Computer Science and Application 计算机科学与应用, 2022, 12(10), 2268-2274
Published Online October 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1210231

Hans汉斯

针对音频源分离平台的鲁棒性提升

李明圆

天津工业大学，天津

收稿日期：2022年9月8日；录用日期：2022年10月7日；发布日期：2022年10月17日

摘　要

尽管基于神经网络的音频源分离方法具有优异的性能和广泛的应用范围，但其对故意攻击的鲁棒性在很大程度上被忽视了。本文在音频源分离平台rl_singing_voice-master的基础上提出了一种新的分离平台结构，该分离平台引入了自注意力机制(self-attention)并使用变分丢弃法(Variational Dropout)对其进行正则化处理。实验结果表明，在MUSDN18数据集上，改进后的音频源分离平台相较于原分离平台，在面对对抗性样本的故意攻击时，鲁棒性也得到了明显提升，分离性能也得到了优化。

关键词

音频源分离，鲁棒性，自注意力机制

Robustness Improvement for Audio Source Separation Platform

Mingyuan Li

Tiangong University, Tianjin

Received: Sep. 8th, 2022; accepted: Oct. 7th, 2022; published: Oct. 17th, 2022

Abstract

Although the neural network based audio source separation method has excellent performance and a wide range of applications, its robustness to intentional attacks has been largely ignored. In this paper, the audio source separation platform rl_singing_voice-master. On the basis of voice master, a new separation platform structure is proposed, which introduces self attention mechanism and regularizes it using variational drop. The experimental results show that compared with the original separation platform, the improved audio source separation platform on the MUSDN18

文章引用: 李明圆. 针对音频源分离平台的鲁棒性提升[J]. 计算机科学与应用, 2022, 12(10): 2268-2274.
DOI: 10.12677/csa.2022.1210231

dataset has significantly improved robustness and separation performance when facing intentional attacks on adversarial samples.

Keywords

Audio Source Separation, Robustness, Self-Attention Mechanism

1. 引言

音频源分离在下游任务中得到了广泛的研究和应用。例如，各种音乐信息检索任务，包括歌词识别与对齐[1] [2] [3]、音乐转录[4] [5]、乐器分类[6]、歌音生成[7]，都依赖于音频源分离(MSS)。同样，自动语音识别也得益于语音增强和语音分离。现如今提出的各种源分离方法的最新进展极大地提高了分离精度，一些方法的性能与理想掩码方法相当，甚至优于理想掩码方法，并将其作为理论上基线[8]-[13]。除了出色的性能，深度神经网络最近被证明易受一种特定类型的攻击，最常被称为对抗性攻击。这些方法主要包括 PGD [14]、Deep Fool [15]、FGSM [16]及 JSMA [17]等。他们只对输入数据进行细微的更改，但导致网络性能出现重大故障，然而这些更改几乎不会被人类观察者注意到。此外，在[18]中已经表明，这种对抗性攻击倾向于很好地泛化跨模型。这种可移植性特性只会增加攻击的可能性，因为攻击者可能不需要知道特定被攻击网络的结构就可以欺骗它。

因而在音源分离领域，针对故意攻击的鲁棒性研究非常重要。首先，如果有人以感知不到的方式恶意操纵音频，从而使分离质量严重下降，那么所有下游任务都有可能失败；其次，如果音频的创作者们不希望他们的音频内容被分离和重用，那么这样的操作可以保护内容不被分离，对原始内容造成最小和难以察觉的干扰。前者被视为针对分离平台攻击的一种防御，后者被视为针对分离信号滥用的内容版权保护。但是，源分离模型在对抗攻击下的鲁棒性在很大程度上被忽视了。

本文对音频源分离平台的鲁棒性进行了研究。对音频源分离平台 rl_singing_voice-master 中编码器的结构做出改进，在编码器中添加自注意力机制(Self-Attention)并使用变分丢弃法(Variational Dropout)进行正则化处理防止过拟合。实验证明，在相同对抗性样本的攻击下，改进后的分离平台鲁棒性明显优于原分离平台。

