Origins of Creativity in Attention-Based Diffusion Models

Emma Finn[1,2], T. Anderson Keller[1,2], Manos Theodosis[1], Demba E. Ba[1]

[1] *The Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, USA*

[2] *CRISP Lab at Harvard University SEAS, USA*

Motivation: The Paradox of Creativity

Diffusion models have had remarkable success over the last decade in generating a hugely diverse set of visually plausible images. These models work by transforming the data to a centered Gaussian and then learning the reverse process, often by training a neural network to approximate the gradient of the log density of the underlying distribution (the score). Prior work has demonstrated that **if this score function is learned exactly, the model would only learn to regenerate training data** (Kamb et al. 2024, Brioli et al.). Why are diffusion models so good at generating novel but "correct" samples and how can we encourage this behavior? We build on existing work by Kamb and Ganguli (2024) to explain this paradox by analyzing a CNN+attention score approximation and corroborating it on a custom toy dataset.

Background: Learning the Score

Diffusion models learn to de-noise data by training a neural net: $f_\theta(\phi_x, t) \approx \nabla_{\phi_x} \log \pi_t(\phi)$.

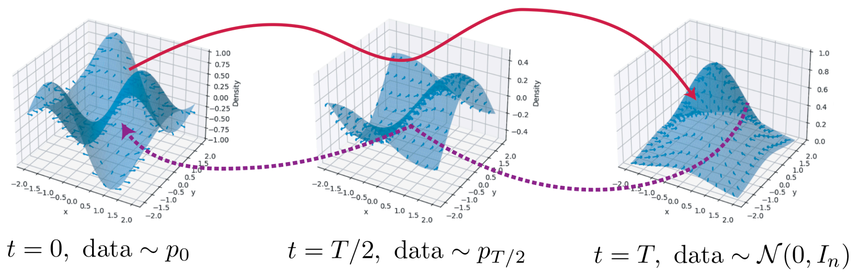


Figure 1: Transformation learned by diffusion.

Diffusion models are often trained by minimizing the "score matching loss."

$$\mathscr{L} = \sum_x \mathbb{E}_{\phi \sim \pi_t} || f_\theta[\phi](x) - \underbrace{s_t[\phi](x)}_{\text{true score}} ||^2$$

However, if the score is learned exactly, diffusion models fail to generalize and only generate elements of the training set. The architecture of the network f_θ alters the optimal solution. In particular, the inductive biases of f_θ create a series of errors in approximating the score that allow the model to generalize.

Related Work: Creativity in CNNs

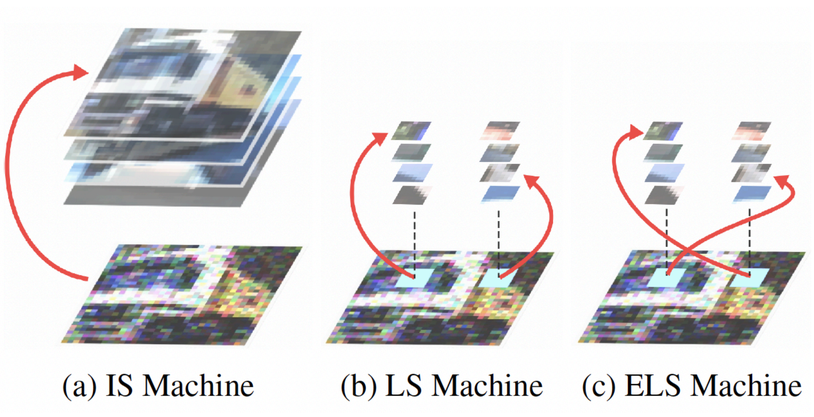


Figure 2: From Kamb and Ganguli [2024]

