Generalization in diffusion models arises from geometry-adaptive harmonic representations

Zahra Kadkhodaie September 2024

Florentin Guth Eero Simoncelli Stéphane Mallat

- Sample from a **learned** density [Song & Ermon 2019; Ho et al 2020]
- How is this possible, given the "Curse of dimensionality" ?!

Diffusion models embed densities

generated by a diffusion model

[Ho et al, 2022]

Memorization vs. generalization

A collection of delta functions

Underlying distribution

[Carlini et al, 2023]

[Somepalli et al, 2023]

A collection of delta functions

A continuous model of the underlying distribution

A point in training set Underlying distribution

Memorization vs. generalization

A collection of delta functions

A continuous model of the underlying distribution

A point in training set Underlying distribution

1. Can diffusion models generalize?

2. If so, how?

Memorization vs. generalization

denoised image (y)

Denoiser is applied iteratively and partially

Diffusion models and denoising Published as a conference paper at ICLR 2020

noisy image $y = x + z$ $z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

> which in turn depend on the input vector *y*. [Kadkhodaie & Simoncelli arXiv2020, NeurlPS2021]

Deep Neural Network

Diffusion models and denoising Published as a conference paper at ICLR 2020

^x ² ^R*^N* is the original image, containing *^N* pixels, *ⁿ* is an image of i.i.d. samples of Gaussian noise tariar dictribution. The noisy observation. The noisy of \mathcal{V} mean of the posterior distribution

Deep Neural Network

 (y) denoised image

 $x|y = \int x p(x|y) dx = [f^{\star}(y) = y + \sigma^2 \nabla \log p_{\sigma}^{\star}]$ **optimal denoiser**

noisy image $y = x + z$
 $z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$ both trained over noise levels in the range 2 [0*,* 10] (image intensities are in the range [0*,* 255]). The CNN performs poorly at high noise levels (Ω , far beyond the training range), whereas Ω $\mathbf{P} \left(\begin{array}{c} \mathbf{P} & \mathbf{P} \end{array} \right)$ 2γ is the similar results of γ In the past decade, purely data-driven models based on convolutional neural networks (LeCun et al., 2015 have come to dominate all previous methods in terms of performance. These models consist of performance. These models consist of performance. These models consists of performance. These models consists of performan α can value of convolutional filters, and rectified nonlinearities, which are capable of representing are diverse and powerful set of functions. Training such architectures to minimize mean square error over large databases of noisy natural-image patches achieves current state-of-the-art results (Zhang $\Omega_{\rm{obs}} = 1$ and $\Omega_{\rm{obs}}$ We assume a measurement model in which images are corrupted by additive noise: *y* = *x* + *n*, where *Minimize* ⋆

[Tweedie, via Robbins, 1956; Miyasawa, 1961]

Diffusion models and denoising Published as a conference paper at ICLR 2020

 $\mathbf r$ the noise standard deviation, , is unknown, the expectation, $\mathbf r$ distribution of . This problem is often called *blind denoising* in the literature. In this work, we study the generalization performance of CNNs *across* noise levels , i.e. when they are tested on noise $\Gamma \sigma \backslash \mathcal{Y}$ if $\Gamma \backslash \mathcal{Y}$ activation pattern of the ReLUs, the effect of the input is a cascade of linear transformations (convolutional or fully connected layers, *Wk*), additive constants (*bk*), and pointwise multiplications by a binary mask corresponding to the fixed activation pattern (*R*). Since each of

 (y) denoised image

$$
= \boxed{f^\star(y) = y + \sigma^2 \nabla \log p^\star_\sigma(y)}
$$
 optimal
denoise

[Tweedie, via Robbins, 1956; Miyasawa, 1961]

$$
\phi_{\sigma}^{\star}(y) = \int p(y|x) \, \dot{p}(x) \, dx = \int g_{\sigma}(y-x) \, \dot{p}(x) \, dx
$$

