MTD421: Bachelor's Thesis Project
Final Presentation
Deep Neural Network approximation for Image Denoising

Bhumika Chopra
Supervised by: Prof. Sivananthan Sampath

Apr 13, 2022

Problem Statement

Image Denoising:
It is the process of removing noise from a noisy image, so as to restore the true image

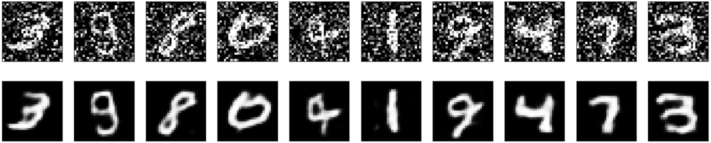


Figure: Example of image denoising

Recap

- Deep Neural Networks
- CNNs (Convolutional Neural Networks)
- Dimensionality Reduction
- Johnson-Lindenstrauss Lemma
- GANs: a brief introduction and training process

Noise Models

1. Additive Noise

$$w(x,y) = s(x,y) + n(x,y)$$

2. Multiplicative Noise

$$w(x,y) = s(x,y) * n(x,y)$$

$s(x,y)$ is the original image intensity at location (x,y)
$n(x,y)$ is the noise at location (x,y)

Classical Methods

1. Spatial domain filtering
   make use of low pass filter to filter high frequency spectrum. e.g., average/median filter
2. Transform domain filtering
   transform image from one domain to another and then apply spatial filtering. e.g., BM3D
3. Wavelet Thresholding method
   apply wavelet transform to a signal and remove coefficients below a certain threshold. e.g., hard or soft thresholding

Stochastic Gradient Descent(SGD)

- optimize a convex function F over a convex domain W
- simple and highly scalable
- commonly used in DNNs to learn the parameters
- Iterates of the SGD algorithm are defined as follows:

$$w_{t+1} = \prod_{W}(w_t - \eta_t \hat{g}_t)$$

where $g_t \in \partial F(w_t)$[1] and $\mathbb{E}\hat{g}_t = g_t$

[1] set of subgradients of F

Convergence of Stochastic Gradient Descent

- F is a non-smooth convex/strongly-convex function over a closed convex domain W
- $W \leftarrow$ a subset of some Hilbert space with induced norm $||.||$
- $w^* = \underset{w}{\arg\min} F$
  minimize optimization error $:= F(\bar{w}) - F(w^*)$

Convergence of Stochastic Gradient Descent

Theorem 1. Suppose F is λ- strongly convex [1] and that $\mathbb{E}[||\hat{g}_t||^2] \leq G^2 \; \forall \; t$. Consider SGD with step sizes $\eta_t = \frac{1}{\lambda t}$. Then for any $T > 1$, it holds that

$$\mathbb{E}[F(w_T) - F(w^*)] \leq \frac{17G^2(1 + log(T))}{\lambda T}$$

[1] $F(w') \geq F(w) + \langle g, w' - w \rangle + \frac{\lambda}{2}||w' - w||^2 \; \forall \; w, w' \in W$ and any g

Convergence of Stochastic Gradient Descent

Proof: By convexity of W(and definition of SGD iterates), for any $w \in W$:

$$\mathbb{E}[||w_{t+1} - w||^2] = \mathbb{E}[||\prod_W (w_t - \eta_t \hat{g}_t) - w||^2]$$

$$\leq \mathbb{E}[||w_t - w||^2] - 2\eta_t \mathbb{E}[\langle g_t, w_t - w\rangle] + \eta_t^2 G^2$$

- Rearranging and summing over $t = T - k, ..., T$
- Construct a lower bound for $\langle g_t, w_t - w\rangle$ using $F(w_t) - F(w)$ by the definition of subgradient g_t
- Substitute $\eta_t = \frac{1}{\lambda t}$

We get:

$$\mathbb{E}\Big[\sum_{t=T-k}^{T} \Big(F(w_t) - F(w)\Big)\Big] \leq \frac{\lambda(T-k)}{2}\mathbb{E}[||w_{T-k} - w||^2] +$$

$$\frac{\lambda}{2}\sum_{t=T-k+1}^{T} \mathbb{E}[||w_t - w||^2] + \frac{G^2}{2\lambda}\sum_{t=T-k}^{T}\frac{1}{t}$$