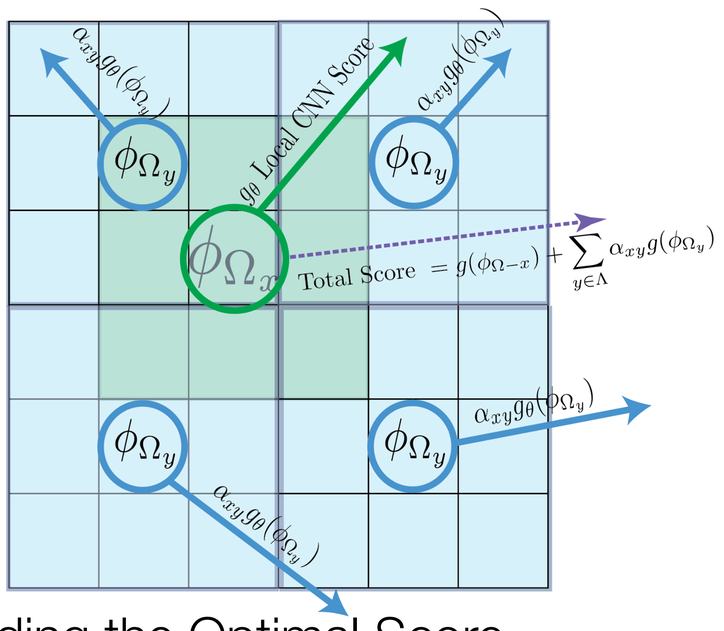
In prior work, Kamb and Ganguli [2024] showed that the **inductive biases of a CNN (namely translation equivariance and locality) produced a "patch mosaic" form of creativity.** They demonstrated that for CNN-backed diffusion, they can "partially predict the results" of pre-trained diffusion models. However, state of the art diffusion models rely on U-Nets with self-attention to parametrize their score function. Attention strongly violates the assumption of locality and the assumption of translation equivariance weakly.

Creativity in CNNs+Attention Layer

We take a preliminary step towards bridging the gap between the existing theory and state of the art models. **We derive a simple theoretical example that suggests that self attention enforces "global self-consistency."** We assume the following CNN+Attention parametrization of the score where ϕ is an arbitrary image, x is a pixel location and Λ is the set of all such locations. We let g_θ be a CNN and add a final self attention layer on top

$$\tilde{g}[\phi](x) = g_\theta(\phi_{\Omega_x}) + \sum_{y \in \Lambda} \frac{\exp(\langle g(\phi_{\Omega_x}), g(\phi_{\Omega_y}) \rangle}{\sum_{y' \in \Lambda} \exp(\langle g(\phi_{\Omega_x}), g(\phi_{\Omega_{y'}}) \rangle} g(\phi_{\Omega_y})$$

This corresponds to the simplest kind of attention, where we downgrade our key and query matrices from learnable parameters into fixed identity matrices.



Finding the Optimal Score

Figure 3: The learned score is the sum of the local score and context patches weighted by similarity.

Given the functional form of our score, we can optimize over the MSE to solve for the "optimal score function" subject to the constraint that that score is parametrized by a CNN with one attention layer.

$$0 = \sum_x \int \pi_t \left[z_x + \sum_y \alpha_{xy} z_y - s[\phi](x) \right]^T \delta_x \, d\phi$$

$$+ \int \pi_t \left[z_x + \sum_y \alpha_{xy} g_y - s[\phi](x) \right]^T \sum_y \alpha_{xy} \left(I + (z_x - \mu_x)\, \delta_y \right) d\phi$$

where

- $z_p = g(\phi_{\Omega_p})$ is the embedding of the patch centered at point p in the image ϕ
- δ_p is shorthand for the indicator that the patch centered at point p is equal to the random patch selected Φ,
- $\alpha_{xy} = \text{softmax}\langle z_x, z_y \rangle$ attention weight from patch x to patch y
- $\mu_x := \sum_p \alpha_{xp}(\phi)\, z_p$ is the attention weighted average embedding at x.

We intuitively interpret this as follows:

- **The first term (with δ_x)**
  - corresponds to when Φ is the query
  - matches the CNN-only case
  - encourages $g(\phi_{\Omega_x})$ to match the true score of ϕ_{Ω_x} considered as a purely local patch
- **The second term (with attention)**
  - corresponds to when Φ is the key or value
  - encourages $g(\phi_{\Omega_x})$ to provide more information about the image at position y
  - represents the **contextual correction** provided by our single attention head

Toy Dataset

We verify this approach by **showing that a simple attention-based diffusion model can learn to reproduce self-consistent structures in images while a CNN-based diffusion model struggles.** We evaluate the capacity to construct consistent samples by measuring how often the samples generated by our trained diffusion model obey the rules of the dataset. We construct a dataset to test this consistency property.

