Image Denoising survey

Monday, November 4, 2024 8:59 AM

paper: Image Denoising: The Deep Learning Revolution and Beyond — A Survey Paper by Michael Elad, Bahjat Kawar, Gregory Vaksman
SIAM J. Imaging Science, 2023

① What is image denoising?

$y = x + v$, all vectors in $\mathbb{R}^N$ (eg. $N = n_x \times n_y$)

y: observed, x: true image, v: noise

* note we wlog use vectors not matrices

It's the simplest inverse problem ($y = A \cdot x + v$) ie. $A = I$

- Assume $v \sim N(0, \sigma^2 I)$ and σ^2 known. } "AWGN" Additive, White Gaussian Noise
  - ... the simplest noise model.
    - ↳ real cameras (CCD pixel arrays) have a tiny amount of AWGN. Mostly "shot noise" due to quantum nature of discrete photons, following a Poisson distribution (partly addressed via Anscombe or other variance stabilizing distribution, model as Gaussian if photon count high)
- Why so simple?
  - Fundamental: backbone of more realistic setups ($A \neq I$, and other noise) } indirect
  - Tractable, understood
  - Paper argues its a key ingredient in fancier setups } direct
- Goal

Build a denoiser D, return estimate

$$\hat{x} = D(y, \sigma)$$

Image Denoising survey (p. 2)

Monday, November 4, 2024 9:11 AM

- What criterion ?

For now, for good reason, use MSE (equivalently, PSNR)

*Careful: before, "MSE" is $\frac{1}{\#\text{pixels}} \sum_{i=1}^{\#\text{pixels}} (\hat{x}_i - x)^2$

and "mean" referred to this average ← actual constant of little importance if it never changes

$\|\hat{x} - x\|_2^2$

Now,

$$\text{MSE} = \mathbb{E}\,\|\hat{x} - x\|_2^2 := \int \|\hat{x} - x\|_2^2 \cdot p(x)\,dx$$

($\hat{x} = D(y, \sigma)$, $y = x + v$, so $\hat{x}$ is a function of x)

i.e. "mean" is used in a different sense.

That is, we're assigning a prior distribution to the set of images.

$p(X = $  $) = 0.013$, $p(X = $  $) = 10^{-20}$

This is the critical ingredient ...
without a prior, there's no hope

Setting $\hat{x} = y$ is the best you can do.

So this is Bayesian ?
not Frequentist ?

Note to future years:
lecture was 1 day before 2024 presidential election

Yes, sort of...

but we don't (anymore) write down a formula for $p(X)$

(these days, we'll learn it from our dataset)

Image Denoising survey (p. 3)

Monday, November 4, 2024 9:21 AM

Estimation 101 $y = x + v$, P_X is prior on X

$P_{X,Y}$ is joint on $X \times Y$

Whether Bayesian or not, Bayes' rule is true:

$$P_{X|Y=y}(x) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \frac{P_{Y|X=x}(y)\, P_X(x)}{P_Y(y)}$$

"X" refers to r.v.
"x" refers to possible value of r.v.

Maximum Likelihood Estimation (MLE) [not "Bayesian"] Very popular

Find x to maximize $p(y|x)$ ("likelihood")) or minimize negative log likelihood

Given x, $y = x + v$ i.e. $y \sim N(x, \sigma^2 I)$

So $p(y|x) = \text{const} \cdot \exp(-\|y-x\|^2/2\sigma^2)$

So

$$\hat{x}_{MLE} = \operatorname{argmin}_x -\log(p(y|x)) = \operatorname{argmin}_x \frac{1}{2}\frac{1}{\sigma^2}\|y-x\|^2 = y$$

$\hat{x}_{MLE} = y$ not useful! (if $A \neq I$ it can be useful)

Maximum 'a posteriori' Estimation (MAP) [Bayesian]