 $p_{\sigma}(y)$ is a blurred (diffused) version of $p(x)$

Diffusion models and denoising Published as a conference paper at ICLR 2020

noisy image $y = x + z$ $z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

diverse and powerful set of functions. Training such architectures to minimize mean square error over large databases of noisy natural-image patches achieves current state-of-the-art results (Zhang We assume a measurement model in which images are corrupted by additive noise: *y* = *x* + *n*, where **optimal denoiser** $f^{\star}(y) = y + \sigma^2 \nabla \log p_{\sigma}^{\star}(y)$ **learned**

 (y) denoised image

[Tweedie, via Robbins, 1956; Miyasawa, 1961]

̂

̂

Diffusion models and denoising

$$
f^{\star}(y) = y + \sigma^2 \nabla \log p_{\sigma}^{\star}(y)
$$
\n
$$
\hat{f}(y) = y + \sigma^2 \nabla \log \hat{p}_{\sigma}(y)
$$
\n
$$
\downarrow
$$
\n $$

Samples from this

Training set size: 1 10 100 1,000 10,000 100,000

Transition from memorization to generalization

Closest training example from A:

Transition from memorization to generalization

Closest training example from A:

Transition from memorization to generalization

Samples, model trained on set B **(same seed):**

Closest training example from A:

Transition from memorization to generalization

Samples, model trained on set B **(same seed):**

Closest training example from B:

Closest training example from A:

Transition from memorization to generalization

Samples, model trained on set B **(same seed):**

Closest training example from B:

Closest training example from A:

Transition from memorization to generalization

Strong generalization in LSUN bedroom dataset

Closest image from S_1 :

Generated by models trained on S_1 :

Generated by models trained on S_2 :

Closest image from S_2 :

 $N=1$

Strong generalization in BF-CNN architecture

Closest image from S_1 :

Generated by models trained on S_1 :

Generated by models trained on S_2 :

Closest image from S_2 :

 $N=1$

 $N = 10$

 $N = 100$

 $N = 1000$

 $N = 10000$

How do diffusion models generalize? What are inductive biases of the denoiser?

Denoising as shrinkage in a basis

1. Transform the noisy image to a basis where noise and signal are separable

- Classical framework for denoising:
-
- 2. Suppress the noise (shrinkage)
- 3. Transform back to pixel domain

the mean squared error : *^f* = arg min*^g ^E||^x ^g*(*y*)*||*², where the expectation is taken over some distribution over images, *x*, as well as over the distribution of noise realizations. In deep learning, the \langle τ ρ , \rangle sanction by some the network. So the optimization is over the optimization is over the optimization is over the optimization is over the optimization in the optimization is over the optimization in th parameters. If the noise standard deviation, , is unknown, the expectation must also be taken over a distribution of . This problem is often called *blind denoising* in the literature. In this work, we study the generalization performance of CNNs *across* noise levels , i.e. when they are tested on noise **frequencies same power in all** $z < z, e_k > z$

Denoising as shrinkage in a basis Fixed basis, fixed shrinkage

distribution over images, *x*, as well as over the distribution of noise realizations. In deep learning, the \langle τ ρ , \rangle sanction by some the network. So the optimization is over the optimization is over the optimization is over the optimization is over the optimization in the optimization is over the optimization in th parameters. If the noise standard deviation, , is unknown, the expectation must also be taken over a distribution of . This problem is often called *blind denoising* in the literature. In this work, we study the generalization performance of CNNs *across* noise levels , i.e. when they are tested on noise **frequencies same power in all**

Denoising as shrinkage in a basis Fixed basis, fixed shrinkage

distribution over images, *x*, as well as over the distribution of noise realizations. In deep learning, the denoising function *g* is parameterized by the weights of the network, so the optimization is over these [Donoho & Johnstone 94] distribution of . This problem is often called *blind denoising* in the literature. In this work, we study

