An Analytic Theory of Creativity in Convolutional Diffusion Models

Mason Kamb [1] Surya Ganguli [1]

[1]Department of Applied Physics, Stanford University

Main contribution

Score-matching diffusion models can generate highly original images that lie far from their training data. However, optimal score-matching theory suggests that these models should only be able to produce memorized training examples. To understand how diffusion models can exhibit 'creative' characteristics, we need to understand *why the underfit their training objective*, and *what they do instead*. We study the problem of approximating the score using **Fully-Convolutional Diffusion Models.** These models have two key implicit biases:

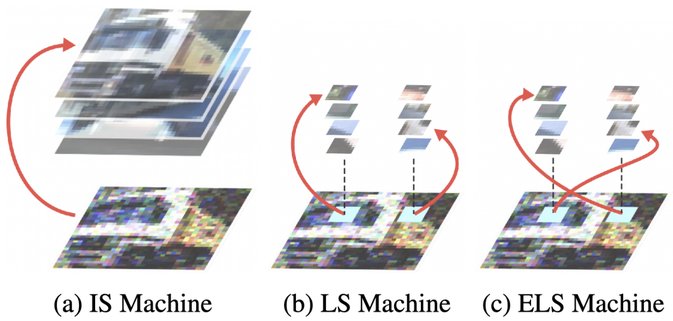
- **Translational equivariance**: applying the model to a translated version of an input image results in an equally translated output.
- **Locality:** the convolutional filters used are typically very narrow. For a finite-depth network, this means that only the pixels in a *local region* around the pixel can be used to estimate the noise.

We develop an analytic theory of optimal score-matching under these constraints. The resulting solution is:

- **Inherently 'creative,'** automatically mixing and matching elements of the training data rather than memorizing.
- **Highly predictive** of the behavior of trained diffusion models, achieving SOTA theory/experiment of $\mathbf{r^2 \sim 0.77 - 0.96}$, typically > 0.90, across different architectures and datasets.

Theory

Diffusion models are trained by taking training images, 'diffusing' them by adding noise, and then training a model to denoise the image by evolving images uphill along the 'score:' the log-probability gradient $s_t(\phi) = \nabla \log \pi_t(\phi)$.



(a) IS Machine (b) LS Machine (c) ELS Machine

- The *ideal* score (IS) can be written as a linear combination of the displacement from each training sample, times a *global* Bayes weight for each data point:

$$s_t(\phi, x) \propto \underbrace{\sum_{\varphi}}_{\text{sum over data}} \underbrace{(\phi(x) - \varphi(x))}_{\text{displacement}} \underbrace{P(\varphi|\phi)}_{\text{Bayes weight}} \tag{1}$$

- The ideal *local* approximation to the score (LS): each pixel x has *its own belief state* about which image it came from, *based only on the information in its local neighborhood*, Ω_x.

$$s_t(\phi, x) \propto \underbrace{\sum_{\varphi}}_{\text{sum over data}} \underbrace{(\phi(x) - \varphi(x))}_{\text{displacement}} \underbrace{P(\varphi|\phi \in \Omega_x)}_{\text{local Bayes weight}} \tag{2}$$

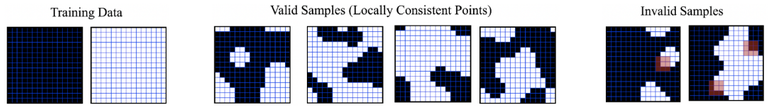
- The ideal *equivariant, local* approximation to the score (ELS): dataset augmented with *all possible translations* of the original dataset.

$$s_t(\phi, x) \propto \underbrace{\sum_{T(\varphi)}}_{\text{sum over data + translations}} \underbrace{(\phi(x) - \varphi(x))}_{\text{displacement}} \underbrace{P(\varphi|\phi \in \Omega_x)}_{\text{local Bayes weight}} \tag{3}$$

Key takeaway:

- Under the *ideal score*, all pixels share the same belief state, and move in lockstep towards most probable image.
- Under a *local approximation to the score*, pixel beliefs decouple. Each pixel moves uphill towards the *locally* most probable image, resulting in images that *mix and match* training set patches.

The 'pixel mosaic' model



Whereas optimal denoising is guaranteed to produce exact copies of the training data, optimal *local* denoising is guaranteed to produce images consisting of *locally consistent points*, where the center pixel of each patch matches the center pixel of the l_2-nearest patch in the training set.

A time-dependent locality scale and the curse of dimensionality

Our theory has one hyperparameter: the locality scale at each time in the reverse process. We find empirically that neural networks use a **coarse-to-fine** procedure, with large scales at high noise levels and small scales at low noise. The time-dependent locality scale provides a key mechanism by which these models can avoid the curse of dimensionality and memorization. By projecting to progressively lower dimensional subspaces, the data stay 'close together' at any level of noise, ensuring that new points are 'in distribution' and preventing the belief state from localizing to any one data point (i.e. memorizing).

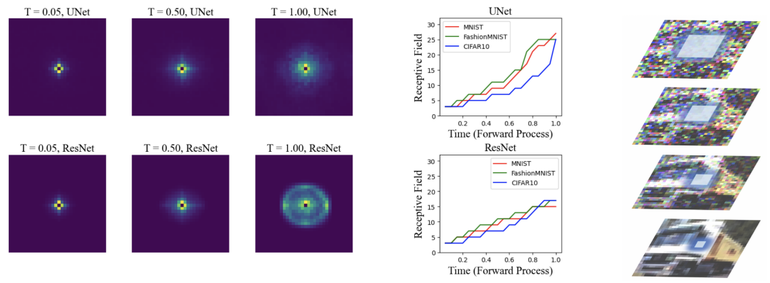


Figure 1. Left: receptive field sizes for different models at different times in the reverse process. Middle: Optimal scales monotonically decrease as the noise level decreases across all models and datasets. Right: schematic depiction of time-dependent locality scale.