Convergence of Stochastic Gradient Descent

- Put $w = w_{T-k}$
- Using a result from Rakhlin et. al., for any $t \geq T-k$

$$\mathbb{E}[||w_t - w_{T-k}||^2] \leq \frac{16G^2}{\lambda^2(T-k)} \leq \frac{32G^2}{\lambda^2 T}$$

- define S_k, expected average value of the last (k+1) iterates

$$S_k = \frac{1}{k+1}\sum_{t=T-k}^{T}\mathbb{E}[F(w_t)]$$

also using:

$$k\mathbb{E}[S_{k-1}] = (k+1)\mathbb{E}[S_k] - \mathbb{E}[F(w_{T-k})]$$

We get:

$$\mathbb{E}[S_{k-1}] \leq \mathbb{E}[S_k] + \frac{G^2}{2\lambda}\Big(\frac{32}{kT} + \sum_{t=T-k}^{T}\frac{1}{k(k+1)t}\Big)$$

Convergence of Stochastic Gradient Descent

- Sum over $k = 1, .., \lfloor \frac{T}{2} \rfloor$
- Upper bound the terms on the right

$$\sum_{k=1}^{\lfloor T/2 \rfloor} \frac{1}{k} \leq 1 + log(\frac{T}{2})$$

$$\sum_{k=1}^{\lfloor T/2 \rfloor} \sum_{t=T-k}^{T} \frac{1}{k(k+1)t} \leq \frac{1 + log(T)}{T}$$

We get:

$$\mathbb{E}[F(w_T) - F(w^*)] \leq \frac{17G^2(1 + log(T))}{\lambda T}$$

Hence proved.

- $O(\frac{log(T)}{T})$ convergence

Convergence of Stochastic Gradient Descent

Theorem 2. Suppose that F is convex, and that for some constants D, G, it holds that $\mathbb{E}[||\hat{g}_t||] \leq G^2$ for all t, and $\sup_{w,w' \in W}||w - w'|| \leq D$. Consider SGD with step sizes $\eta_t = \frac{c}{\sqrt{t}}$ for some constant $c > 0$. Then for any $T > 1$, it holds that

$$\mathbb{E}[F(w_T) - F(w^*)] \leq \left(\frac{D^2}{c} + cG^2\right)\frac{2 + log(T)}{\sqrt{T}}$$

Convergence of Stochastic Gradient Descent

Proof: Similar to proof of Theorem 1, we sum

$$\mathbb{E}[\langle g_t, w_t - w\rangle] \leq \frac{\mathbb{E}[||w_t - w||^2]}{2\eta_t} - \frac{\mathbb{E}[||w_{t+1} - w||^2]}{2\eta_t} + \frac{\eta_t G^2}{2}$$

over $t = T - k, .., T$

- Put $\eta_t = \frac{c}{\sqrt{t}}$
- Given that: $||w - w'|| \leq D \;\; \forall w, w' \in W \Rightarrow ||w - w'||^2 \leq D^2$

We get:

$$\mathbb{E}\Big[\sum_{t=T-k}^{T} F(w_t) - F(w_{T-k})\Big] \leq \Big(\frac{D^2}{2c} + cG^2\Big)\frac{k+1}{\sqrt{T}}$$

Convergence of Stochastic Gradient Descent

Use:

- $S_k = \frac{1}{k+1}\sum_{t=T-k}^{T} \mathbb{E}[F(w_t)]$, expected average value of the last (k+1) iterates
- $k \cdot \mathbb{E}[S_{k-1}] = (k+1)\mathbb{E}[S_k] - \mathbb{E}[F(w_{T-k})]$

We get:

$$\mathbb{E}[S_{k-1}] \leq \mathbb{E}[S_k] + \left(\frac{D^2}{2c} + cG^2\right) \cdot \frac{1}{k\sqrt{T}}$$

- Sum the above ineq. over $k = 1, 2, .., T-1$

$$\mathbb{E}[F(w_T)] = \mathbb{E}[S_0] \leq \frac{1}{\sqrt{T}}\left(\frac{D^2}{2c} + cG^2\right)\sum_{k=1}^{T-1}\frac{1}{k} + \mathbb{E}[S_{T-1}]$$