Find x to maximize $p(x|y)$

$$p(x|y) = \frac{p(y|x)\, p(x)}{p(y)}$$

$p(y)$ unimportant

$$\min_x -\log(p(x|y)) = \min_x \underbrace{-\log(p(y|x))}_{\text{MLE term}} \underbrace{-\log(p(x))}_{\rho(x)}$$

if $p(x) =$ constant (ie. uniform, uninformative) then this has no effect so MLE = MAP in this case

exploits a prior

assume $p(x) \sim e^{-\rho(x)}$

data fitting term

regularization term

Image Denoising survey (p. 4)

Monday, November 4, 2024 11:02 AM

Ex Suppose $p(x) = \text{const} \cdot \exp\left(-\rho(x)\right)$ where $\rho(x) = \frac{1}{2}\|x\|_2^2$ i.e. we prioritize small values of pixels

Then MAP:

$$\min_x \; -\log(p(y|x)) - \log(p(x))$$

$$= \min_x \; \frac{1}{2\sigma^2}\|x-y\|^2 + \frac{1}{2}\|x\|^2$$

Tikhonov regularization / Ridge regression

MMSE minimum MSE estimator [Bayesian]

Find $\hat{x}$ to minimize MSE $\mathbb{E}\|\hat{x}-x\|^2$

$$\hat{x}_{MMSE} = \mathbb{E}[x|y]$$

"Closed form" misleading

not helpful directly except in simplest cases

(2) "Classical" Techniques

... most (not all) involved choosing priors $p(x)$, or, equivalently, choosing $-\log(p(x)) =: \rho(x)$

$p(x) = c \cdot \exp(-\rho(x))$ "Gibbs distribution form"

↳ Regularization based ("years" is opinion of authors)

Table 3.1: Evolution of priors for images.

Years	Core concept	Formulae for $\rho(\cdot)$
$\sim$ 1970	Energy regularization	$\|\mathbf{x}\|_2^2$
1975-1985	Spatial smoothness	$\|\mathbf{Lx}\|_2^2$ or $\|\mathbf{D}_v\mathbf{x}\|_2^2 + \|\mathbf{D}_h\mathbf{x}\|_2^2$
1980-1985	Optimally Learned Transform	$\|\mathbf{Tx}\|_2^2 = \mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}$ (via PCA)
1980-1990	Weighted smoothness	$\|\mathbf{Lx}\|_\mathbf{W}^2$
1990-2000	Robust statistics	$\mathbf{1}^T\mu\{\mathbf{Lx}\}$ *e.g.*, Hubber-Markov
1992-2005	Total-Variation	$\int_{v\in\Omega}\|\nabla\mathbf{x}(v)\|dv = \mathbf{1}^T\sqrt{\|\mathbf{D}_v\mathbf{x}\|^2 + \|\mathbf{D}_h\mathbf{x}\|^2}$
1987-2005	Other PDE-based options	$\int_{v\in\Omega} g\left[\nabla\mathbf{x}(v), \nabla^2\mathbf{x}(v)\right] dv$
2005-2009	Field-of-Experts	$\sum_k \lambda_k \mathbf{1}^T\mu_k\{\mathbf{L}_k\mathbf{x}\}$
1993-2005	Wavelet sparsity	$\|\mathbf{Wx}\|_1$
2000-2010	Self-similarity	$\sum_k\sum_{j\in\Omega(k)} d\{\mathbf{R}_k\mathbf{x}, \mathbf{R}_j\mathbf{x}\}$
2002-2012	Sparsity methods	$\|\alpha\|_0$ *s.t.* $\mathbf{x} = \mathbf{D}\alpha$
2010-2017	Low-Rank assumption	$\sum_k \|\mathbf{X}_{\Omega(k)}\|_*$

Image Denoising survey (p. 5)

Wednesday, November 6, 2024 6:18 AM

Other classical denoising

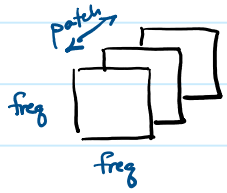
optimization method approaches ("MAP") were of great interest academically but almost never made it into Photoshop (slow, parameters to tune) ...though some variants are used in MRI ...

