310.04799*, 2023.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Jackson, P. and Haq, S. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.

Jin, R., Hou, B., Xiao, J., Su, W., and Shen, L. Fine-tuning linear layers only is a simple yet effective way for task arithmetic. *arXiv preprint arXiv:2407.07089*, 2024.

Kalyan, P., Rao, P., Jyothi, P., and Bhattacharyya, P. Emotion arithmetic: Emotional speech synthesis via weight space interpolation.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018.

Ortiz-Jimenez, G., Favero, A., and Frossard, P. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

Pham, M., Marshall, K. O., Hegde, C., and Cohen, N. Robust concept erasure using task vectors. *arXiv preprint arXiv:2404.03631*, 2024.

Pichora-Fuller, M. K. and Dupuis, K. Toronto emotional speech set (TESS), 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Ramesh, G., Audhkhasi, K., and Ramabhadran, B. Task vector algebra for asr models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12256–12260. IEEE, 2024.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Su, H., Farn, H., Chen, S.-T., and Lee, H.-y. Syn2real: Leveraging task arithmetic for mitigating synthetic-real discrepancies in asr domain adaptation. *arXiv preprint arXiv:2406.02925*, 2024.

Tzanetakis, G., Essl, G., and Cook, P. Automatic musical genre classification of audio signals, 2001. URL http://ismir2001.ismir.net/pdf/tzanetakis.pdf.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yang, E., Wang, Z., Shen, L., Liu, S., Guo, G., Wang, X., and Tao, D. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*, 2023.

Zhang, J., Chen, S., Liu, J., and He, J. Composing parameter-efficient modules with arithmetic operation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=5r3e27I9Gy.