Source Separation with Deep Generative Priors

Vivek Jayaram [*] **John Thickstun** [*]

Abstract

Despite substantial progress in signal source separation, results for richly structured data continue to contain perceptible artifacts. In contrast, recent deep generative models can produce authentic samples in a variety of domains that are indistinguishable from samples of the data distribution. This paper introduces a Bayesian approach to source separation that uses generative models as priors over the components of a mixture of sources, and noise-annealed Langevin dynamics to sample from the posterior distribution of sources given a mixture. This decouples the source separation problem from generative modeling, enabling us to directly use cutting-edge generative models as priors. The method achieves state-of-the-art performance for MNIST digit separation. We introduce new methodology for evaluating separation quality on richer datasets, providing quantitative evaluation of separation results on CIFAR-10. We also provide qualitative results on LSUN.

1. Introduction

The single-channel source separation problem (Davies & James, 2007) asks us to decompose a mixed signal $\mathbf{m} \in \mathcal{X}$ into a linear combination of k components $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathcal{X}$ with scalar mixing coefficients $\alpha_i \in \mathbb{R}$:

$$\mathbf{m} = g(\mathbf{x}) \equiv \sum_{i=1}^{k} \alpha_i \mathbf{x}_i. \tag{1}$$

This is motivated by, for example, the "cocktail party problem" of isolating the utterances of individual speakers $\mathbf{x}_i$ from an audio mixture $\mathbf{m}$ captured at a busy party, where multiple speakers are talking simultaneously.

[*] Equal contribution. Paul G. Allen School of Computer Science and Engineering, University of Washington. Correspondence to: Vivek Jayaram <vjayaram@cs.washington.edu>, John Thickstun <thickstn@cs.washington.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

With no further constraints or regularization, solving Equation (1) for $\mathbf{x}$ is highly underdetermined. Classical "blind" approaches to single-channel source separation resolve this ambiguity by privileging solutions to (1) that satisfiy mathematical constraints on the components $\mathbf{x}$, such as statistical independence (Davies & James, 2007) sparsity (Lee et al., 1999) or non-negativity (Lee & Seung, 1999). These constraints can be be viewed as weak priors on the structure of sources, but the approaches are blind in the sense that they do not require adaptation to a particular dataset.

Recently, most works have taken a data-driven approach. To separate a mixture of sources, it is natural to suppose that we have access to samples $\mathbf{x}$ of individual sources, which can be used as a reference for what the source components of a mixture are supposed to look like. This data can be used to regularize solutions of Equation (1) towards structurally plausible solutions. The prevailing way to do this is to construct a supervised regression model that maps an input mixture $\mathbf{m}$ to components $\mathbf{x}_i$ (Huang et al., 2014; Halperin et al., 2019). Paired training data $(\mathbf{m}, \mathbf{x})$ can be constructed by summing randomly chosen samples from the component distributions $\mathbf{x}_i$ and labeling these mixtures with the ground truth components.

Instead of regressing against components $\mathbf{x}$, we use samples to train a generative prior $p(\mathbf{x})$; we separate a mixed signal $\mathbf{m}$ by sampling from the posterior distribution $p(\mathbf{x}|\mathbf{m})$. For some mixtures this posterior is quite peaked, and sampling from $p(\mathbf{x}|\mathbf{m})$ recovers the only plausible separation of $\mathbf{m}$ into likely components. But in many cases, mixtures are highly ambiguous: see, for example, the orange-highlighted MNIST images in Figure 1. This motivates our interest in sampling, which explores the space of plausible separations. In Section 3 we introduce a procedure for sampling from the posterior, an extension of the noise-annealed Langevin dynamics introduced in Song & Ermon (2019), which we call Bayesian Annealed SIgnal Source separation: "BASIS" separation.

Ambiguous mixtures pose a challenge for traditional source separation metrics, which presume that the original mixture components are identifiable and compare the separated components to ground truth. For ambiguous mixtures of rich data, we argue that recovery of the original mixture components is not a well-posed problem. Instead, the problem

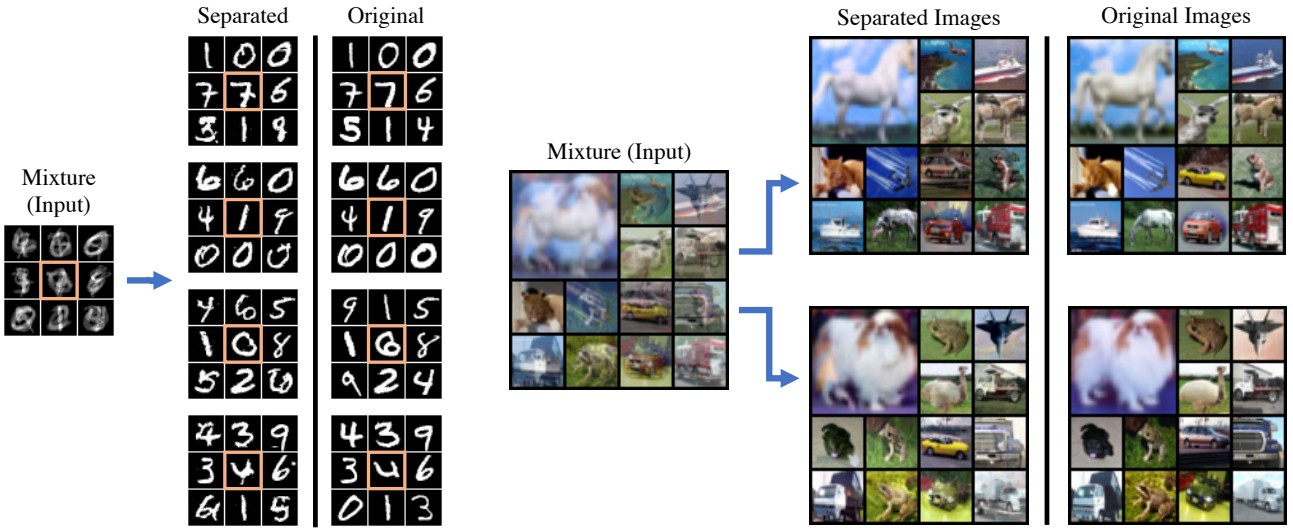

Figure 1. Separation results for mixtures of four images from the MNIST dataset (Left) and two images from the CIFAR-10 dataset (Right), using BASIS with the NCSN (Song & Ermon, 2019) generative model as a prior over images. We draw attention to the central panel of the MNIST results (highlighted in orange), which shows how a mixture can be separated in multiple ways.

we aim to solve is finding components of a mixture that are consistent with a particular data distribution. Motivated by this perspective, we discuss evaluation metrics in Section 4.

