TOWARDS UNSUPERVISED SINGLE-CHANNEL BLIND SOURCE SEPARATION USING ADVERSARIAL PAIR UNMIX-AND-REMIX

Yedid Hoshen

Hebrew University of Jerusalem and Facebook AI Research

ABSTRACT

Blind single-channel source separation is a long standing signal processing challenge. Many methods were proposed to solve this task utilizing multiple signal priors such as low rank, sparsity, temporal continuity etc. The recent advance of generative adversarial models presented new opportunities in signal regression tasks. The power of adversarial training however has not yet been realized for blind source separation tasks. In this work, we propose a novel method for blind source separation (BSS) using adversarial methods. We rely on the independence of sources for creating adversarial constraints on pairs of approximately separated sources, which ensure good separation. Experiments are carried out on image sources validating the good performance of our approach, and presenting our method as a promising approach for solving BSS for general signals.

Index Terms— BSS, GANs, Source Separation, Adversarial Training, Unmixing

1. INTRODUCTION

The task of single-channel blind source separation (BSS) sets to reconstruct each of several sources (typically additively) mixed together. The task is poorly determined as more information needs to be reconstructed than the number of observations. BSS methods therefore need to rely on strong signal priors in order to constrain source reconstruction. Many priors were proposed for this task each giving rise to different optimization criteria. Source priors include: sparsity in time-frequency, non-Gaussian distribution of sources and low rank of sources. Recently, deep neural network methods that learn high quality signal representations (a form of prior learning) made much progress on single-channel source separation for cases where clean samples of each of the sources were available in training. This allowed creating synthetically mixed datasets, where random clean samples from each source are sampled and additively mixed. A deep neural network is then used to regress each of the components from the synthetic mixture. Such approaches are very effective due to learning source priors, rather than using generic hand-specified priors. Recent work was carried out to reduce the supervision required to having clean samples of only a single source, however when only mixed source samples are available (and no clean samples), classical methods are still used.

In this paper, we introduce a machine learning-based approach for the single-channel BSS case i.e. when no clean source samples are available at training time. Our method is based on generative adversarial networks (GANs) and uses a mixture of distributional, energy and cycle constraints to achieve high-quality unsupervised source separation. Our method makes the assumption of distributional independence between sources. In this work, we concentrate on the case where we are given mixed images (which is similar to having short audio clips) and do not take into account temporal priors (e.g. HMM models), which are left to future work. Our method is experimentally shown to outperform state-of-the-art single-channel BSS methods for image signals. Due to the strong performance on image signal separation, we believe that our approach presents a novel and promising direction for solving the long-standing task of single-channel BSS for general signals.

2. PREVIOUS WORK

Single-channel BSS has received much attention. The best results are typically obtained by using strong priors about the signals. Robust-PCA [1] separates instrumental and vocal sources by assuming that one source is low-rank while the vocal source is sparse. Results are improved with supervision [2]. For repetitive signals, Kernel Additive Models [3] may be used. Using temporal continuity was exploited by Roweis [4] and Virtanen [5]. Work was also done on designing priors for image mixture separation e.g. Levin and Weiss [6] used the presence of corners, although this method required access to clean signals.

Machine learning methods take away some of the difficulty in manual prior design. Speaker source separation was achieved by deep neural networks by [7]. Permutation-invariant training [8] allows speaker independent separation using deep learning methods. The above methods were used in a supervised source separation context i.e. when clean samples of each source are available. Supervised deep methods were also used for image separation (e.g. in the context of reflection removal [9, 10]). Generative adversarial networks (GANs) [11] were used by some researchers in a supervised

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

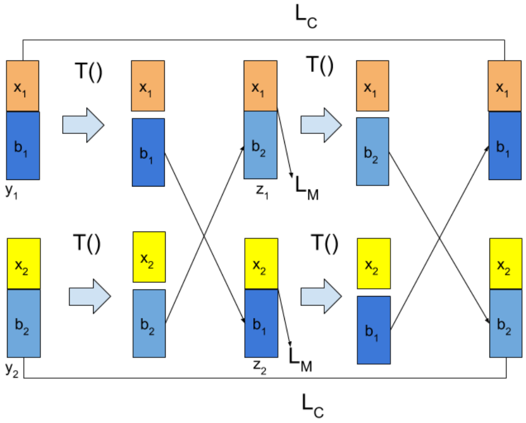

Fig. 1. A schematic of our architecture: we select a pair of samples y_1 and y_2, and separate them into estimated sources. We then flip the combination of sources between the two signals to synthesize z_1, z_2. We optimize $T(y)$ (implemented as $y \cdot M(y)$) to make the new mixtures z_1, z_2 indistinguishable from the original mixtures y_1, y_2. We then separate and remix the new mixtures again. The optimal separation function will recover the original mixtures y_1, y_2.

setting, for learning a better loss function [12, 13], typically with modest gains. They have been similarly used in image separation tasks with more significant gains [9].