- non-local means (Baudes, Coll, Morel CVPR '05)

  fancy way to average over nearby (or, similar) pixels

- PDE-inspired methods

- BM3D (Dabov et al. '07) Gold standard (prior to deep learning)
  - reliable (not too much tuning required)
  - fast
  - good
  - like JPEG, takes 2D DCT of a patch
  - groups similar patches together (cube)
  - force joint sparsity on cube



- K-SVD (Elad's): sparsity in a learned "basis"/dictionary

- Improvements were starting to be marginal '09 paper, "Is Denoising Dead?"

Classical ("mathematical") methods largely ignored color images "WLOG let x be a B+W image..."
RGB, YUV, YCbCr
luma ← chroma

③ Machine Learning Techniques

- Don't explicitly pick a (simple) prior $p(x)$... implicitly learn one from data. ie. We can learn a denoiser specialized for astronomy, etc.

- Simplest approach: supervised learning w/ hybrid real/synthetic data
  - ML needs a lot of data
  - We do have a lot of image data $\{x_k\}_{k=1}^{M}$ ← "real" data
    - To get labeled data, we can artificially add noise

      $y_k = \underbrace{x_k}_{\text{real}} + \underbrace{v_k}_{\text{synthetic}}$ so we have $\{(y_k, x_k)\}_{k=1}^{M}$

      - Simplest case: σ = constant (see paper for some info on case when we want to do "blind denoising", i.e., σ unknown)
  - Define a neural net architecture
  - Train using ERM framework (Empirical Risk Minimization) ... as usual

    What loss function $\text{dist}(x, \hat{x})$? $\hat{x} = D_\theta(y, \sigma)$

    Usually MSE $\|x - \hat{x}\|_2^2$

- Variants:
  - σ unknown
  - No access to clean images $\{x_k\}$ → eg. if you want to tailor to a very specific type of image. "Unsupervised"
  - Beyond AWGN
  - Boosting (transfer learning) to specialized datasets (MRI, astronomy, cartoon...)
  - SSIM loss function ~. Target perceptual loss

Image Denoising survey (p. 6)

Wednesday, November 6, 2024 6:19 AM

- Some ML architectures are inspired by (or "unfolding") classical methods
  - Elad calls it "unrolling", slightly similar but not the same as the unrolling we'll discuss when $A \neq I$ "LISTA" Gregor, Le Cun ICML '10
  - "Deep K-SVD" and "LIDIA" (like BM3D), see §5
- "Deep image prior": CNN trained on a single image to fit itself (like INR a little)
  CNN architecture implicitly regularizes.
  Do early stopping! (some disadvantage as classical Lucy-Richardson deblurring!)

(4) Beyond Image Denoising The exciting part

(A) Denoisers ($A = I$) can directly help general inverse problems ($A \neq I$)

$$y = Ax + v$$

(B) Denoisers can directly help generative models to synthesize new images (with applications to (A))

(C) We can go beyond "MAP" and "MMSE" to get a denoised image w/ much better perceptual quality

(A) $y = Ax + v$, A linear Ex: deblurring, inpainting, demosaicing, superresolution, tomography (CT, MRI, ...)

MAP $\hat{x}_{MAP} = \arg\min_x \frac{1}{2}\frac{1}{\sigma^2}\|Ax - y\|^2 + c \cdot \rho(x)$ $\rho(x) = -\log(p(x))$ as usual

Plug-n-Play idea (Venkatakrishnan, Bouman, Wohlberg, '13 ... though I think it goes back earlier)

Suppose $\rho(x) = \|x\|_1$.

How would people solve the optimization problem?

Two common approaches: (1) ADMM or (2) proximal gradient descent

both work equally well to motivate key idea.

Paper uses (1), I'll use (2)

Image Denoising survey (p. 7)

Wednesday, November 6, 2024 7:04 AM

"Proximal Gradient Descent"

Goal: iteratively minimize $\underbrace{f(x),\ \nabla f(x) = \frac{1}{\sigma^2} A^T(Ax-y)}_{\frac{1}{2\sigma^2}\|Ax-y\|^2} + c\cdot\rho(x)$ via:

$$x_{k+1} = \arg\min_x \underbrace{f(x_k) + \langle \nabla f(x_k), x - x_k\rangle}_{\text{Taylor series of } f} + \underbrace{\frac{1}{2\eta}\|x-x_k\|^2}_{\text{don't go too far! We don't trust Taylor Series}} + c\cdot\rho(x)$$