Formulating the source separation problem in a Bayesian framework decouples the problem of source generation from source separation. This allows us to leverage pre-trained, state-of-the-art, likelihood-based generative models as prior distributions, without requiring architectural modifications to adapt these models for source separation. Examples of source separation using noise-conditioned score networks (NCSN) (Song & Ermon, 2019) as a prior are presented in Figure 1. Further separation results using NCSN and Glow (Kingma & Dhariwal, 2018) are presented in Section 5.

2. Related Work

Blind separation. Work on blind source separation is data-agnostic, relying on generic mathematical properties to privilege particular solutions to (1) (Comon, 1994; Bell & Sejnowski, 1995; Davies & James, 2007; Huang et al., 2012). Because blind methods have no access to sample components, they face the challenging task of modeling the distribution over unobserved components while simultaneously decomposing mixtures into likely components. It is difficult to fit a rich model to latent components, so blind methods often rely on simple models such as dictionaries to capture the structure of these components.

One promising recent work in the blind setting is Double-DIP (Gandelsman et al., 2019). This work leverages the unsupervised Deep Image Prior (Ulyanov et al., 2018) as a prior over signal components, similar to our use of a trained generative model. But the authors of this work document fundamental obstructions to applying their method to single-channel source separation; they propose using multiple image frames from a video, or multiple mixtures of the same components with different mixing coefficients α. This multiple-mixture approach is common to much of the work on blind separation. In contrast, our approach is able to separate components from a single mixture.

Supervised regression. Regression models for source separation learn to predict components for a mixture using a dataset of mixed signals labeled with ground truth components. This approach has been extensively studied for separation of images (Halperin et al., 2019), audio spectrograms (Huang et al., 2014; 2015; Nugraha et al., 2016; Jansson et al., 2017), and raw audio (Lluis et al., 2019; Stoller et al., 2018b; Défossez et al., 2019), as well as more exotic data domains, e.g. medical imaging (Nishida et al., 1999). By learning to predict components (or equivalently, masks on a mixture) this approach implicitly builds a generative model of the signal components. This connection is made more explicit in recent work that uses GAN's to force components emitted by a regression model to match the distribution of a given dataset (Zhang et al., 2018; Stoller et al., 2018a).

The supervised approach takes advantage of expressive deep models to capture a strong prior over signal components. But it requires specialized model architectures trained specifically for the source separation task. In contrast, our approach leverages standard, pre-trained generative models for source separation. Furthermore, our approach can directly exploit ongoing advances in likelihood-based generative modeling to improve separation results.

Signal Dictionaries. Much work on source separation is based on the concept of a signal dictionary, most notably the line of work based on non-negative matrix factorization (NMF) (Lee & Seung, 2001). These approaches model signals as combinations of elements in a latent dictionary. Decomposing a mixture into dictionary elements can be used for source separation by (1) clustering the elements of the dictionary and (2) reconstituting a source using elements of the decomposition associated with a particular cluster.

Dictionaries are typically learned from data of each source type and combined into a joint dictionary, clustered by source type (Schmidt & Olsson, 2006; Virtanen, 2007). The blind setting has also been explored, where the clustering is obtained without labels by e.g. k-means (Spiertz & Gnann, 2009). Recent work explores more expressive decomposition models, replacing the linear decompositions used in NMF with expressive neural autoencoders (Smaragdis & Venkataramani, 2017; Venkataramani et al., 2017).

When the dictionary is learned with supervision from labeled sources, dictionary clusters can be interpreted as implicit priors on the distributions over components. Our approach makes these prior explicit, and works with generic priors that are not tied to the dictionary model. Furthermore, our method can separate mixed sources of the same type, whereas mixtures of sources with similar structure present a conceptual difficulty for dictionary-based methods.

Generative adversarial separation. Recent work by Subakan & Smaragdis (2018) and Kong et al. (2019) explores the intriguing possibility of optimizing $\mathbf{x}$ given a mixture $\mathbf{m}$ to satisfy (1), where components $\mathbf{x}_i$ are constrained to the manifold learned by a GAN. The GAN is pre-trained to model a distribution over components. Like our method, this approach leverages modern deep generative models in a way that decouples generation from source separation. We view this work as a natural analog to our likelihood-based approach in the GAN setting.

Likelihood-based approaches. Our approach is similar in spirit to older ideas based on maximum a posteriori estimation (Geman & Geman, 1984) likelihood maximization (Pearlmutter & Parra, 1997; Roweis, 2001) and Bayesian source separation (Benaroya et al., 2005). We build upon their insights, with the advantage of increased computational resources and modern expressive generative models.

3. BASIS Separation

We consider the following generative model of a mixed signal $\mathbf{m}$, relaxing the mixture constraint $g(\mathbf{x}) = \mathbf{m}$ to a soft Gaussian approximation:

$$\mathbf{x} \sim p, \tag{2}$$

$$\mathbf{m} \sim \mathcal{N}\left(g(\mathbf{x}), \gamma^2 I\right). \tag{3}$$

Algorithm 1 BASIS Separation

Input: $\mathbf{m} \in \mathcal{X}$, $\{\sigma_i\}_{i=1}^L$, δ, T
Sample $\mathbf{x}_1, \ldots, \mathbf{x}_k \sim \text{Uniform}(\mathcal{X})$
for $i \leftarrow 1$ **to** L **do**
  $\eta_i \leftarrow \delta \cdot \sigma_i^2 / \sigma_L^2$
  for $t = 1$ **to** T **do**
    Sample $\varepsilon_t \sim \mathcal{N}(0, I)$
    $\mathbf{u}^{(t)} \leftarrow \mathbf{x}^{(t)} + \eta_i \nabla_{\mathbf{x}} \log p_{\sigma_i}(\mathbf{x}^{(t)}) + \sqrt{2\eta}\varepsilon_t$
    $\mathbf{x}^{(t+1)} \leftarrow \mathbf{u}^{(t)} - \frac{\eta_i}{\sigma_i^2} \text{Diag}(\alpha) \left(\mathbf{m} - g(\mathbf{x}^{(t)})\right)$
  end for
end for

This defines a joint distribution $p_\gamma(\mathbf{x}, \mathbf{m}) = p(\mathbf{x}) p_\gamma(\mathbf{m}|\mathbf{x})$ over signal components $\mathbf{x}$ and mixtures $\mathbf{m}$, and a corresponding posterior distribution

$$p_\gamma(\mathbf{x}|\mathbf{m}) = p(\mathbf{x}) p_\gamma(\mathbf{m}|\mathbf{x}) / p_\gamma(\mathbf{m}). \tag{4}$$

In the limit as $\gamma^2 \to 0$, we recover the hard constraint on the mixture $\mathbf{m}$ given by Equation (1).

BASIS separation (Algorithm 1) presents an approach to sampling from (4) based on the discussion in Sections 3.1 and 3.2. In Section 3.3 we discuss the behavior of the gradients $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, which motivates some of the hyperparameter choices in Section 3.4. We describe a procedure to construct the noisy models p_{σ_i} required for BASIS in Section 3.5.