Convergence of Stochastic Gradient Descent

Upper bounding the terms on the RHS:

- $$\sum_{k=1}^{T-1} \frac{1}{k} \leq (1 + log(T))$$
- $$\mathbb{E}[S_{T-1}] - F(w^*) \leq \left(\frac{D^2}{c} + cG^2\right) \frac{1}{\sqrt{T}}$$

We get:

$$\mathbb{E}[F(w_T) - F(w^*)] \leq \left(\frac{D^2}{c} + cG^2\right) \cdot \frac{2 + log(T)}{\sqrt{T}}$$

Hence proved.

- $O(\frac{log(T)}{\sqrt{T}})$ convergence

Minimax Loss function

We define, the loss function

$$L = \min_{G} \max_{D} V(D, G)$$

$$V(D, G) = E_{x \sim p_{data(x)}}[\log(D(x))] + E_{z \sim p_z(z)}[\log(1 - D(G(z))]$$

KL divergence and JS divergence

$$D_{KL}(P||Q) = \sum_{x \sim X} P(x) \log \frac{P(x)}{Q(x)}$$

if P and Q are discrete distributions, or

$$D_{KL}(P||Q) = \int_{-\inf}^{inf} P(x) \log \frac{P(x)}{Q(x)} dx$$

if P and Q are continuous distributions.

$$JSD(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

here

$$M = \frac{1}{2}(P + Q)$$

Convergence of GANs

Proposition - Given the generator, G the optimal discriminator is given by-

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

Convergence of GANs

Proof: Given a generator, G, the discriminator will try to maximize, V(D,G).

$$V(D,G) = \int_x p_{data}(x)\log(D(x))dx + \int_z p_z(z)\log(1-D(G(z)))dz$$

$$V(D,G) = \int_x p_{data}(x)\log(D(x)) + p_g(x)\log(1-D(x))dx$$

Hence, we get that $V(G,D)$ is maximum for $D_G^*(x)$ which is given by

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}$$

Convergence of GANs

Proposition - Jensen-Shannon divergence is always non-negative.

Convergence of GANs

Proof:

$$-D_{KL}(P||Q) = \sum_{x \sim X} P(x) \log \frac{Q(x)}{P(x)}$$

$$-D_{KL}(P||Q) \leq \sum_{x \sim X} P(x)(\frac{Q(x)}{P(x)} - 1) = \sum_{x \sim X} Q(x) - \sum_{x \sim X} P(x) = 1 - 1 = 0$$

$$D_{KL}(P||Q) \geq 0$$

Convergence of GANs

Theorem The global minimum of the virtual training criterion $C(G)$ is achieved only when $p_g = p_{data}$ and the minimum value is $-\log 4$.

$$C(G) = E_{x \sim p_{data}}[\log(D_G^*(x))] + E_{x \sim p_g}[\log(1 - D_G^*(x))]$$

$$C(G) = E_{x \sim p_{data}}[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}] + E_{x \sim p_g}[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}]$$

Convergence of GANs

Proof: At $p_g = p_{data}$ we get $C(G) = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4$. To see that this is the least possible value of $C(G)$, reached only for $p_g = p_{data}$, observe that

$$C(G) = -\log 4 + D_{KL}(p_{data}||\frac{p_{data} + p_g}{2}) + D_{KL}(p_g||\frac{p_{data} + p_g}{2})$$

$$C(G) = -\log 4 + 2JSD(p_{data}||p_g)$$

Hence, $C(G)$ =-$\log 4$ is the global minimum of C(G) and is attained only when is $p_g = p_{data}$, i.e., the generative model perfectly replicates the data generating process.