There have been few attempts to apply deep methods for the unsupervised regime. In a recent work [14], we proposed using deep learning methods for semi-supervised separation (when samples of one source are available but not of the other). One of our baselines proposed a GAN-based method for learning the masking function. This method was outperformed by our main method - Neural Egg Separation which is non-adversarial. In this paper, we deal with the more challenging scenario, where no clean examples are available for any of the sources.

The architecture in DRIT [15] used for image style and content disentanglement bears some relation to ours as it uses cycle and pair-adversarial constraints . Our approach is different in key ways: we operate in the input rather than latent domain, we use masking rather than a set of encoders significantly constraining the network and improving results. We also introduce the energy equity term which is critical for the success of our approach.

3. ADVERSARIAL UNMIX-AND-REMIX

In the following, we denote the set of mixed signals as $y_1, y_2..y_N$. For ease of explanation, our formulation will assume two sources (however in Sec. 6 we explain why there is no loss of generality). We name our sources, $\mathcal{X}$ and $\mathcal{B}$ such that every mixed signal y_i consists of separate sources x_i and b_i, but no examples of such sources are given in the training set.

Our objective is to learn separation function $T()$ which separates a mixed signal y into its sources x and b. We parametrize the separation function by a multiplicative masking operation $M()$ as shown in Eq. 1:

$$T(y) = y \cdot M(y) \tag{1}$$

The separated sources are therefore given by $y \cdot M(y)$ and $y \cdot (1 - M(y))$. The masking function is learned as part of training.

Our method begins by sampling two mixed signals, which we will denote y_1 and y_2 (with no loss of generality). We operate the masking function on each mixture obtaining:

$$\begin{aligned} \widetilde{x_1} &= T(y_1) \quad & \widetilde{b_1} &= y_1 - T(y_1) \\ \widetilde{x_2} &= T(y_2) \quad & \widetilde{b_2} &= y_2 - T(y_2) \end{aligned} \tag{2}$$

We make the assumption of independence between the two sources $\mathcal{X}$ and $\mathcal{B}$. This assumption is valid for many interesting mixtures of signals such as images and reflections or foreground and background noise.

With this assumption, we can now synthesize new mixed signals z_1 and z_2, which are obtained by flipping the source combinations between the two pairs:

$$\begin{aligned} z_1 &= \widetilde{x_1} + \widetilde{b_2} \\ z_2 &= \widetilde{x_2} + \widetilde{b_1} \end{aligned} \tag{3}$$

Although the new mixed signals will be different from y_1 and y_2, we make the observation that their *distribution* should be the same as that of y_1 and y_2, if the separation works correctly. Therefore to encourage correct separation, we require the distribution of $\mathcal{Y}$ and $\mathcal{Z}$ to be identical. This can be enforced using an adversarial domain confusion constraint. Specifically this works by training a discriminator $D()$ to attempt to identify if a specific signal comes from $\mathcal{Y}$ or from $\mathcal{Z}$. The discriminator is trained using the following LS-GAN [16] loss function:

$$arg\min_{D} L_D = \sum_{y \in \mathcal{Y}} (D(y) - 1)^2 + \sum_{z \in \mathcal{Z}} D(z)^2 \quad (4)$$

We co-currently train the masking function $M()$ so that it acts to fool the discriminator by making the mixed signals $\mathcal{Z}$ as similar as possible to $\mathcal{Y}$:

$$arg\min_{M} L_M = \sum_{z \in \mathcal{Z}} (D(z) - 1)^2 \quad (5)$$

Where z iterates over all z_1 and z_2.

Although perfect separation is one possible solution of the distribution matching equation, another acceptable by unwanted solution is $\tilde{x} = y$ and $\tilde{b} = 0$. This trivial solution satisfies the distributional matching perfectly, but obviously achieves no separation. To combat this trivial solution, we add another loss term which favors solutions that give non-zero weights to the different sources:

$$L_E = \sum_{y \in \mathcal{Y}} (y \cdot M(y))^2 + (y \cdot (1 - M(y)))^2 \quad (6)$$

A further constraint on the separation can be obtained by another application of the separation function of the synthetic mixture signal pair $z_!$ and z_2. We perform the same unmixing and remixing operation as performed in the first stage:

$$\begin{array}{ll} \overline{x_1} = T(z_1) & \overline{b_2} = z_1 - T(z_1) \\ \overline{x_2} = T(z_2) & \overline{b_1} = z_2 - T(z_2) \end{array} \quad (7)$$