 $f(y) = \sum_{k} \lambda_{k}(y) \langle y, e_{k} \rangle e_{k}$ Wavelet basis

Feedforward neural networks with rectified linear units (ReLUs) are piecewise affine: for a given activation pattern of the Relation of the Relation of the input is a cascalled of the input is a cascalled of formations (convolutional or fully connected layers, *Wk*), additive constants (*bk*), and pointwise Coefficients fall faster in wavelet basis. More compact representation of signal. these is affine, the entire cascade implements a single affine transformation. For a fixed noisy input **Sparse signal** *f*(*y*) = *WLR*(*W^L*¹*...R*(*W*1*y* + *b*1) + *...b^L*¹) + *b^L* = *Ayy* + *by,* (1) Easier separation between noise and signal with sparse signal

Denoising as shrinkage in a basis Fixed basis, adaptive shrinkage

distribution over images, *x*, as well as over the distribution of noise realizations. In deep learning, the denoising function *g* is parameterized by the weights of the network, so the optimization is over these [Donoho & Johnstone 94] distribution of . This problem is often called *blind denoising* in the literature. In this work, we study

Feedforward neural networks with rectified linear units (ReLUs) are piecewise affine: for a given activation pattern of the Relation of the Relation of the input is a cascalled of the input is a cascalled of formations (convolutional or fully connected layers, *Wk*), additive constants (*bk*), and pointwise Coefficients fall faster in wavelet basis. More compact representation of signal. these is affine, the entire cascade implements a single affine transformation. For a fixed noisy input **Sparse signal** *f*(*y*) = *WLR*(*W^L*¹*...R*(*W*1*y* + *b*1) + *...b^L*¹) + *b^L* = *Ayy* + *by,* (1) Easier separation between noise and signal with sparse signal

Denoising as shrinkage in a basis Fixed basis, adaptive shrinkage

^x ² ^R*^N* is the original image, containing *^N* pixels, *ⁿ* is an image of i.i.d. samples of Gaussian noise

Adaptive basis, adaptive shrinkage Denoising as shrinkage in a basis

Jacobian w.r.t. Input y $I(q) = \nabla I(q)$ $\mathsf{function} \quad \mathsf{J} \setminus \mathsf{J} \mathsf{J} \qquad \mathsf{J} \setminus \mathsf{J} \setminus \mathsf{J}$ cascades of convolutional filters, and rectified non-linearities, which are capable of representing and diverse and powerful set of functions. Training such architectures to minimize mean square error over large databases of noisy natural-image patches achieves current state-of-the-art results (Zhang We assume a measurement model in which images are corrupted by additive noise: *y* = *x* + *n*, where Locally linear $\hat{f}(a) = \nabla \hat{f}$ function: $\frac{J(9)}{9}$ Nearly symmetric

parameters. If the noise standard deviation, , is unknown, the expectation must also be taken over a distribution of . This problem is often called *blind denoising* in the literature. In this work, we study the generalization performance of CNNs *across* noise levels , i.e. when they are tested on noise [Mohan*, Kadkhodaie*, Simoncelli, Fernandez-Granda, ICLR 2020]

Eigen decomposition of Jacobian \boldsymbol{k} Shrinkage factors **Eigen basis**

Adaptive basis, adaptive shrinkage Denoising as shrinkage in a basis

$I(u) = \nabla u$ urely data-driven models based on convolutional networks (LeCun et al., 2011) $\mathsf{function} \quad \mathsf{J} \setminus \mathsf{J} \mathsf{J} \qquad \mathsf{J} \setminus \mathsf{J} \mathsf{J} \mathsf{J}$ cascades of convolutional filters, and rectified non-linearities, which are capable of representing are capable of representing and representing are capable of representing and representing are capable of representing and Locally linear function:

parameters. If the noise standard deviation, , is unknown, the expectation must also be taken over a distribution of . This problem is often called *blind denoising* in the literature. In this work, we study the generalization performance of CNNs *across* noise levels , i.e. when they are tested on noise [Mohan*, Kadkhodaie*, Simoncelli, Fernandez-Granda, ICLR 2020]