2. 相关知识

2.1. 自注意力机制(self-attention)

在注意力机制的基础上减少了对外部信息的依赖，使得设计的模型能够更好地关注特征之间的相关性或数据相关性。通过人工神经网络获得的输出特征本质上是由卷积层通过卷积核和原始特征的线性组合获得的。因此通常使用叠加卷积层的方法使获得的效果更加优化。事实上，这种方法效果并不突出。语义信息不足是导致模型应用于计算机视觉领域中性能不佳的一大原因，而自注意力机制则是拥有更大的感受野来对全局信息进行捕获，包含了更多的上下文信息。上下文信息的作用在许多视觉任务中显得格外重要，例如目标检测和语义分割等。自注意力机制为此提供了一种有效的建模方法，通过 q、k 和 v 的三元组来对全局上下文信息有效捕获，如图 1 所示。

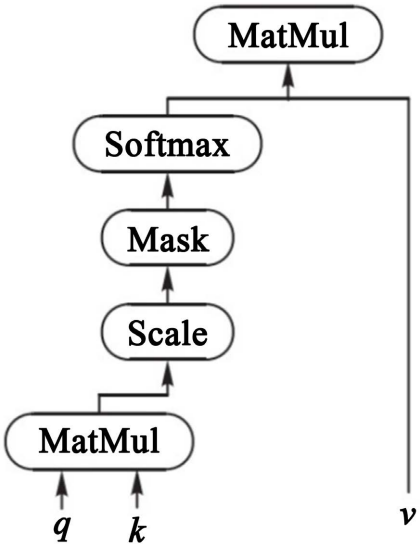

Figure 1. Schematic diagram of self-attention mechanism
图 1. 自注意力机制原理图

2.2. 变分丢弃法(Variational Dropout)

泛化问题向来是深度学习模型的关键问题。因为神经网络拥有着超强的拟合能力，所以通常会以很高的效率降低训练集的错误率，从而引发过拟合。那么就可以想到，在对深度神经网络的训练过程中，我们可以适当的对一些神经元(及对应的连接边)进行随机丢弃操作，以防止过度拟合。这种方法称为丢弃法(Dropout Method)。循环神经网络因为其自身具有网络记忆能力，如果直接采用随即丢弃法，即直接随机丢弃网络中某个时刻的隐藏状态，可能在时间维度上会导致循环网络中的记忆存储能力遭到不可控的破坏。这时，随机丢弃非时间维连接(非循环连接)不失为一种更加简单有效的方法。如图 2，细箭头代表着随机丢弃操作，不同的丢弃掩码使用颜色不同的箭头表示。

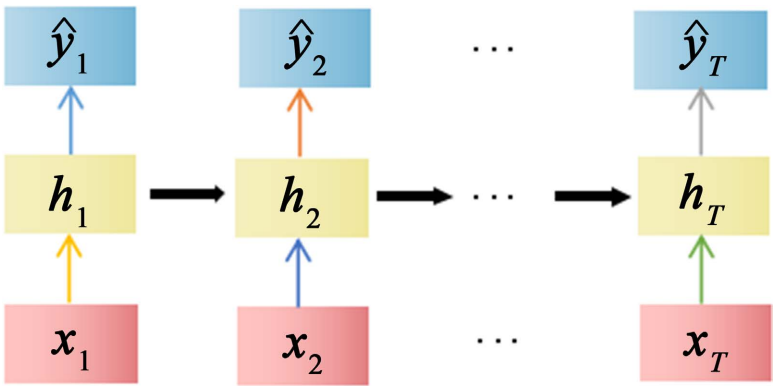

Figure 2. Schematic diagram of dropout method
图 2. 丢弃法原理图

非时间维连接丢弃法主要是依据贝叶斯中对丢弃方法解释，即丢弃法可以看作是对参数的采样，并且在任何时刻都需保证采样的参数是固定的。所以，当丢弃法被应用在循环神经网络防止其过拟合的情况时，有必要随机丢弃参数矩阵中的元素，同时始终保持丢弃掩码固定不变，这就是所谓的变分丢弃法(Variational Dropout)。如图 3 所示，可以看出丢弃掩码在不同时刻都是始终相同的。