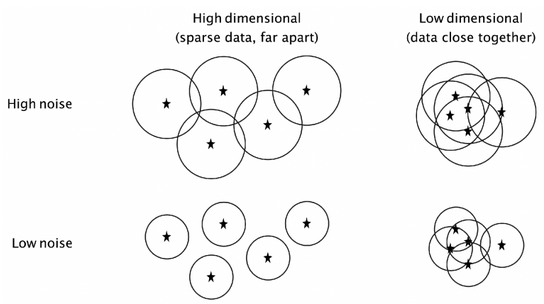
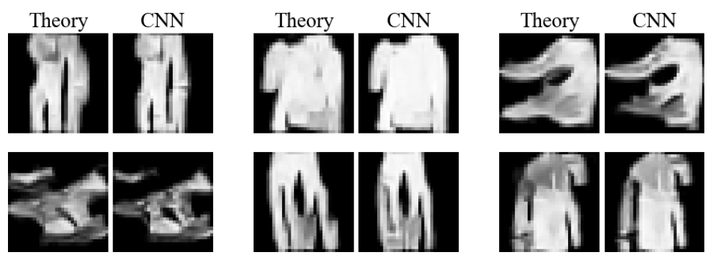


Figure 2. As the noise level decreases, the locality scale must be decreased commensurately in order to ensure that the model does not transition to a pathological memorizing state.

The local origin of excess limbs

Spatial consistency issues (e.g. incorrect numbers of digits and limbs) are ubiquitous in diffusion models. Our theoretical model reproduces these behaviors and *explains them* as a consequence of excessive locality.



Samples

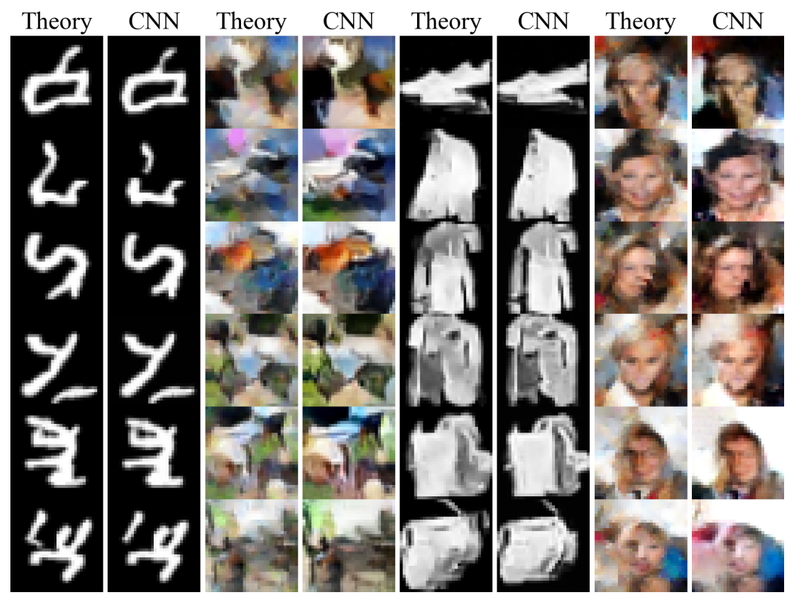


Figure 3. Our analytic theory (left columns) can accurately predict on a *case by case basis* the outputs of convolutional diffusion models (right columns), with UNet or ResNet architectures trained on MNIST, CIFAR10, FashionMNIST, and CelebA (left to right), even when these outputs are highly original and far from the training data.

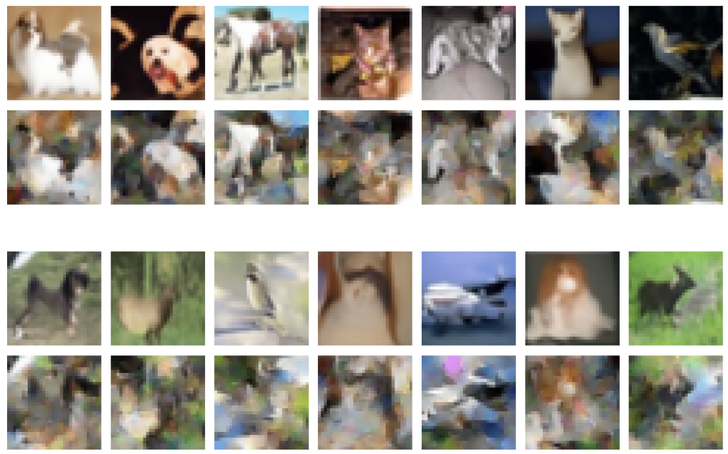


Figure 4. In practice, our theory is also moderately predictive ($r^2 \sim 0.77$) of a (small, simple) Attention-enabled model on CIFAR10. However, self-attention enables the model to 'carve out' semantically coherent foreground objects.

Dataset	Arch.	ELS Corr.	LS Corr.	IS Corr.
MNIST	UNet	**0.89**	0.88	0.70
CIFAR10	UNet	**0.90**	0.87	0.41
FashionMNIST	UNet	**0.93**	0.93	0.80
CelebA	UNet	0.85	**0.90**	0.55
MNIST	ResNet	**0.94**	0.82	0.61
CIFAR10	ResNet	**0.95**	0.90	0.42
FashionMNIST	ResNet	**0.94**	0.88	0.68
CelebA	ResNet	**0.96**	0.90	0.47
CIFAR10	UNet + SA	**0.77**	0.77	0.48

Table 1. A summary of the experimental results of the paper for different datasets and model configurations for each architecture across each dataset.