3.1. Langevin dynamics

Sampling from the posterior distribution $p_\gamma(\mathbf{x}|\mathbf{m})$ looks formidable; just computing Equation (4) requires evaluation of the partition function $p_\gamma(\mathbf{m})$. But using Langevin dynamics (Neal et al., 2011; Welling & Teh, 2011) we can sample $\mathbf{x} \sim p_\gamma(\cdot|\mathbf{m})$ while avoiding explicit computation of $p_\gamma(\mathbf{x}|\mathbf{m})$. Let $\mathbf{x}_0 \sim \text{Uniform}(\mathcal{X})$, $\varepsilon_t \sim \mathcal{N}(0, I)$, and define a sequence

$$\begin{aligned} \mathbf{x}^{(t+1)} &\equiv \mathbf{x}^{(t)} + \eta \nabla_{\mathbf{x}} \log p_\gamma(\mathbf{x}^{(t)}|\mathbf{m}) + \sqrt{2\eta}\varepsilon_t \\ &= \mathbf{x}^{(t)} + \eta \nabla_{\mathbf{x}} \left(\log p(\mathbf{x}^{(t)}) + \tfrac{1}{2\gamma^2} \|\mathbf{m} - g(\mathbf{x}^{(t)})\|^2 \right) + \sqrt{2\eta}\varepsilon_t. \end{aligned} \tag{5}$$

Observe that $\nabla_{\mathbf{x}} \log p_\gamma(\mathbf{m}) = 0$, so this term is not required to compute (5). By standard analysis of Langevin dynamics, as the step size $\eta \to 0$, $\lim_{t\to\infty} D_{KL}(\mathbf{x}_t \parallel \mathbf{x}|\mathbf{m}) = 0$, under regularity conditions on the distribution $p_\gamma(\mathbf{x}|\mathbf{m})$.

If the prior $p(\mathbf{x})$ is parameterized by a neural model, then gradients $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ can be computed by automatic differentiation with respect to the inputs of the generator network. This family of likelihood-based models includes autoregressive models (Salimans et al., 2017; Parmar et al., 2018), the variational autoencoder (Kingma & Welling, 2014; van den Oord et al., 2017), or flow-based models (Dinh et al., 2017;

Kingma & Dhariwal, 2018). Alternatively, if gradients of the distribution are modeled (Song & Ermon, 2019), then $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ can be used directly.

3.2. Accelerated mixing

To accelerate mixing of (5) we adopt a simulated annealing schedule over noisy approximations to the model $p(\mathbf{x})$, extending the unconditional sampling algorithm proposed in Song & Ermon (2019) to accelerate sampling from the posterior distribution $p_\gamma(\mathbf{x}|\mathbf{m})$. Let $p_\sigma(\mathbf{x})$ denote the distribution of $\mathbf{x} + \epsilon_\sigma$ for $\mathbf{x} \sim p$ and $\epsilon_\sigma \sim \mathcal{N}(0, \sigma^2 I)$. We define the noisy joint likelihood $p_{\sigma,\gamma}(\mathbf{x}, \mathbf{m}) \equiv p_\sigma(\mathbf{x}) p_\gamma(\mathbf{m}|\mathbf{x})$, which induces a noisy posterior approximation $p_{\sigma,\gamma}(\mathbf{x}|\mathbf{m})$. At high noise levels σ, $p_\sigma(\mathbf{x})$ is approximately Gaussian and irreducible, so the Langevin dynamics (5) will mix quickly. And as $\sigma \to 0$, $D_{KL}(p_\sigma \parallel p) \to 0$. This motivates defining the modified Langevin dynamics

$$\mathbf{x}^{(t+1)} \equiv \mathbf{x}^{(t)} + \eta \nabla_{\mathbf{x}} \log p_{\sigma,\gamma}(\mathbf{x}^{(t)}|\mathbf{m}) + \sqrt{2\eta}\varepsilon_t. \quad (6)$$

The dynamics (6) approximate samples from $p(\mathbf{x}|g(\mathbf{x}) = \mathbf{m})$ as $\eta \to 0$, $\gamma^2 \to 0$, $\sigma^2 \to 0$, and $t \to \infty$. An implementation of these dynamics, annealing η, γ^2, and σ^2 as $t \to \infty$ according to the hyper-parameter settings presented in Section 3.4, is presented in Algorithm 1.

We anneal η, γ^2, and σ^2 using a heuristic introduced in Song & Ermon (2019): the idea is to maintain a constant signal-to-noise ratio (SNR) between the expected size of the posterior log-likelihood gradient term $\eta \nabla_{\mathbf{x}} \log p_{\sigma,\gamma}(\mathbf{x}|\mathbf{m})$ and the expected size of the Langevin noise $\sqrt{2\eta}\varepsilon$:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim p_\sigma}\left[\left\| \frac{\eta \nabla_{\mathbf{x}} \log p_{\sigma,\gamma}(\mathbf{x}|\mathbf{m})}{\sqrt{2\eta}} \right\|^2\right] = \frac{\eta}{4} \mathop{\mathbb{E}}_{\mathbf{x}\sim p_\sigma}\left[\left\| \nabla_{\mathbf{x}} \log p_\gamma(\mathbf{m}|\mathbf{x}) + \nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x}) \right\|^2\right]. \quad (7)$$

Assuming that gradients w.r.t. to the likelihood and the prior are uncorrelated, the SNR is approximately

$$\frac{\eta}{4} \mathop{\mathbb{E}}_{\mathbf{x}\sim p_\sigma}\left[\left\| \nabla_{\mathbf{x}} \log p_\gamma(\mathbf{m}|\mathbf{x}) \right\|^2\right] + \frac{\eta}{4} \mathop{\mathbb{E}}_{\mathbf{x}\sim p_\sigma}\left[\left\| \nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x}) \right\|^2\right]. \quad (8)$$

Observe that $\log p_\gamma(\mathbf{m}|\mathbf{x})$ is a concave quadratic with smoothness proportional to $1/\gamma^2$; it follows analytically that $\mathbb{E}\left[\left\| \nabla_{\mathbf{x}} \log p_\gamma(\mathbf{m}|\mathbf{x}) \right\|^2\right] \propto 1/\gamma^2$. Song & Ermon (2019) found empirically that $\mathbb{E}\|\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})\|^2 \propto 1/\sigma^2$ for the NCSN model; we observe similar behavior for the flow-based Glow model (Kingma & Dhariwal, 2018) and in Section 3.3 we propose a possible explanation for this behavior. Therefore, to maintain a constant SNR, it suffices to set both γ^2 and σ^2 proportional to η.

