MULTI-TASK LEARNING FOR SOUND EVENT DETECTION USING VARIATIONAL AUTOENCODERS

Technical Report

Petros Giannakopoulos[*]

Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens, Greece
petrosgk@di.uoa.gr

Aggelos Pikrakis[†]

Dept. of Informatics
University of Piraeus, Greece
pikrakis@unipi.gr

ABSTRACT

This technical report presents a multi-task learning model based on recurrent variational autoencoders (VAEs). The proposed method employs recurrent VAEs with shared parameters to simultaneously learn the tasks of strong labeling, weak labeling and feature sequence reconstruction. During the training stage, the model receives as input strongly labeled, weakly labeled data and unlabeled data and it simultaneously optimizes frame-based and file-based cross-entropy losses for strongly labeled and weakly labeled data, respectively, as well as the reconstruction loss for the unlabeled data. Using a shared posterior among all task branches, the model projects the input data for each task into a common latent space. The decoding of latents sampled from this common latent space, in combination with the shared parameters among task branches act jointly as a regularizer that prevents the model from overfitting to the individual tasks. The proposed method is evaluated on the DCASE-2022 Task4 dataset on which it achieves an event-based macro F1 score of 32.5% on the validation set and 31.8% on the public evaluation set.

Index Terms— Sound event detection, multi-task learning, variational autoencoder, semi-supervised learning

1. INTRODUCTION

Sound Event Detection (SED) is the process of identifying sounds in the environment, such as a human speaking, a dog barking, a vaccum cleaner etc. [1]. Besides the understanding of the environment that it provides, it can also be used as feedback to other systems that are capable of taking actions, as it is the case with the triggering of an alarm. In recent years, neural networks have contributed to notable improvements in the performance of SED systems. Convolutional Neural Networks (CNNs) [2, 3], Recurrent Neural Networks (RNNs) [4, 5], Convolutional RNNs (CRNNs) [6, 7] and Transformers [8, 9, 10] have been used with success as the backbone of SED systems.

The main drawback of neural-network based approaches is that a large amount of labeled data is required during a supervised training stage. There are two main SED variations, i.e., *strong* and *weak* audio event tagging. In the case of *strong* tagging, an SED system must detect both the audio event type *and* the respective endpoints. In the case of *weak* tagging, the SED system must only detect the presence of the audio event. The *strong* tagging task requires audio data to be annotated with timestamps that provide the beginning and end of each audio event occurrence. This type of data, known as *strongly labeled data*, are difficult, time-consuming and costly to collect in amounts that are sufficient to effectively train neural-network based approaches via supervised learning. Emphasis has therefore been placed on developing training methods which reduce the requirements for strongly annotated data, while remaining effective. These range from simple data augmentation techniques to weakly-supervised and semi-supervised learning methods. Data augmentation has proved to be an effective technique to improve the generalization capabilities of SED models by performing random or targeted processing on existing data to artificially generate new data samples [2, 3]. Furthermore, several SED model architectures and training schemes have been proposed which can take advantage of *weakly labeled* and/or *unlabeled* data to improve generalization while reducing the requirements for *strongly labeled* data [8, 11, 12].

Multi-Task Learning (MTL) [13] is a method where a model can learn to solve multiple tasks simultaneously, while exploiting possible common characteristics and differences across them. Such a model can achieve improved performance on each individual task compared to a model that learns to solve each problem in isolation. MTL has

[*] Research in this paper was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty Members & Researchers and the Procurement of High-Cost Research Equipment Grant" (Project Number: 3449).

[†] We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

been applied to the domain of weakly-supervised and semi-supervised SED [14, 15, 16] with promising results. Previous works have combined MTL with Variational Auto-Encoders (VAEs) [17, 18, 19, 20] and showed that projecting input features for each task into latent representations sampled from the posterior of a variational encoder can improve regularization of shared features for downstream tasks and is more robust to noise and outliers in the input features.