In this case, we notice that the result should be identical to the original unmixed signals y_1 and y_2:

$$\begin{array}{l} \overline{y_1} = \overline{x_1} + \overline{b_1} \\ \overline{y_2} = \overline{x_2} + \overline{b_2} \end{array} \quad (8)$$

We therefore introduce a "cycle" loss term, ensuring that the double application of unmixing and remix operation of a pair of mixed signals recovers the original signals:

$$L_C = \sum_{y \in \mathcal{Y}} \|\overline{y}, y\| \quad (9)$$

To summarize, our method optimizes the separation function $T()$ (which is implemented using multiplicative masking function $M()$ as described in Eq. 1). The loss function to be optimized is the combination of the domain confusion loss L_M, the energy equity loss L_E and the cycle reconstruction loss L_C:

$$arg\min_{M} L_{Total} = L_C + \alpha \cdot L_M + \beta \cdot L_E \quad (10)$$

We also adversarially optimize the discriminator $D()$ as described in Eq. 4.

4. IMPLEMENTATION

We implemented the masking function $M()$ by an architecture that follows DiscoGAN [17] with 64 channels (at the layer before last, each preceding layer having twice the number of channels). The discriminator followed a standard DC-GAN [18] architecture with 64 channels. We used a learning rate of 0.0001. Optimization was carried out by SGD with the ADAM update rule. We carried out 4 mask update steps for every $D()$ update. We used $\alpha = 5$ for the adversarial loss L_M and $\beta = 5$ for the energy equity loss L_E.

5. EVALUATION

In this section, we evaluate the effectiveness of our method for image separation tasks against other state-of-the-art unsupervised single channel source separation methods.

Datasets: We use the following image datasets in our experiments:

MNIST: The MNIST dataset [19] consists of 50000 training and 10000 validation images of hand written digits $0 - 9$. The images are roughly evenly distributed between the different classes. The original image resolution is 28×28. In order to use standard generative architectures, we pad the images by 2 pixels from each direction to have a size of 32×32. We split the dataset into two sources: the images of the digits from $0 - 4$ and the images of the digits from $5 - 9$. A random image is sampled from each source, and then combined with equal weights. We sampled $25k$ training mixture images (from the training sets), and $5k$ validation images from the validation set.

Shoes and Bags: The Shoes dataset [20] first collected by Yu and Grauman consists of color images of different types of shoes. We rescale the image resolution to 64×64. The Handbags dataset [21] collected by Zhu et al. consists of color images of a variety of handbags. We also rescale these images to a resolution of 64×64. The two datasets are often used in image generative modeling tasks. As masking works better when the background has 0 value, we run our experiments on the inverted intensity images (i.e. from image I, we use $255 - I$). Our sampling procedure is to randomly sample a shoe image and a handbag image (without replacement) and mix them with equal weights. This is repeated $10k$ times to form our training set. We similarly sample $5k$ mixture test images. No source image is repeated between the train and test sets.

Methods: Separating two images from arbitrary image classes does not satisfy the requirements for any of the typical priors as there are no obvious temporal, sparsity or low-rank constraints. We compare against RPCA [1] which is representative of methods that use strong priors. To represent decompositional methods we compare against GLO, a generative model (which in [14] was preferable to NMF). To have an upper bound for the quantitative comparison, we also give the

Fig. 2. A Qualitative Comparison of MNIST and Shoes/Bags Separation

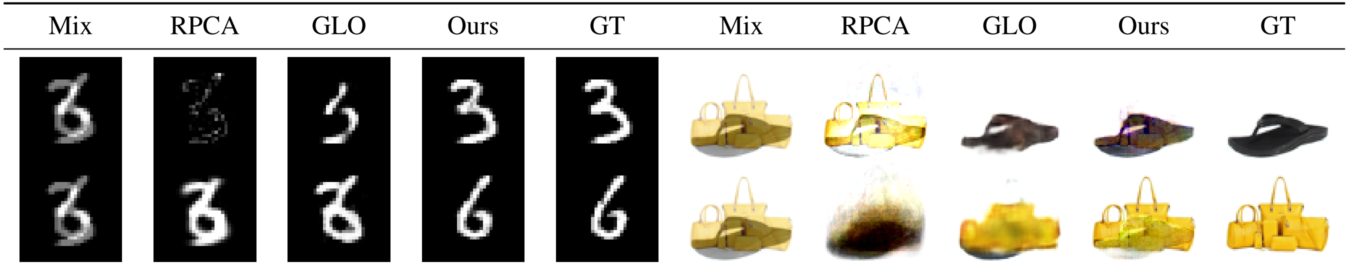

Table 1. Separation Accuracy (PSNR)

Dataset	*RPCA*	*GLO*	*Ours*	*Sup*
MNIST	11.5	13.0	**20.4**	24.4
Shoes and Bags	7.9	12.0	**19.0**	22.9

Table 2. Separation Accuracy (SSIM)

Dataset	*RPCA*	*GLO*	*Ours*	*Sup*
MNIST	0.36	0.74	**0.90**	0.96
Shoes and Bags	0.18	0.51	**0.73**	0.86

fully supervised performance (using the same masking function architecture that we used). We stress however that our method is fully unsupervised, and we do not expect to do better than the fully supervised method.