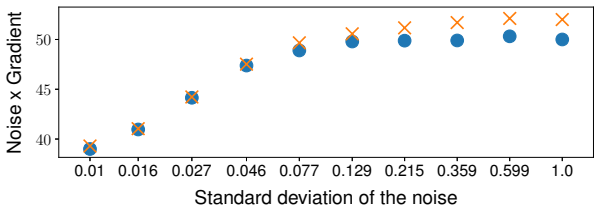

Figure 2. The behavior of $\sigma \times \|\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})\|$ in expectation for the NCSN (orange) and Glow (blue) models trained on CIFAR-10 at each of 10 noise levels as σ decays geometrically from 1.0 to 0.01. For large σ, $\|\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})\| \approx 50/\sigma$. This proportional relationship breaks down for smaller σ. Because the expected gradient of the noiseless density $\log p(\mathbf{x})$ is finite, its product with σ must asymptotically approach zero as $\sigma \to 0$.

3.3. The gradients of the noisy prior

We remark that the empirical finding $\mathbb{E}\|\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})\|^2 \propto 1/\sigma^2$ discussed in Section 3.2, and the consistency of this observation across models and datasets, could be surprising. Gradients of the noisy densities p_σ can be described by convolution of p with a Gaussian kernel:

$$\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x}) = \nabla_{\mathbf{x}} \log \mathop{\mathbb{E}}_{\epsilon\sim\mathcal{N}(0,I)} \left[p(\mathbf{x} - \sigma\epsilon)\right]. \quad (9)$$

From this expression, assuming p is continuous, we clearly see that the gradients are asymptotically independent of σ:

$$\lim_{\sigma\to 0} \nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}). \quad (10)$$

Maintaining proportionality $\mathbb{E}\|\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})\|^2 \propto 1/\sigma^2$ requires the gradients to grow unbounded as $\sigma \to 0$, but the gradients of the noiseless distribution $\log p(\mathbf{x})$ are finite. Therefore, proportionality must break down asymptotically and we conclude that–even though we turn the noise σ^2 down to visually imperceptible levels–we have not reached the asymptotic regime.

We conjecture that the proportionality between the gradients and the noise is a consequence of severe non-smoothness in the noiseless model $p(\mathbf{x})$. The probability mass of this distribution is peaked around plausible images $\mathbf{x}$, and decays rapidly away from these points in most directions. Consider the extreme case where the prior has a Dirac delta point mass. The convolution of a Dirac delta with a Gaussian is itself Gaussian so, near the point mass, the noisy distribution p_σ will be proportional to a Gaussian density with variance σ^2. If p_σ were exactly Gaussian then analytically

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim p_\sigma}\left[\|\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})\|^2\right] = \frac{1}{\sigma^4} \mathop{\mathbb{E}}_{\mathbf{x}\sim p_\sigma}\left[\mathbf{x}^2\right] = \frac{1}{\sigma^2}. \quad (11)$$

Because the distribution $p(\mathbf{x})$ does not contain actual delta spikes–only approximations thereof–we would expect this proportionality to eventually break down as $\sigma \to 0$. Indeed, Figure 2 shows that both for NCSN and Glow models of

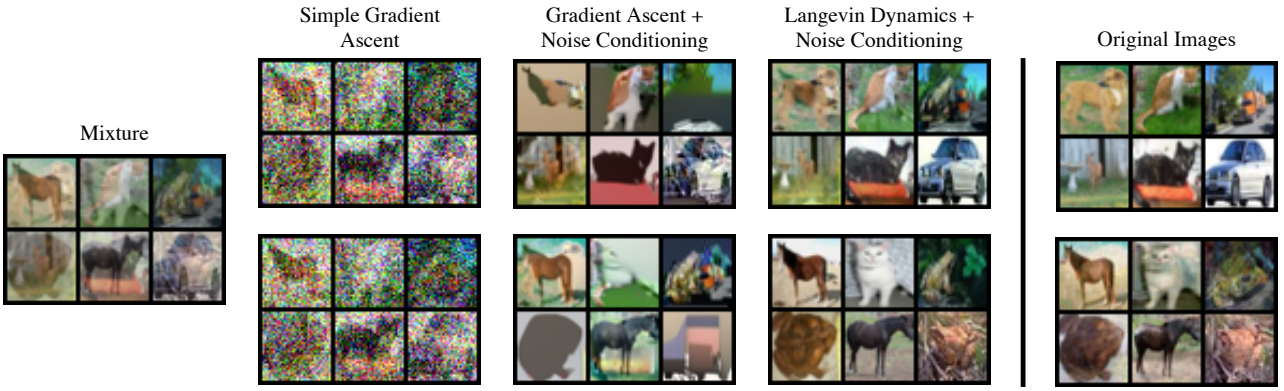

Figure 3. Non-stochastic gradient ascent produces sub-par results. Annealing over smoothed-out distributions (Noise Conditioning) guides the optimization towards likely regions of pixel space, but gets stuck at sub-optimal solutions. Adding Gaussian noise to the gradients (Langevin dynamics) shakes the optimization trajectory out of bad local optima.

CIFAR-10, after maintaining a very consistent proportionality $\mathbb{E}\left[\|\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})\|^2\right] \propto 1/\sigma^2$ at the higher noise levels, the decay of σ^2 to zero eventually outpaces the growth of the gradients.

3.4. Hyper-parameter settings

We adopt the hyper-parameters proposed by Song & Ermon (2019) for annealing σ^2, the proportionality constant δ, and the iteration count T. The noise σ is geometrically annealed from $\sigma_1 = 1.0$ to $\sigma_L = 0.01$ with $L = 10$. We set $\delta = 2 \times 10^{-5}$, and $T = 100$. We find that the same proportionality constant between σ^2 and η also works well for γ^2 and η, allowing us to set $\gamma^2 = \sigma^2$. We use these hyper-parameters for both the NCSN and Glow models, applied to each of the three datasets MNIST, CIFAR-10, and LSUN.

3.5. Constructing noise-conditioned models

For noise-conditioned score networks, we can directly compute $\nabla_{\mathbf{x}} \log p_\sigma(\mathbf{x})$ by evaluating the score network at the desired noise level. For generative flow models like Glow, these noisy distributions are not directly accessible. We could estimate the distributions $p_\sigma(\mathbf{x})$ by training Glow from scratch on datasets perturbed by each of the required noise levels σ^2. But this not practical; Glow is expensive to train, requiring thousands of epochs to converge and consuming hundreds of gpu-hours to obtain good models even for small low-resolution datasets.

Instead of training models $p_\sigma(\mathbf{x})$ from scratch, we apply the concept of fine-tuning from transfer learning (Yosinski et al., 2014). Using pre-trained models of $p(\mathbf{x})$ published by the Glow authors, we fine-tune these models on noise-perturbed data $\mathbf{x} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Empirically, this procedure quickly converges to an estimate of $p_\sigma(\mathbf{x})$, within about 10 epochs.