(no approximation)

(if $\rho(x)\equiv 0$, ... $= \overbrace{x_k - \eta\cdot\nabla f(x_k)}^{z_k}$ which is gradient descent!)

$$\star = \arg\min_x \frac{1}{2}\|x - \underbrace{z_k}_{\text{separable}}\|^2 + c\cdot\rho(x)$$

(so again, $\rho = 0$, ... $= z_k$ clearly)

$$= \text{prox}_{c\cdot\rho}(z_k).$$

This generalizes projected gradient descent

Classical choices of ρ were made such that prox_ρ is easy to compute

Ex: $\rho(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$ is separable so whole prox problem is separable ($\min_{x_1,x_2} f(x_1) + g(x_2)$ you can do separately?)

Wait a minute... $\star$ is the MAP estimator for denoising ($A = I$)! We (approximately) linearize, denoise, then repeat.

Insight Rather than choose ρ and then calculate prox_ρ, lets choose prox_ρ (our denoiser $= D(\cdot,\sigma)$) and leave ρ implicitly defined!

- Advantage: leverage modern denoisers tailored to realistic/specialized datasets
- Disadvantage: math. In interesting cases, we have no reason to believe it must converge...

Image Denoising survey (p. 8)

Wednesday, November 6, 2024 7:18 AM

RED idea "Regularization by Denoising" (Romano, Elad, Milanfar '17) award-winning

Somewhat similar spirit to Plug-n-Play

$$D(y;\sigma) = \operatorname{argmin}_x \frac{1}{2\sigma^2}\|x-y\|^2 + \rho(x)$$

ρ: implicitly defined by D now

RED instead defines ρ via D in a different way: $\rho(x) := x^T(x - D(x;\sigma))$

with some motivation.

Under conditions on D (smooth, "passive", local homogeneity) then this ρ is a convex function!

So we get back nice math, convergence.

Exploring this idea further eg. which denoiser to use? which σ?

$$x_{MAP} = \operatorname{argmin}_x \frac{1}{2}\|Ax-y\|^2 + \underbrace{\sigma^2 \cdot c}_{\tilde{c}} \cdot \overbrace{-\log(p(x))}^{\text{aka } \rho}$$

if ρ is differentiable, could solve via gradient descent

$$x_{k+1} = x_k - \eta\left(A^T(Ax_k - y) - c\cdot \underbrace{\nabla_x \log(p(x))}_{\text{"score function" in statistics}}\right)$$

Going back to 1960's, it's known — optimal MMSE denoiser $E(x|y)$

$$\nabla_y \log(p(y)) = \frac{D(y,\sigma_0) - y}{\sigma_0^2}$$

This fits w/ RED definition (if we had optimal denoiser), and also suggests using a denoiser trained w/ MSE loss.

[RED Algo.] Parameters: stepsize η, weight c, noise level σ

$$x_{k+1} = x_k - \eta\left(A^T(Ax_k - y) + c\cdot(x_k - D(x_k,\sigma))\right)$$

Image Denoising survey (p. 9)

Wednesday, November 6, 2024 7:39 AM

(B) Generative AI : image synthesis via denoisers

Idea: learn $G_\theta(z)$ to generate an image $x \sim p(x)$

↑ random input, $z \sim N(0, I)$

like (multivariate) inverse CDF sampling... but we won't have explicit formulas

Ex: Variational Autoencoders (VAE)

Ex: Generative Adversarial Networks (GAN) ← were dominant method for generative AI in image denoising

Ex: Diffusion Models (now dominant)

"Denoising diffusion probabilistic models" Ho, Jain, Abbeel. NeurIPS '20

Sohl-Dickstein et al. ICML '15

Song + Ermon, NeurIPS '19 or "Score-based generative models"

How to sample from $p(x)$? (assuming you know $p(x)$)

ex:

- rejection sampling. Very inefficient, essentially inapplicable in higher dim.
- better: Markov Chain Monte Carlo MCMC