Convergence of GANs

Theorem Given G and D have enough capacity and at each step of the algorithm, D is allowed to reach its optimum and p_g is updated so as to minimize $C(G)$ then p_g converges to p_{data}.

$$C(G) = E_{x \sim p_{data}}[\log(D_G^*(x))] + E_{x \sim p_g}[\log(1 - D_G^*(x))]$$

$$C(G) = V(G, D_G^*) = U(p_g, D_G^*)$$

Convergence of GANs

Proof: [1] We get that, $C(G)$ is convex in p_g.
Why will the gradient exist?
'The subderivatives of a supremum of convex functions include the derivative of the function at the point where the maximum is attained.'
Make a plot of $U(p_g, D)$ vs p_g and at each point let the discriminator achieve its optimum value D^*_G.
This implies that if we compute the set of derivatives of $U(p_g, D^*_G)$ w.r.t p_g, it will include all the partial derivatives of $U(p_g, D)$ w.r.t p_g at the locations where $U(p_g, D)$ is maximum for a given p_g.
This is equivalent to computing the gradient descent update for p_g at at the optimal D^*_G given the generator, G. We have proven that $C(G)$ is convex and attains a unique global minimum when $p_g = p_{data}$. Hence, with sufficiently small updates to p_g it will converge to p_{data}. This concludes the proof.

[1] The function $F(x) = a \log x + b \log(1 - x)$ where $a, b \in [0, 1]$ is convex.

Application of Image Denoising

Aim - Given an image of a person with a mask, remove the mask and return an output image of the same person without the mask.

UNet architecture

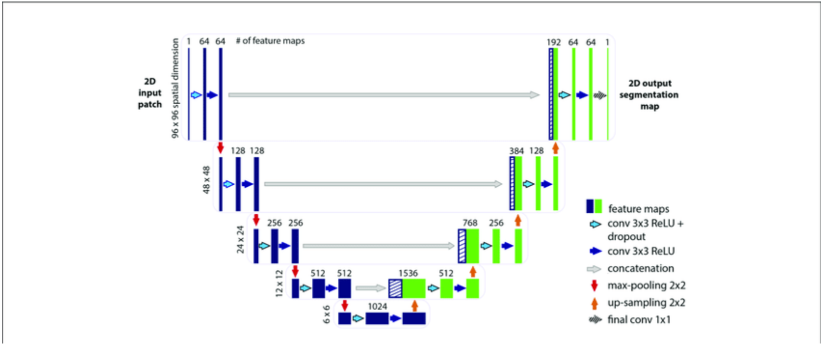


Figure: U-Net

- UNets perform classification of each pixel in the input image
- Make use of skip connections.

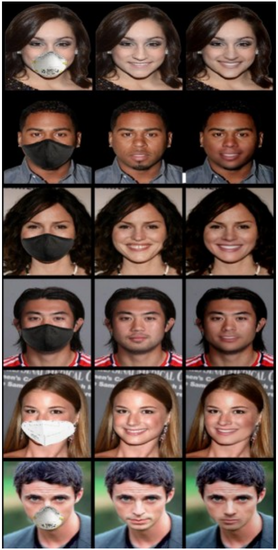
Loss function

Structural Similarity Index Measure (SSIM) compares luminance, contrast and structure between the two images and returns a weighted combination of the three values between 0 and 1.

Training details

We trained a UNet+RESNET based model on the CelebA dataset. On fine-tuning the parameters, we obtained the best results when the input image has size 256x256x3, kernel size is 3x3, number of epochs is 20, batch size is 12 with ADAM optimizer.

Results



Improvements

1. The generated images under the mask are not as sharp and refined as normal human face images.
2. The model is designed to work on front-facing human faces and does not work properly on sideways profiles.
3. Improvements in human face points detection (dlib) and application on face can drastically improve results.
4. Portability of the model.

References

1. Fan L et al. “Brief review of image denoising techniques”. In: Vis. Comput. Ind. Biomed. Art 2, 7 (2019)
2. Shamir, Ohad and Zhang, Tong, ”Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes”, In: International Conference on Machine Learning (2013)
3. Ian J. Goodfellow et al. “Generative Adversarial Networks”. In: Advances in Neural Information Processing Systems (2014)
4. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convo- lutional Networks for Biomedical Image Segmentation”. In: Medical Im- age Computing and Computer-Assisted Intervention – MICCAI (2015), pp. 234–241.
5. https://github.com/strvcom/strv-ml-mask2face
6. Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: Computer Vision and Pattern Recognition - CVPR (2015).

Thank you