3.6. The importance of stochasticity

We remark that adding Gaussian noise to the gradients in the BASIS algorithm is essential. If we set aside the Bayesian perspective, it is tempting to simply run gradient ascent on the pixels of the components to maximize the likelihood of these components under the prior, with a Lagrangian term to enforce the mixture constraint $g(\mathbf{x}) = \mathbf{m}$:

$$\mathbf{x} \leftarrow \mathbf{x} + \eta \nabla_{\mathbf{x}} \left[\log p(\mathbf{x}) - \lambda \|g(\mathbf{x}) - \mathbf{m}\|^2\right]. \tag{12}$$

But this does not work. As demonstrated in Figure 3, there are many local optima in the loss surface of $p(\mathbf{x})$ and a greedy ascent procedure simply gets stuck. Pragmatically, the noise term in Langevin dynamics can be seen as a way to knock the greedy optimization (12) out of local maxima.

In the recent literature, pixel-space optimizations by following gradients $\nabla_{\mathbf{x}}$ of some objective are perhaps associated more with adversarial examples than with desirable results (Goodfellow et al., 2015; Nguyen et al., 2015). We note that there have been some successes of pixel-wise optimization in texture synthesis (Gatys et al., 2015) and style transfer (Gatys et al., 2016). But broadly speaking, pixel-space optimization procedures often seem to go wrong. We speculate that noisy optimizations (6) on smoothed-out objectives like p_σ could be a widely applicable method for making pixel-space optimizations more robust.

4. Evaluation Methodology

Many previous works on source separation evaluate their results using peak signal-to noise ratio (PSNR) or structural similarity index (SSIM) (Wang et al., 2004). These metrics assume that the original sources are identifiable; in probabilistic terms, the true posterior distribution $p(\mathbf{x}|\mathbf{m})$ is presumed to have a unique global maximum achieved by the

ground truth sources (up to permutation of the sources). Under the identifiability assumption, it is reasonable to measure the quality of a separation algorithm by comparing separated sources to ground truth mixture components. PSNR, for example, evaluates separations by computing the mean-squared distance between pixel values of the ground truth and separated sources on a logarithmic scale.

For CIFAR-10 source separation, the ground truth source components of a mixture are not identifiable. As evidence for this claim, we call the reader's attention to Figure 4. For each mixture depicted in Figure 4, we present separation results that sum to the mixture and (to our eyes) look plausibly like CIFAR-10 images. However, in each case the separated images exhibit high deviation from the ground truth. This phenomenon is not unusual; Figure 5 shows an un-curated collection of samples from $p(\mathbf{x}|\mathbf{m})$ using BASIS, illustrating a variety of plausible separation results for each given mixture. We will later see evidence again of non-identifiability in Figure 7. If we accept that the separations presented in Figures 4, 5, and 7 are reasonable, then source separation on this dataset is fundamentally underdetermined; we cannot measure success using metrics like PSNR that compare separation results to ground truth.

Instead of comparing separations to ground truth, we propose instead to quantify the extent to which the results of a source separation algorithm look like samples from the data distribution. If a pair of images sum to the given mixture and look like samples from the data distribution, we deem the separation to be a success. This shift in perspective from identifiability of the latent components to the quality of the separated components is analogous to the classical distinction in the statistical literature between estimation and prediction (Shmueli et al., 2010; Bellec et al., 2018). To this end, we borrow the Inception Score (IS) (Salimans et al., 2016) and Frechet Inception Distance (FID) (Heusel et al., 2017) metrics from the generative modeling literature to evaluate CIFAR-10 separation results. These metrics attempt to quantify the similarity between two distributions given samples. We use them to compare the distribution of components produced by a separation algorithm to the distribution of ground truth images.

In contrast to CIFAR-10, the posterior distribution $p(\mathbf{x}|\mathbf{m})$ for an MNIST model is demonstrably peaked. Moreover, BASIS is able to consistently identify these peaks. This constitutes a constructive proof that components of MNIST mixtures are identifiable, and therefore comparisons to the ground-truth components make sense. We report PSNR results for MNIST, which allows us to compare the results of BASIS to other recent work on MNIST image separation (Halperin et al., 2019; Kong et al., 2019).

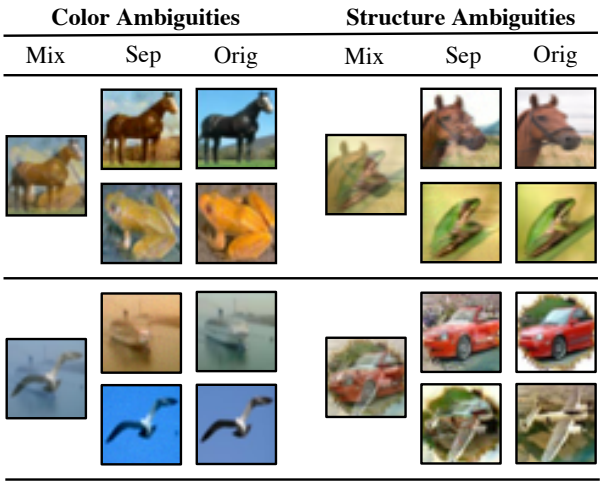

Figure 4. A curated collection of examples demonstrating color and structural ambiguities in CIFAR-10 mixtures. In each case, the original components differ substantially from the components separated by BASIS using NCSN as a prior. But in each case, the separation results also look like plausible CIFAR-10 images.

5. Experiments

We evaluate results of BASIS on 3 datasets: MNIST (LeCun et al., 1998) CIFAR-10 (Krizhevsky, 2009) and LSUN (Yu et al., 2015). For MNIST and CIFAR-10, we consider both NCSN (Song & Ermon, 2019) and Glow (Kingma & Dhariwal, 2018) models as priors, using pre-trained weights published by the authors of these models. For LSUN there is no pre-trained NCSN model, so we consider results only with Glow. For Glow, we fine-tune the weights of the pre-trained models to construct noisy models p_σ using the procedure described in Section 3.5. Code and instructions for reproducing these experiments is available online.[1]

Baselines. On MNIST we compare to results reported for the GAN-based "S-D" method (Kong et al., 2019) and the fully supervised version of Neural Egg separation "NES" (Halperin et al., 2019). Results for MNIST are presented in Section 5.1. To the best of our knowledge there are no previously reported quantitative metrics for CIFAR-10 separation, so as a baseline we ran Neural Egg separation on CIFAR-10 using the authors' published code. CIFAR-10 results are presented in Section 5.2. We present additional qualitative results for 64×64 LSUN in Section 5.3, which demonstrate that BASIS scales to larger images.