Denoising as shrinkage in an adaptive basis Adaptive basis, adaptive shrinkage

Denoising as shrinkage in an adaptive basis Geometry Adaptive Harmonic Basis (GAHBs)

1.Adaptive 2.oscillatory

 $\lambda_{53} = 0.973$

Some top Eigenvectors

hypothesis:

DNN denoisers have inductive biases towards learning GAHBs

Denoising as shrinkage in an adaptive basis Geometry Adaptive Harmonic Basis (GAHBs)

hypothesis:

- -

DNN denoisers have inductive biases towards learning GAHBs

Denoising as shrinkage in an adaptive basis Geometry Adaptive Harmonic Basis (GAHBs)

How to test this? Synthetic images

[Korostelev & Tsybakov, 1993; Donoho, 1999; Peyré & Mallat, 2008]

Geometric C^{α} images

GAHBs are optimal for denoising these

geometry adaptive harmonic basis

Geometric C^{α} images

Geometric C^{α} images

Optimal denoiser on C^{α} images has slope . *Cα α α* + 1 [Korostelev & Tsybakov, 1993] [Peyré & Mallat, 2008]

Denoising performance

Geometric C^{α} images

Optimal denoiser on C^{α} images has slope . *Cα α α* + 1

Denoising performance

five-dimensional *curved* manifold 1.Vertical position

2.Horizontal position 3.Radius/size 4.Foreground intensity 5.Background intensity

Interim summary

- Diffusion models can transition from **memorization** to **generalization** with large enough training set size
- Generalization is **strong:** two denoisers trained on nonoverlapping training sets converge to nearly the same function
- Generalization due to an **inductive bias** corresponding to shrinkage in a Geometry Adaptive Harmonic Basis (GAHB)

Kadkhodaie Z, Guth F, Simoncelli EP, Mallat S. "*Generalization in diffusion models arises from geometry-adaptive harmonic representation".*ICLR 2024.

- We are in good shape with learning densities
- Can we reduce the size of training set required for generalization?

- Image resolution
- Network size
- Complexity of image dataset

- Image resolution
- Network size
- Complexity of image dataset

Hundreds of millions or billions of parameters with **global receptive fields**

Network receptive field

Synthesis fails without global receptive fields!

Synthesis fails without global receptive fields!

Do bigger images require bigger models?

Wavelet decomposition

 $p(x_0) = p(x_1, \bar{x}_1)$

Wavelet decomposition

 $p(x_0) = p(x_1) p(\bar{x}_1 | x_1)$

Factorization of the prior

Global Local $p(x_0) = p(x_1) p(\bar{x}_1 | x_1)$

Conditional densities are local

Multi-scale wavelet representation

 x_0

Clean low-pass

Conditional denoisers

Low pass denoiser

Low-pass denoiser with global receptive field

CNN

Noisy low-pass

Denoised low-pass

Multi-scale wavelet conditional denoiser

How local?

Pixel-domain denoiser

Pixel-domain denoiser Wavelet-domain denoiser

How local?

-
-

 $RF = 43x43$ RF = 23x23 RF = 9x9 RF = 43x43 RF = 23x23 RF = 9x9

Synthesis example

Synthesis example

Total number of parameters \sim 1 million

Synthesis example

To sum up:

- We can model probability of large images with small networks.
- The global structure is captured by a global prior over a small low-pass image.
- Details can be modeled using local (Markov) conditional probability distributions in the wavelet domain.

Thank you!

Kadkhodaie, Guth, Simoncelli, Mallat, "*Generalization in diffusion models arises from geometryadaptive harmonic representation".* ICLR 2024

Kadkhodaie, Guth, Mallat, & Simoncelli, "*Learning multi-scale local conditional probability models of images.*" ICLR 2023