K_1	$f(K_1)$	K_2	$f(K_2)$
K_3	$f(K_3)$	K_4	$f(K_4)$

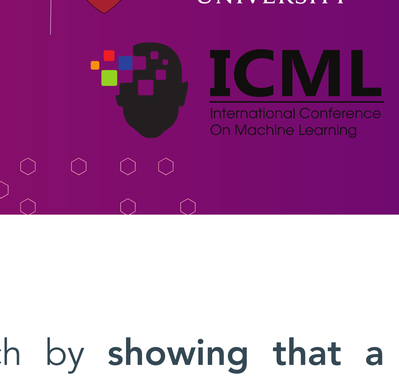


Figure 4: Toy Dataset Structure

To generate an image, we first select four keys K_1, K_2, K_3, K_4 uniformly at random with replacement from a list of four colors red, green, blue, and yellow. Then, we sample a function also uniformly at random from the set of functions $f : (R, G, B, Y) \to (R, G, B, Y)$, where each f associates one output color to one input color. **We say a generated image is "consistent" if the key-value pairs correspond to a *valid one-to-one* function in that space.**

Experiments

Model	Consistency
Random Baseline	5.38%
CNN	10.88%
CNN + Top 1 Attention	21.64%
CNN + Identity Attention	25.44%
CNN + Attention	64.03%

We train four simple diffusion models on this dataset: a pure CNN-based model, a CNN-based model with top-1 attention, a CNN-based model with attention with identity key and query matrices, and a CNN + Attention model. We implement our score network as a simple 2 layer CNN. Our CNN+Attention model has exactly the same CNN base but includes a single-headed self-attention block. Samples generated by attention-based models demonstrated increased consistency and higher quality image generation overall.

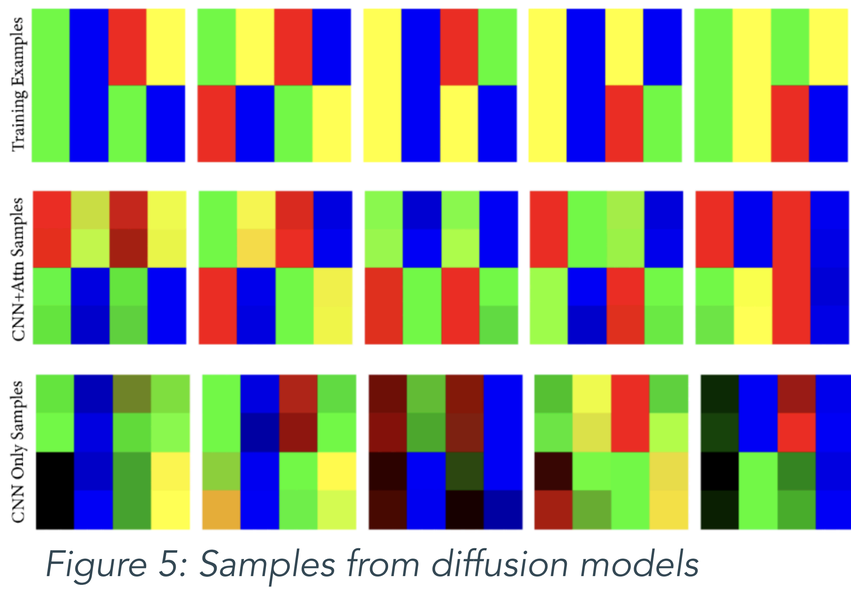


Figure 5: Samples from diffusion models

References:

Preprint

Biroli, G., Bonnaire, T., De Bortoli, V., and Mézard, M. Dynamical regimes of diffusion models. arXiv preprint arXiv:2402.18491, 2024.

Kamb, M., and Ganguli, S. An analytic theory of creativity in convolutional diffusion models, 2024. URL https://arxiv.org/abs/2412.20292.