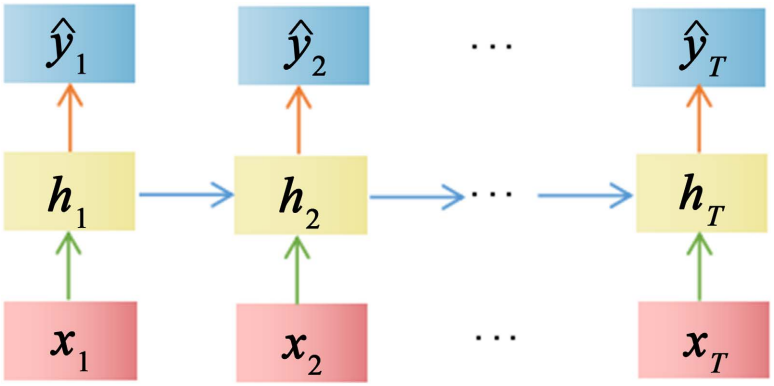

Figure 3. Schematic diagram of variational dropout
图 3. 变分丢弃法原理图

3. 模型介绍

本文对 SI Mimilakis [19]等人提出的基于无监督表征学习的音源分离平台 rl_singing_voice-master (RSVM)做出改进。RSVM 模型使用编码器和解码器进行音源分离实现。其中编码器在解码器的帮助下学习表征形式，解码器使用一个简单的正弦模型作为解码函数来重建唱歌的声音。如下图 4 所示。

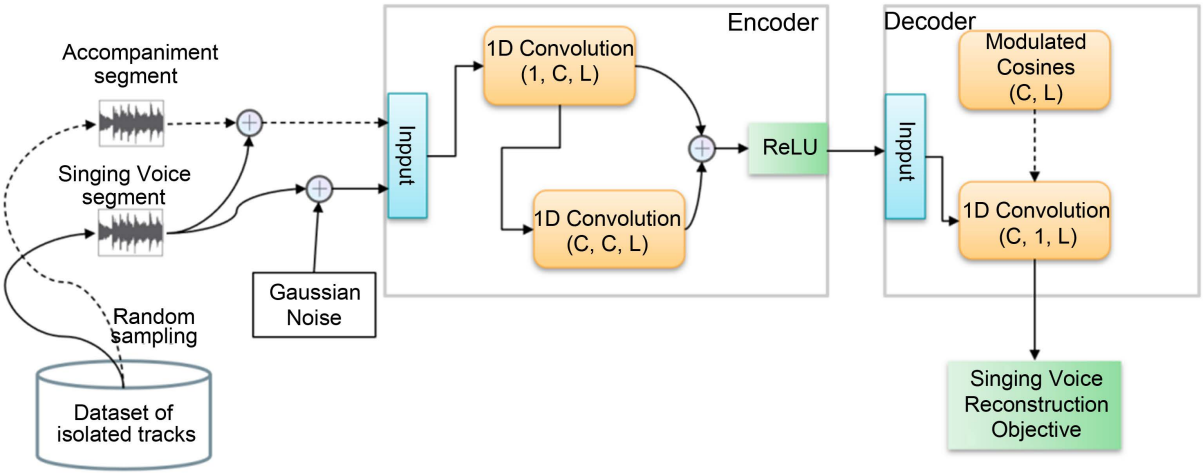

Figure 4. RSVM model structure diagram
图 4. RSVM 模型结构图

RSVM 模型虽然在分离效率及分离性能上表现出色，但其在面对故意攻击时的表现却并不突出。本文对 RSVM 模型的编码器部分进行改进，在两个卷积层之间添加自注意力机制，减少了模型在特征提取时对于外部信息的依赖，将关注度更多放在有效信息上。并使用变分丢弃法对其进行正则化处理，提升模型的泛化性。从而达到抵抗外部噪声故意攻击的目的。改进后的编码器如图 5 所示。