Qualitative Results: A qualitative comparison is presented in Fig. 2. We observe that RPCA completely fails on this task, as the sparse/low-rank prior is not suitable for arbitrary images. GLO tended to result in uneven separation - one generator containing a part of one source, while the other generator containing a mixture of the sources. Our method, generally resulted in clean separation of the sources. In highly textured regions, we sometimes saw some ”dripping” of the texture to the other source.

Quantitative Results: We present a quantitative comparison on MNIST and Shoes/Bags . The metrics are PSNR (in Tab. 1) and SSIM [22] (in Tab.2), which are standard image reconstruction quality metrics. In both cases we can observe that GLO performed much better than RPCA (due to the prior in RPCA being unsuitable for this more general task). Our method far outperformed both baseline methods, due to our careful separation design. The performance of our method approaches the supervised separation performance, however there still is a significant performance gap due to supervision, which is unsurprising. In ablation experiments, we found that the adversarial loss and the energy equity loss were essential for the convergence of our method to the correct solution. The cycle constraint was found to only slightly increase stability of convergence and did not increase accuracy. Overall, we can conclude that the results validate the strong performance of our method for separating image sources.

6. DISCUSSION

We make several comments about our work:

Priors: Our method was shown to be effective at separating mixtures of images, using no image specific priors such as repetition, sparsity or low-rank. It is therefore potentially extensible to all 2D signals. We make the general assumption that the distributions of the two signals are independent.

Spectrograms: Preliminary experiments on spectrograms were not able to match the success of the method for image separation. We think that this is due to GAN modeling of images being more developed than that of spectrograms. We believe that with future progress in adversarial architecture for spectrograms, our technique will be able to separate audio clips.

Multiple Sources: Although the formulation in this work only dealt with 2 sources, it can be applied to a larger number of sources, by a applying the method in a binary tree-like structure (recursively applying our method on each separated ”source” until reaching the leaves - the clean sources). We note however that the binary tree-like structure will need to have a stopping criterion detecting when a clean source has been found (similar to a leaf in a tree). We leave this to future work.

7. CONCLUSION

In this paper, we introduced a novel method for the single-channel separation of sources without seeing any clean examples of the individual sources. Previous methods have been able to achieve this either by learning strong priors from clean data or by carefully hand-crafting priors for particular sources. Our method makes very few assumptions on the sources, making it applicable to signals for which strong priors are not known. We demonstrated that our method works well on separating mixtures of images. Future work on adversarial training for spectrograms is needed to extend our approach to audio sources.

8. REFERENCES

[1] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *ICASSP*, 2012.

[2] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *ICASSP*, 2015.

[3] Antoine Liutkus, Derry Fitzgerald, Zafar Rafii, Bryan Pardo, and Laurent Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.

[4] Sam T Roweis, "One microphone source separation," in *NIPS*, 2001.

[5] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *TASLP*, 2007.

[6] Anat Levin, Assaf Zomet, and Yair Weiss, "Separating reflections from a single image using local features," in *CVPR*, 2014.

[7] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Deep learning for monaural speech separation," in *ICASSP*, 2014.

[8] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017.

[9] Zhixiang Chi, Xiaolin Wu, Xiao Shu, and Jinjin Gu, "Single image reflection removal using deep encoder-decoder network," in *arXiv preprint arXiv:1802.00094*, 2018.

[10] Xuaner Zhang, Ren Ng, and Qifeng Chen, "Single image reflection separation with perceptual losses," in *CVPR*, 2018.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[12] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *ICASSP*, 2018.

[13] Cem Subakan and Paris Smaragdis, "Generative adversarial source separation," in *ICASSP*, 2018.

[14] Tavi Halperin, Ariel Ephrat, and Yedid Hoshen, "Neural separation of observed and unobserved distributions," in *ICLR Submission*, 2018.

[15] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018.

[16] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.

[17] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017.

[18] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[19] Yann LeCun and Corinna Cortes, "MNIST handwritten digit database," 2010.

[20] Aron Yu and Kristen Grauman, "Fine-grained visual comparisons with local learning," in *CVPR*, 2014.

[21] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros, "Generative visual manipulation on the natural image manifold," in *ECCV*, 2016.

[22] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*, 2003.