We also consider results for a simple baseline, "Average," that separates a mixture $\mathbf{m}$ into two 50% masks $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{m}/2$. This is a surprisingly competitive baseline. Observe that if we had no prior information about the distribution of components, and we measure separation quality by PSNR, then by a symmetry argument setting $\mathbf{x}_1 = \mathbf{x}_2$ is the optimal

[1] https://github.com/jthickstun/basis-separation

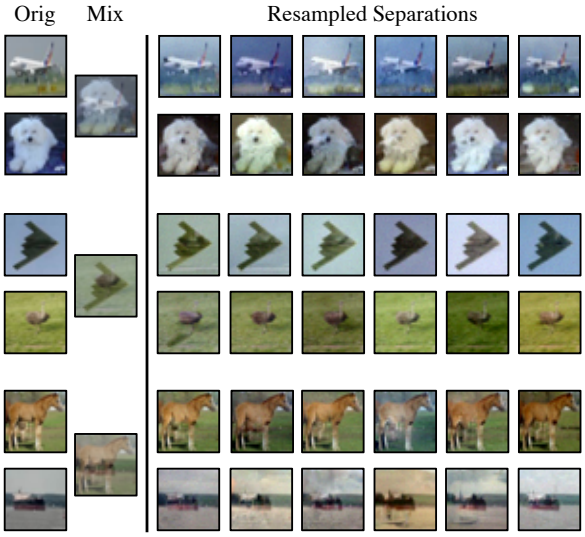

Figure 5. Repeated sampling using BASIS with NCSN as a prior for several mixtures of CIFAR-10 images. While most separations look reasonable, variation in color and lighting makes comparative metrics like PSNR unreliable. This challenges the notion that the ground truth components are identifiable.

separation strategy in expectation. In principle we would expect Average to perform very poorly under IS/FID, because these metrics purport to measure similarity of distributions and mixtures should have little or no support under the data distribution. But we find that IS and FID both assign reasonably good scores to Average, presumably because mixtures exhibit many features that are well supported by the data distribution. This speaks to well-known difficulties in evaluating generative models (Theis et al., 2016) and could explain the strength of "Average" as a baseline.

We remark that we cannot compare our algorithm to the separation-like task reported for CapsuleNets (Sabour et al., 2017). The segmentation task discussed in that work is similar to source separation, but the mixtures used for the segmentation task are constructed using the non-linear threshold function $h(\mathbf{x}) = \max(\mathbf{x}_1 + \mathbf{x}_2, 1)$, in contrast to our linear function g. While extending the techniques of this paper to non-linear relationships between $\mathbf{x}$ and $\mathbf{m}$ is intriguing, we leave this to future this work.

Class conditional separation. The Neural Egg separation algorithm is designed with the assumption that the components $\mathbf{x}_i$ are drawn from different distributions. For quantitative results on MNIST and CIFAR-10, we therefore consider two slightly different tasks. The first is class-agnostic, where we construct mixtures by summing randomly selected images from the test set. The second is class-conditional, where we partition the test set into two groupings: digits $0-4$ and $5-9$ for MNIST, animals and machines for CIFAR-10. The former task allows us compare to S-D results on MNIST, and the latter task allows us to compare to Neural Egg separation on MNIST and CIFAR-10.

There are two different ways to apply a prior for class-conditional separation. First observe that, because $\mathbf{x}_1$ and $\mathbf{x}_2$ are chosen independently,

$$p(\mathbf{x}) = p(\mathbf{x}_1, \mathbf{x}_2) = p_1(\mathbf{x}_1)p_2(\mathbf{x}_2). \tag{13}$$

In the class agnostic setting, $\mathbf{x}_1$ and $\mathbf{x}_2$ are drawn from the same distribution (the empirical distribution of the test set) so it makes sense to use a single prior $p = p_1 = p_2$. In the class conditional setting, we could potentially use separate priors over components $\mathbf{x}_1$ and $\mathbf{x}_2$. For the MNIST and CIFAR-10 experiments in this paper, we use pre-trained models trained on unconditional distribution of the training data for both the class agnostic and class conditional setting. It is possible that better results could be achieved in the class conditional setting by re-training the models on class conditional training data. For LSUN, the authors of Glow provide separate pre-trained models for the Church and Bedroom categories, so we are able to demonstrate class-conditional LSUN separations using distinct priors in Section 5.3.

Sample Likelihoods. Although we do not directly model the posterior likelihood $p(\mathbf{x}|\mathbf{m})$, we can compute the log-likelihood of the output samples $\mathbf{x}$. The log-likelihood is a function of the artificial variance hyper-parameter γ, so it is more informative to look at the unweighted square error $\|\mathbf{m} - g(\mathbf{x})\|^2$; this quantity can be interpreted as a reconstruction error, and measures how well we approximate the hard mixture constraint. Because we geometrically anneal the variance γ, by the end of optimization the mixture constraint is rigorously enforced; per-pixel reconstruction error is smaller than the quantization level of 8-bit color, resulting in pixel-perfect visual reconstructions.

For Glow, we can also compute the log-probability of samples under the prior. How do the probabilities of sources $\mathbf{x}_{\text{BASIS}}$ constructed by BASIS separation compare to the probabilities of data $\mathbf{x}_{\text{test}}$ taken directly from a dataset's test set? Because we anneal the noise to a fixed level $\sigma_L > 0$, we find it most informative to ask this question using the minimal-noise, fine-tuned prior $p_{\sigma_L}(\mathbf{x})$. As seen in Table 1, the outputs of BASIS separation are generally comparable in log-likelihood to test set images; BASIS separation recovers sources deemed typical by the prior.

Table 1. The mean log-likelihood under the minimal-noise Glow prior $p_{\sigma_L}(\mathbf{x})$ for the test set $\mathbf{x}_{\text{test}}$, and for samples of 100 BASIS separations $\mathbf{x}_{\text{BASIS}}$. The log-likelihood of each test set under the noiseless prior $p(\mathbf{x}_{\text{test}})$ is reported for reference.

Dataset	$p(\mathbf{x}_{\text{test}})$	$p_{\sigma_L}(\mathbf{x}_{\text{test}})$	$p_{\sigma_L}(\mathbf{x}_{\text{BASIS}})$
MNIST	0.5	3.6	3.6
CIFAR-10	3.4	4.5	4.7
LSUN (bed)	2.4	4.2	4.4
LSUN (crh)	2.7	4.4	4.4

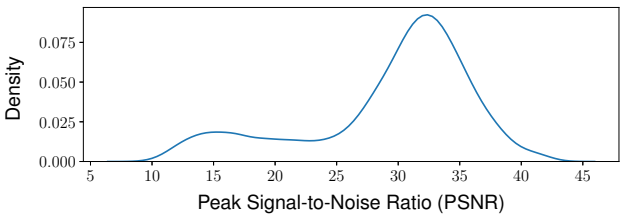

Figure 6. The empirical distribution of PSNR for 5,000 class agnostic MNIST digit separations using BASIS with the NCSN prior (see Table 2 for comparison of the central tendencies of this and other separation methods).

5.1. MNIST separation

Quantitative results for MNIST image separation are reported in Table 2, and a panel of visual separation results are presented in Figure 1. For quantitative results, we report mean PSNR over separations of $12,000$ separated components. The distribution of PSNR for class agnostic MNIST separation is visualized in Figure 6. We observe that approximately $2/3$ of results exceed the mean PSNR of 29.5, which to our eyes is visually indistinguishable from ground truth.