4. 实验

4.1. 实验概况

为了验证本文提出的音频源分离平台是否可以有效地提升分离平台的鲁棒性，本文采用了常用的攻击方法 PGD 对 RSVM 模型及改进后的模型进行有意攻击。为了最大限度保证实验结果的可靠性，在相同的实验环境中使用相同的训练参数对 RSVM 模型及改进后的模型进行训练。在测试模型分离性能时，

对其输入样本添加同等强度的攻击噪声测试模型的抗干扰能力。本实验采用的 CPU 为 Inter Core i9-10850k；GPU 为单卡 NVIDIA Geforce RTX 3090；内存为 24 GB；使用的操作系统为 ubuntu 18.04。实验中的参数设置-nb-workers=4 来加载 batch 数据，对 GPU 的利用率大约为 95%。

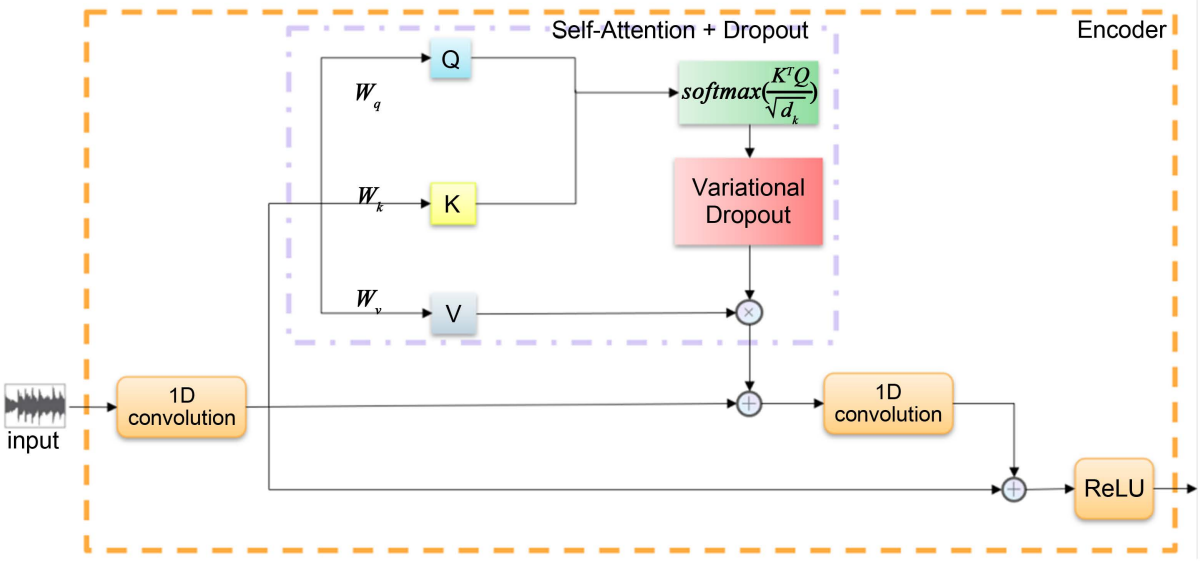

Figure 5. Partial structure diagram of improved RSVM model encoder
图 5. 改进后的 RSVM 模型编码器部分结构图