  Apply a Markov chain ($x_{k+1} | x_k \perp x_{k-1}, x_{k-2}, \ldots$)

  so that its stationary distribution is $p(x)$

Welling, Teh '10 popularized in ML (for other purposes)

- Our models are based on a variant of MCMC:

Langevin Dynamics originally from physics $z_t \sim N(0, I)$

$$x_{t+1} = x_t + \alpha \underbrace{\nabla_x \log(p(x_t))}_{\text{score}} + \sqrt{2\alpha}\, z_t$$

α = stepsize; z_t keeps it from being MAP

and recall $-\nabla_x \log(p(x))$ (or $\nabla_x p(x)$) is implicitly done via our denoiser, $\nabla_x p = 2(x - D(x, \sigma_0))$

one motivation: stochastic differential eq'n

$$dX(t) = -\nabla p(X(t))\, dt + \sqrt{2/\beta}\, dB(t)$$

β = temperature, $B(t)$ = Brownian motion

then discretize, "Euler-Maruyama" Converges to $X \sim p$

(under suitable conditions) this gives samples $x_t \sim p$

but need tricks to make it practical:

- we initialize w/ some x_0, but $p(x_0) = 0$ (or very close)

  and $\nabla \log(p(x_0))$ will explode to $-\infty$ ↑ most matrices are not images!

  Solution "blur", i.e. add fudge factor, equivalent to Tikhonov style...

- slow convergence

  Solution: "annealed Langevin Dynamics" ALD (Song, Ermon '19)

  Start w/ large σ, run for a while, then decrease σ, ..., until reach target σ

Image Denoising survey (p. 10)

Wednesday, November 6, 2024 10:25 AM

Algo: ALD Parameters $\{\sigma_i\}_{i=0}^{L}$, ε, T

$x_0 \sim N(0, I)$ // size of (vectorized) image

for $i = 1, \ldots, L$

$\alpha_i = \varepsilon \cdot (\sigma_i / \sigma_L)^2$

for $t = 1, \ldots, T$

$z_t \sim N(0, I)$

$x_t \leftarrow x_{t-1} + \alpha_i / \sigma_i^2 \left(D(x_{t-1}, \sigma_i) - x_{t-1} \right) + \sqrt{2\alpha_i}\, z_t$ (denoiser)

$x_0 \leftarrow x_T$

end

end

return x_0

(c) "High perceptual quality image recovery" : beyond point estimators

"MAP" is Bayesian ... but not fully in the Bayesian spirit, since it's a point estimator

MMSE is $\underset{\hat{x}}{\operatorname{argmin}}\, \mathbb{E} \|x - \hat{x}\|^2$ (i.e. over $x \sim p$)

Issue: loss function is MSE, and, averaging over all possible inputs x

Ex: I want to own a dog (one dog), my partner doesn't.

MMSE solution is to own half a dog

Consequence: MMSE results $\hat{x}$ might not look realistic

"perception – distortion tradeoff"

realistic? MMSE

One approach: $\hat{x}$ is a sample from $p(x|y)$ Conceptual shift: random output

- Can be done via GANS (generator vs. discriminator)
- or via diffusion models

See Eq (9.2) for details... it's slightly more complicated.

Bayes' Rule

$$\nabla_x \left(\log \left(p(x|y) = \frac{p(y|x)p(x)}{p(y)} \right) \right) = \nabla_x p(y|x_t) + \nabla_x p(x)$$

($p(y)$: irrelevant)

Conditional score function

$\nabla_x p(y|x_t)$: via Eq. (9.4) in Elad et al.

$\nabla_x p(x)$: usual score fcn (via denoiser)

- now MMSE denoiser used to create realistic denoiser!

Image Denoising survey (p. 11)

Wednesday, November 6, 2024 10:48 AM

This high-perceptual quality denoiser can be extended to solve general inverse problems ($y = Ax + v$) as well.

See §9.2. Basically some SVD tricks w/ A, and properties of multivariate normal distribution. "SNIPS" approach

Other approaches exist

Some handle when "A" isn't explicitly defined ("stylization", "JPEG deblocking")

Fancier extensions: text-to-image generation (DALL-E)