A natural approach to improve separation performance is to sample multiple $\mathbf{x} \sim p(\cdot|\mathbf{m})$ for a given mixture $\mathbf{m}$. A major advantage of models like Glow, that explicitly parameterize the prior $p(\mathbf{x})$, is that we can approximate the maximum of the posterior distribution with the maximum over multiple samples. By construction, samples from BASIS approximately satisfy $g(\mathbf{x}) = \mathbf{m}$, so for the noiseless model we simply declare $p(\mathbf{m}|\mathbf{x}) = 1$ and therefore $p(\mathbf{x}|\mathbf{m}) \propto p(\mathbf{x})$. We demonstrate the effectiveness of resampling in Table 2 (Glow, 10x) by comparing the expected PSNR of $\mathbf{x} \sim p(\cdot|\mathbf{m})$ to the expected PSNR of $\arg\max_i p(\mathbf{x}_i)$ over 10 samples $\mathbf{x}_1, \ldots, \mathbf{x}_{10} \sim p(\cdot|\mathbf{m})$. Even moderate resampling dramatically improves separation performance. Unfortunately this approach cannot be applied to the otherwise superior NCSN model, which does not model explicit likelihoods $p(\mathbf{x})$.

Table 2. PSNR results for separating 6,000 pairs of equally mixed MNIST images. For class split results, one image comes from label $0-4$ and the other comes from $5-9$. We compare to S-D (Kong et al., 2019), NES (Halperin et al., 2019), convolutional NMF (class split) (Halperin et al., 2019) and standard NMF (class agnostic) (Kong et al., 2019).

Algorithm	Class Split	Class Agnostic
Average	14.8	14.9
NMF	16.0	9.4
S-D	-	18.5
BASIS (Glow)	22.9	22.7
NES	24.3	-
BASIS (Glow, 10x)	27.7	27.1
BASIS (NCSN)	**29.5**	**29.3**

Without any modification, we can apply BASIS to separate mixtures of $k > 2$ images. We contrast this with regression-based methods, which require re-training to target varying numbers of components. Figure 1 shows the results of BASIS using the NCSN prior applied to mixtures of four randomly selected images. For more mixture components, we observe that identifiability of ground truth sources begins to break down. This is illustrated by looking at the central item in each panel of Figure 1 (highlighted in orange).

5.2. CIFAR-10

Quantitative results for CIFAR-10 image separation measured are presented in Table 3, and visual separation results are presented in Figure 1.

We can also view image colorization (Levin et al., 2004; Zhang et al., 2016) as a source separation problem by interpreting a grayscale image as a mixture of the three color channels of an image $\mathbf{x} = (\mathbf{x}_r, \mathbf{x}_g, \mathbf{x}_b)$ with

$$g(\mathbf{x}) = (\mathbf{x}_r + \mathbf{x}_g + \mathbf{x}_b)/3. \tag{14}$$

Unlike our previous separation problems, the channels of an image are clearly not independent, and the factorization of p given by Equation 13 is unwarranted. But conveniently, a generative model trained on color CIFAR-10 images itself models the joint distribution $p(\mathbf{x}) = p(\mathbf{x}_r, \mathbf{x}_g, \mathbf{x}_b)$. Therefore, the same pre-trained generative model that we use to separate images can also be used to color them.

Qualitative colorization results are visualized in Figure 7. The non-identifiability of ground truth is profound for this task (see Section 4 for discussion of identifiability). We draw attention to the two cars in the middle of the panel: the white car that is colored yellow by the algorithm, and the blue car that is colored red. The colors of these specific cars cannot be inferred from a greyscale image; the best an

Table 3. Inception Score / FID Score of 25,000 separations (50,000 separated images) of two overlapping CIFAR-10 images using NCSN as a prior. In Class Split one image comes from the category of animals and other from the category of vehicles. NES results using published code from Halperin et al. (2019).

Algorithm	Inception Score	FID
Class Split		
NES	5.29 ± 0.08	51.39
BASIS (Glow)	5.74 ± 0.05	40.21
Average	6.14 ± 0.11	39.49
BASIS (NCSN)	$\mathbf{7.83 \pm 0.15}$	**29.92**
Class Agnostic		
BASIS (Glow)	6.10 ± 0.07	37.09
Average	7.18 ± 0.08	28.02
BASIS (NCSN)	$\mathbf{8.29 \pm 0.16}$	**22.12**

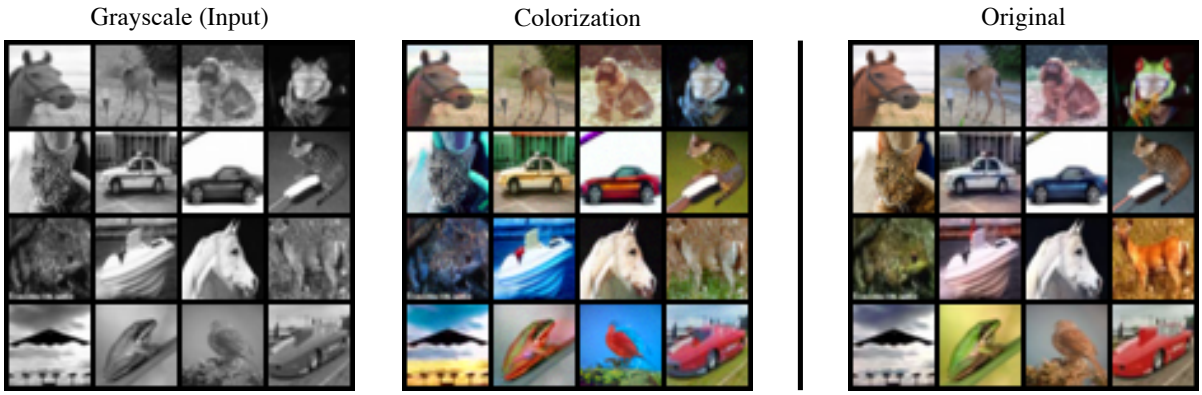

Figure 7. Colorizing CIFAR-10 images. Left: original CIFAR-10 images. Middle: greyscale conversions of the images on the left. Right: imputed colors for the greyscale images, found by BASIS using NCSN as a prior.

Table 4. Inception Score / FID Score of 50,000 colorized CIFAR-10 images. As measured by IS/FID, the quality of NCSN colorizations nearly matches CIFAR-10 itself.

Data Distribution	Inception Score	FID Score
Input Grayscale	8.01 ± 0.10	68.52
BASIS (Glow)	8.69 ± 0.15	28.70
BASIS (NCSN)	$\mathbf{10.53 \pm 0.17}$	**11.58**
CIFAR-10 Original	11.24 ± 0.12	0.00

algorithm can do is to choose a reasonable color, based on prior information about the colors of cars.