4.2. 数据源

训练和测试数据来自 MUSDB18 数据集[20]。MUSDB18 的训练和测试装置各有 100 和 50 条音乐轨道，都是立体声的，采样频率为 44100 赫兹。每个音轨文件由混音和它的四个源音频组成：“人声”、“鼓”、“低音”和“其他”。由于我们评估的为音乐人声分离，因此我们只使用“人声”源音频作为每个混合音轨的分离目标。

4.3. 评估指标

对音频源分离平台输出性能评估的主要方法有信源失真比 (SDR)、信噪比(SAR)以及信源干扰比(SIR) [21]。本文采用 SI-SDR 表示重构时域信号的信源失真比；BM-SISDR 表示音源分离结果的信源失真比(BS)；AT-BM-SISDR 表示加入对抗攻击后音源分离结果的信源失真比(AT-BS)；评估模型鲁棒性的标准本文设置为 BS 与 AT-BS 的差值，即对抗性攻击前后的音源分离结果的信源失真比之差，用 RB 表示。使用 museval 包计算所有 SDR 值，并报告每个轨道上度量值的中位数的所有轨道上的中位数。

4.4. 实验结果

在 MUSDB18 的测试数据集上，本实验对比分析了音频源分离平台 RSVM 及改进后的分离平台在对抗性攻击下的抗干扰能力，并对两平台的音源分离性能做出比较。实验结果如表 1 所示。从表 1 可以明显看出，在使用相同的实验环境及训练参数的情况下，改进后的 RSVM 分离平台的音源分离效果(BS)以及被有意攻击后的音源分离效果(AT-BS)相较于原 RSVM 分离平台均有明显提升。对比攻击前后的各分离平台，原 RSVM 分离平台在受到攻击后分离效果下降了 0.67 dB，而改进后的 RSVM 分离平台分离效果仅下降了 0.41 dB，证明了改进后的 RSVM 分离平台的鲁棒性得到了提升。虽然重构信号的得分(SI-SDR)呈下降趋势，但这是不可避免的，相比于分离性能以及鲁棒性的提升也是可以接受的。

Table 1. Experimental result data
表 1. 实验结果数据

	SI-SDR	BS	AT-BS	RB
RSVM 平台	29.88	4.96	4.29	0.67
改进后的 RSVM	28.52	5.45	5.03	0.41

5. 结束语

针对音频源分离模型在受到故意攻击时鲁棒性较差的问题，本文对 RSVM 分离平台做出改进。通过在分离模型编码器部分添加自注意力机制并对其采用变分丢弃法进行正则化处理，使得改进后的音频源分离台的性能得到优化，并提高了音频源分离平台的鲁棒性。同时验证了自注意力机制及变分丢弃法对鲁棒性提升上具有通用性及有效性。

参考文献

[1] Mesaros, A. and Virtanen, T. (2010) Recognition of Phonemes and Words in Singing. 2010 *IEEE International Conference on Acoustics*, *Speech and Signal Processing*, Dallas, 14-19 March 2010, 2146-2149. https://doi.org/10.1109/ICASSP.2010.5495585

[2] Fujihara, H., Goto, M., Ogata, J. and Okuno, H.G. (2011) Lyric Synchronizer: Automatic Synchronization System between Musical Audio Signals and Lyrics. *IEEE Journal of Selected Topics in Signal Processing*, **5**, 1252-1261. https://doi.org/10.1109/JSTSP.2011.2159577

[3] Sharma, B., Gupta, C., Li, H. and Wang, Y. (2019) Automatic Lyrics-to-Audio Alignment on Polyphonic Music Using Singing-Adapted Acoustic Models. 2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing*, Brighton, 12-17 May 2019, 396-400. https://doi.org/10.1109/ICASSP.2019.8682582

[4] Gillet, O. and Richard, G. (2008) Transcription and Separation of Drum Signals from Polyphonic Music. *The IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, **3**, 529-540. https://doi.org/10.1109/TASL.2007.914120