In this work we propose an SED model based on the MTL-VAE principle and RNNs. We then simultaneously train the model on three audio event tagging tasks, each having its own dataset as provided by the DCASE-Task4 2022 challenge: strong tagging on synthetic audio data, weak tagging on real audio data and strong tagging on real audio data. The model is also simultaneously trained on a fourth task: reconstruction of unlabeled audio features. We demonstrate that the model is able to leverage cross-task information to achieve superior performance on the task of strong audio event tagging on real data, which is the task of interest, compared to the case when it is trained on this task without MTL. We also demonstrate that using a VAE architecture improves generalization performance. Our MLT-VAE SED model achieved 32.5% event-based macro F1 score on the DCASE-Task4 2022 challenge validation set and 31.8% on the public evaluation set, without using data augmentation.

2. PROPOSED METHOD

2.1. Network architecture

The proposed MTL-VAE architecture for SED is demonstrated in Figure 1. It consists of a variational encoder for each task input and all encoders share weights. The resulting outputs of the variational encoders are shared stochastic latent representations of the input features of all downstream tasks. The latent representations are inputs to decoders with shared weights, with each decoder being responsible for a respective task. Each decoder is followed by a classification head which outputs either frame-level predictions for the *strong* audio event tagging or file-level predictions for the *weak* tagging tasks. The decoder responsible for the reconstruction task is followed by a feature reconstruction head.

2.2. Training procedure

For the concurrent training on all four tasks, the final objective that the model must optimize for is the sum of four objectives, one for each task, specifically: 1) frame-level cross-entropy for the *strong synthetic audio event tagging* task, 2) file-level cross-entropy for the *weak real audio event tagging* task, 3) frame-level cross-entropy for the *strong real audio event tagging* task, and 4) reconstruction error for the *unlabeled data reconstruction* task. To that sum we must add the KL-divergence objective between the posterior of the VAE and a Gaussian prior $N(0, 1)$. Based on the above, the final objective is:

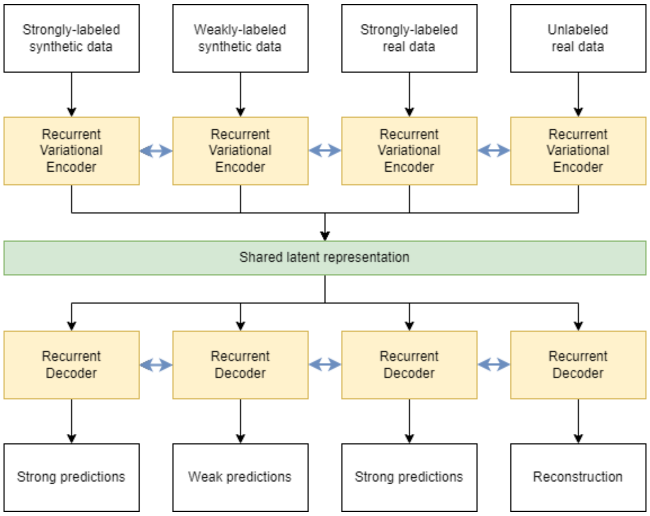

Strongly-labeled synthetic data
Weakly-labeled synthetic data
Strongly-labeled real data
Unlabeled real data
Recurrent Variational Encoder
Recurrent Variational Encoder
Recurrent Variational Encoder
Recurrent Variational Encoder
Shared latent representation
Recurrent Decoder
Recurrent Decoder
Recurrent Decoder
Recurrent Decoder
Strong predictions
Weak predictions
Strong predictions
Reconstruction

Figure 1: Overview of the architecture of our proposed MTL-VAE architecture. Input features for each task are encoded into a shared latent representation by variational encoders with shared weights. Decoders with shared weights perform the tasks of audio event prediction and audio feature reconstruction.

$$L = 2 * BCE_{strong} + BCE_{weak} + MSE + KLD \quad (1)$$

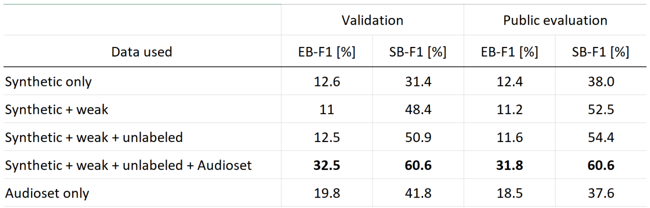
We use the Adam optimizer [21] with a learning rate of $5 * 10^{-4}$ and a batch size of 32. Each batch contains synthetic audio data with strong labels, real audio data with weak labels, real unlabeled audio data, and real audio data with strong labels from the Audioset dataset [22].