Quantitative coloring results for CIFAR-10 are presented in Table 4. We remark that the IS and FID scores for coloring are substantially better than the IS and FID scores of 8.87 and 25.32 respectively reported for unconditional samples from the NCSN model; conditioning on a greyscale image is enormously informative. Indeed, the Inception Score of NCSN-colorized CIFAR-10 is close to the Inception Score of the CIFAR-10 dataset itself.

5.3. LSUN separation

Qualitative results for LSUN separations are visualized in Figure 8. While the separation results in Figure 8 are imperfect, Table 1 shows that the mean log-likelihood of the separated components is comparable to the mean log-likelihood that the model assigns to images in the test set. This suggests that the model is incapable of distinguishing these separations from better results, and the imperfections are attributable to the quality of the model rather than to the separation algorithm. This is encouraging, because it suggests that the artifacts are due to the Glow model rather than the BASIS separation algorithm, and that better separation results will be achievable with improved generative models.

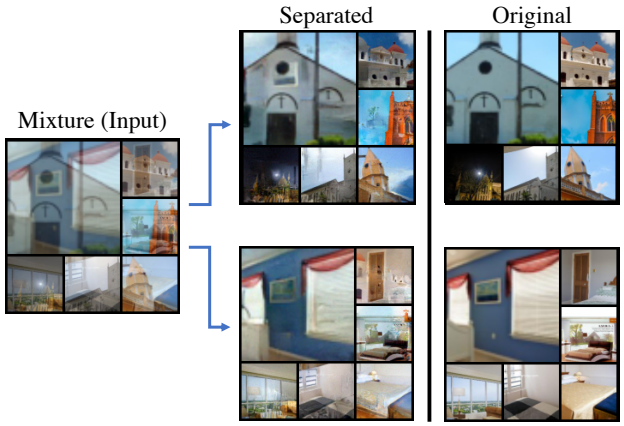

Figure 8. 64×64 LSUN separation results using Glow as a prior. One mixture component is sampled from the LSUN churches category, and the other component is sampled from LSUN bedrooms.

6. Conclusion

In this paper, we introduced a new approach to source separation that makes use of a likelihood-based generative model as a prior. We demonstrated the ability to swap in different generative models for this purpose, presenting results of our algorithm using both NCSN and Glow. We proposed new methodology for evaluating source separation on richer datasets, demonstrating strong performance on MNIST and CIFAR-10. Finally, we presented qualitative results on LSUN that point the way towards scaling this method to practical tasks such as speech separation, using generative audio models like WaveNets (Oord et al., 2016).

Acknowledgements

We thank Zaid Harchaoui, Sham M. Kakade, Steven Seitz, and Ira Kemelmacher-Shlizerman for valuable discussion and computing resources. This work was supported by the National Science Foundation Grant DGE-1256082.

References

Bell, A. J. and Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Bellec, P. C., Lecué, G., Tsybakov, A. B., et al. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

Benaroya, L., Bimbot, F., and Gribonval, R. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):191–199, 2005.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Davies, M. E. and James, C. J. Source separation using single channel ica. *Signal Processing*, 87(8):1819–1832, 2007.

Défossez, A., Usunier, N., Bottou, L., and Bach, F. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *International Conference on Learning Representations*, 2017.

Gandelsman, Y., Shocher, A., and Irani, M. Double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *The IEEE Conference on Computer Vision and Pattern Recognition*, volume 6, pp. 2, 2019.

Gatys, L., Ecker, A. S., and Bethge, M. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 262–270, 2015.

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

Geman, S. and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Halperin, T., Ephrat, A., and Hoshen, Y. Neural separation of observed and unobserved distributions. *Advances in Neural Information Processing Systems*, 2019.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Huang, P.-S., Chen, S. D., Smaragdis, P., and Hasegawa-Johnson, M. Singing-voice separation from monaural recordings using robust principal component analysis. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 57–60. IEEE, 2012.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *International Symposium on Music Information Retrieval*, pp. 477–482, 2014.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, 2015.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. Singing voice separation with deep u-net convolutional networks. 2017.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

Kong, Q., Xu, Y., Jackson, P. J. B., Wang, W., and Plumbley, M. D. Single-channel signal separation and deconvolution with generative adversarial networks. In *International Joint Conference on Artificial Intelligence*, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788, 1999.

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

Lee, T.-W., Lewicki, M. S., Girolami, M., and Sejnowski, T. J. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE signal processing letters*, 6(4):87–90, 1999.

Levin, A., Lischinski, D., and Weiss, Y. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, pp. 689–694. 2004.

Lluis, F., Pons, J., and Serra, X. End-to-end music source separation: is it possible in the waveform domain? *Interspeech*, 2019.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

Nishida, S., Nakamura, M., Ikeda, A., and Shibasaki, H. Signal separation of background eeg and spike by using morphological filter. *Medical engineering & physics*, 21 (9):601–608, 1999.

Nugraha, A. A., Liutkus, A., and Vincent, E. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664, 2016.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A., and Tran, D. Image transformer. *International Conference on Machine Learning*, 2018.

Pearlmutter, B. A. and Parra, L. C. Maximum likelihood blind source separation: A context-sensitive generalization of ica. In *Advances in Neural Information Processing Systems*, pp. 613–619, 1997.

Roweis, S. T. One microphone source separation. In *Advances in Neural Information Processing Systems*, pp. 793–799, 2001.

Sabour, S., Frosst, N., and Hinton, G. E. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pp. 3856–3866, 2017.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *International Conference on Learning Representations*, 2017.

Schmidt, M. N. and Olsson, R. K. Single-channel speech separation using sparse non-negative matrix factorization. In *International Conference on Spoken Language Processing*, 2006.

Shmueli, G. et al. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Smaragdis, P. and Venkataramani, S. A neural network alternative to non-negative audio models. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 86–90. IEEE, 2017.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019.

Spiertz, M. and Gnann, V. Source-filter based clustering for monaural blind source separation. In *International Conference on Digital Audio Effects*, 2009.

Stoller, D., Ewert, S., and Dixon, S. Adversarial semi-supervised audio source separation applied to singing voice extraction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2391–2395. IEEE, 2018a.

Stoller, D., Ewert, S., and Dixon, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *International Symposium on Music Information Retrieval*, 2018b.

Subakan, Y. C. and Smaragdis, P. Generative adversarial source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 26–30. IEEE, 2018.

Theis, L., Oord, A. v. d., and Bethge, M. A note on the evaluation of generative models. *International Conference on Learning Representations*, 2016.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

van den Oord, A., Vinyals, O., et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

Venkataramani, S., Subakan, C., and Smaragdis, P. Neural network alternatives toconvolutive audio models for source separation. In *International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2017.

Virtanen, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, pp. 681–688, 2011.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Zhang, X., Ng, R., and Chen, Q. Single image reflection separation with perceptual losses. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4786–4794, 2018.