[5] Manilow, E., Seetharaman, P. and Pardo, B. (2020) Simultaneous Separation and Transcription of Mixtures with Multiple Polyphonic and Percussive Instruments. *IEEE International Conference on Acoustics*, *Speech and Signal Processing*, Barcelona, 4-8 May 2020, 771-775. https://doi.org/10.1109/ICASSP40776.2020.9054340

[6] Gómez, J.S., Abeßer, J. and Cano, E. (2018) Jazz Solo Instrument Classification with Convolutional Neural Networks, Source Separation, and Transfer Learning. *Proceedings of the* 19*th International Society for Music Information Retrieval Conference*, *ISMIR*, Paris, 23-27 September 2018, 577-584.

[7] Liu, J.-Y., Chen, Y.-H., Yeh, Y.-C. and Yang, Y.-H. (2019) Score and Lyrics-Free Singing Voice Generation.

[8] Jansson, A., Humphrey, E.J., Montecchio, N., Bittner, R.M., Kumar, A. and Weyde, T. (2017) Singing Voice Separation with Deep U-Net Convolutional Networks. 18*th International Society for Music Information Retrieval Conference*, Suzhou, 23-27 October 2017, 745-751.

[9] Takahashi, N. and Mitsufuji, Y. (2017) Multi-Scale Multi-Band DenseNets for Audio Source Separation. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), New Paltz, 15-18 October 2017, 261-265. https://doi.org/10.1109/WASPAA.2017.8169987

[10] Takahashi, N., Goswami, N. and Mitsufuji, Y. (2018) Mmdenselstm: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation. 2018 16*th International Workshop on Acoustic Signal Enhancement* (*IWAENC*), Tokyo, 17-20 September 2018, 106-110. https://doi.org/10.1109/IWAENC.2018.8521383

[11] Lee, J.H., Choi, H.-S. and Lee, K. (2019) Audio Query-Based Music Source Separation. *Proceedings of the* 20*th International Society for Music Information Retrieval Conference*, *ISMIR*, Delft, 4-8 November 2019, 878-885.

[12] Liu, J.-Y. and Yang, Y.-H. (2019) Dilated Convolution with Dilated GRU for Music Source Separation. *Proceedings International Joint Conference on Artificial Intelligence* (*IJCAI*), Macao, 10-16 August 2019, 4718-4724.

[13] Luo, Y. and Mesgarani, N. (2019) Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *The IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, **27**, 1256-1266. https://doi.org/10.1109/TASLP.2019.2915167

[14] Madry, A., Makelov, A., Schmidt, L., *et al*. (2017) Towards Deep Learning Models Resistant to Adversarial Attacks.

[15] Moosavi-Dezfooli, S.M., Fawzi, A. and Frossard, P. (2016) DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2574-2582. https://doi.org/10.1109/CVPR.2016.282

[16] Goodfellow, I.J., Shlens, J. and Szegedy, C. (2014) Explaining and Harnessing Adversarial Examples.

[17] Papernot, N., McDaniel, P., Jha, S., *et al*. (2016) The Limitations of Deep Learning in Adversarial Settings. 2016 *IEEE European Symposium on Security and Privacy*, Saarbruecken, 21-24 March 2016, 372-387. https://doi.org/10.1109/EuroSP.2016.36

[18] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014) Intriguing Properties of Neural Networks. https://arxiv.org/abs/1312.6199

[19] Mimilakis, S.I., Drossos, K. and Schuller, G. (2020) Unsupervised Interpretable Representation Learning for Singing Voice Separation. *EUSIPCO* 2020, Amsterdam, 24-28 August 2020, 1412-1416.

[20] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S.I. and Bittner, R. (2017) MUSDB18—A Corpus for Music Separation. https://hal.inria.fr/hal-02190845

[21] Vincent, E., Gribonval, R. and Févotte, C. (2006) Performance Measurement in Blind Audio Source Separation. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, **14**, 1462-1469. https://doi.org/10.1109/TSA.2005.858005