3. EXPERIMENTS

3.1. Effects of Multi-Task Learning

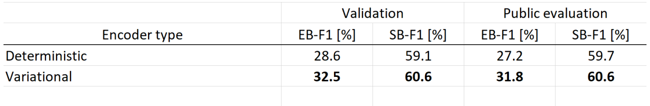
The results of our experiments are summarized in Figure 2. When using all four types of data (synthetic strongly labeled, real weakly labeled, unlabeled, real strongly labeled) to simultaneously learn four tasks (strong event tagging on synthetic audio data, weak event tagging on real audio data, reconstruction of unlabeled data and strong event tagging on real audio data) we observe an event-based macro F1-score of 32.5% on the DCASE-Task4 2022 validation set and a score of 31.8% on the public evaluation set. We also observe a segment-based macro F1-score of 60.6% on both the validation and public evaluation sets.

We conduct an ablation study to assess the impact of each additional learned task to the performance of the multi-task model. We observe that when the model is only trained for *strong event tagging* on synthetic audio data with strong



	Validation		Public evaluation	
Data used	EB-F1 [%]	SB-F1 [%]	EB-F1 [%]	SB-F1 [%]
Synthetic only	12.6	31.4	12.4	38.0
Synthetic + weak	11	48.4	11.2	52.5
Synthetic + weak + unlabeled	12.5	50.9	11.6	54.4
Synthetic + weak + unlabeled + Audioset	**32.5**	**60.6**	**31.8**	**60.6**
Audioset only	19.8	41.8	18.5	37.6

Figure 2: Event-based and segment-based macro F1-scores on the validation and public evaluation sets of our proposed SED model. The best results are obtained when the model is trained on all four tasks (synthetic audio data strong tagging, real audio data weak tagging, unlabeled audio data reconstruction, real audio data strong tagging). Each additional task improves classification performance.



	Validation		Public evaluation	
Encoder type	EB-F1 [%]	SB-F1 [%]	EB-F1 [%]	SB-F1 [%]
Deterministic	28.6	59.1	27.2	59.7
Variational	**32.5**	**60.6**	**31.8**	**60.6**

Figure 3: Performance comparison of the proposed SED model when the encoder is deterministic and when it is variational. A variational encoder seems to improve generalization ability and improves the event-based macro F1 score over using a deterministic encoder.

event labels, it has the worst scores on the classification metrics with an event-based macro F1-score of 12.6% and a segment-based macro F1-score of 31.4% on the validation set, as well as 12.4% and 38.0% respectively on the public evaluation set.

Adding the task of *weak event tagging* on real audio data with weak event labels improves the segment-based F1-score to 48.4% and 52.5% on the validation and public evaluation sets respectively, but the event-based F1 score does not improve. Further adding the task of *reconstruction* of unlabeled audio data improves the segment-based F1-score to 50.9% and 54.4% on the validation and public evaluation sets, respectively.

Finally, the addition of the task of *strong event tagging* on real audio event data with strong event labels (from the Audioset dataset) significantly improves the event-based F1-score to 32.5% on the validation set and 31.9% on the public evaluation set. The segment-based F1-score further improves to 60.6% on both sets. This most likely occurs because the real audio dataset with strong event labels and, consequently, the task of *strong event tagging* on real audio data have the closest domain proximity to the validation and public evaluation datasets, which are also real audio data with strong event labels. Therefore, it is not a surprise that this task has the largest contribution to the information extracted by the MTL model.

However, when training only on the Audioset data and learning only the task of *strong event tagging* on real audio data, the final performance is significantly lower than when training on all tasks using all types of data. The event-based F1 score drops to 19.8% and 18.5% on the validation and public evaluation sets respectively, while the segment-based F1 score becomes 41.8% and 37.6%, respectively. This underlines the effectiveness of MTL and that all data types and their respective tasks contribute to the ability to learn more robust representations that generalize better.

3.2. Contribution of VAEs

Figure 3 compares the event-based and segment-based macro F1 scores achieved on the validation and public evaluation sets by the MTL model when the encoder is deterministic and when it is variational. Using a variational encoder leads to an improvement in the event-based F1 score of approximately 4% and an improvement of 1% in the segment-based F1 score. We conclude that this is due to the better generalization ability of the variational autoencoder architecture. Introducing stochasticity into the latent representations of each encoded task data features and constraining the shared latent space to be close to a Gaussian prior leads to improved regularization of learned task data representations.

4. CONCLUSION

In this work we designed a Multi-Task Learning (MTL) model based on a recurrent autoencoder architecture with variational information bottleneck. We applied this MTL model to the challenge of learning Sound Event Detection when only a limited amount of annotated training data is available, as outlined in DCASE Task4. For each of the four types of data provided by the DCASE Task4 dataset, we assigned a task to be learned: *strong audio event tagging* from the synthetic audio data with strong event labels, *weak audio event tagging* from the real audio data with weak event labels, *reconstruction* of real unlabeled data from the provided real audio data without annotations, and *strong audio event tagging* from the real audio data with strong event labels. The model is trained simultaneously on all tasks and has the ability to exploit cross-task information through parameter (weight) sharing between the autoencoders appointed to each task and through projecting the encoded features for each task data into a shared latent space. We then demonstrate that this MTL scheme significantly improves the model's classification accuracy, as measured by the event-based and segment-based macro F1 scores, in the validation and public evaluation datasets of DCASE Task4, with each additional learned task contributing to improving the model's final performance. We also found that introducing stochasticity into the shared latent representations, by using variational instead of deterministic encoders further improves classification performance through better cross-task generalization, since the stochasticity introduced into latent representations acts as a regularizer.

5. REFERENCES

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[2] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *arXiv preprint arXiv:1604.07160*, 2016.

[3] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[4] T. H. Vu and J.-C. Wang, "Acoustic scene and event recognition using recurrent neural networks," *Detection and Classification of Acoustic Scenes and Events*, vol. 2016, pp. 1–3, 2016.

[5] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Duration-controlled lstm for polyphonic sound event detection," *IEEE/ACM Transactions on ASLP*, vol. 25, no. 11, pp. 2059–2070, 2017.

[6] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on ASLP*, vol. 25, no. 6, pp. 1291–1303, 2017.

[7] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *2017 IEEE ICASSP*. IEEE, 2017, pp. 771–775.

[8] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Weakly-supervised sound event detection with self-attention," in *2020 IEEE ICASSP*. IEEE, 2020, pp. 66–70.

[9] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE ASRU*. IEEE, 2019, pp. 449–456.

[10] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[11] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *2017 ICASSP*. IEEE, 2017, pp. 791–795.

[12] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," *Orange Labs Lannion, France, Tech. Rep*, 2019.

[13] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[14] S. Deshmukh, B. Raj, and R. Singh, "Multi-task learning for interpretable weakly labelled sound event detection," *arXiv preprint arXiv:2008.07085*, 2020.

[15] H. Liang, W. Ji, R. Wang, Y. Ma, J. Chen, and M. Chen, "A scene-dependent sound event detection approach using multi-task learning," *IEEE Sensors Journal*, 2021.

[16] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multitask learning of sound events and scenes with soft scene labels," in *2020 IEEE ICASSP*. IEEE, 2020, pp. 621–625.

[17] W. Qian, B. Chen, Y. Zhang, G. Wen, and F. Gechter, "Multi-task variational information bottleneck," *arXiv preprint arXiv:2007.00339*, 2020.

[18] G. Lu, X. Zhao, J. Yin, W. Yang, and B. Li, "Multi-task learning using variational auto-encoder for sentiment classification," *Pattern Recognition Letters*, vol. 132, pp. 115–122, 2020.

[19] T.-H. Vo, G.-S. Lee, H.-J. Yang, S.-R. Kang, I.-J. Oh, and S.-H. Kim, "Multi-task with variational autoencoder for lung cancer prognosis on clinical data," in *The 9th International Conference on Smart Media and Applications*, 2020, pp. 234–237.

[20] J. Shen, X. Zhen, M. Worring, and L. Shao, "Variational multi-task learning with gumbel-softmax priors," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 031–21 042, 2021.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE ICASSP*, New Orleans, LA, 2017.