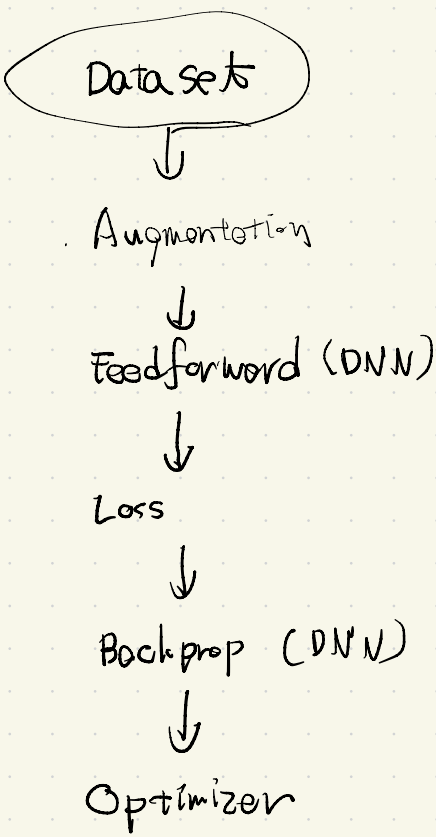
ランダム行列と深層学習　3/23, 24

Day 1

深層学習速習

- framework of deep learning
- examples



Framework

$\mathcal{D} \subset X \times Y \subset \mathbb{R}^n \times \mathbb{R}^{n_0}$

$\mathcal{D}$: finite

μ: prob. dist. on $X \times Y$

$L : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$

Find $f : \mathbb{R}^n \to \mathbb{R}^{n_0}$

s.t. mininize $\int_{(x,y) \in X \times Y} L(f(x), y)\, \mu(dxdy)$ (μ: unknown)

with information $\mathcal{D}$.

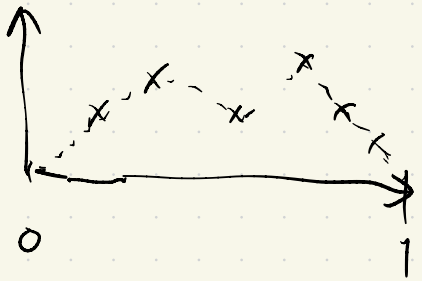
- $\mathcal{D}$ で $X \times Y$ と μ をうまく approx
- $f = f_\theta$ $(\theta \in \Theta \subset \mathbb{R}^p)$ parametrize
- 積分 を approx
- "minimize" を approx

Dataset

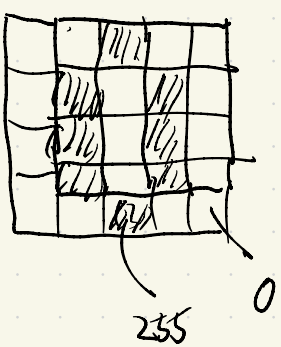
CV

1-dim

$X, Y = [0,1] \subset \mathbb{R}$



Computer Vision



RGB

$H \times W \times C$

$C = 3$

→ vectorize

HWC-dim vector

output : $0 \rightsquigarrow \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = e_1 \in \mathbb{R}^{10}$

↙ 0 ~ 9 → $e_1 \sim e_{10}$

MNIST : 手寫數字.

Natural Language Processing

(NLP)

I am eating pizza.

tokenize

→ [8, 10, 201, 1018]

vector

→ $x \in X$.

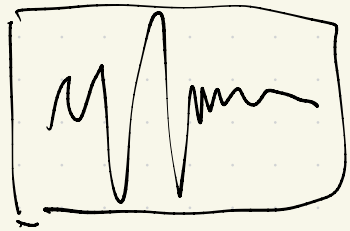
mask

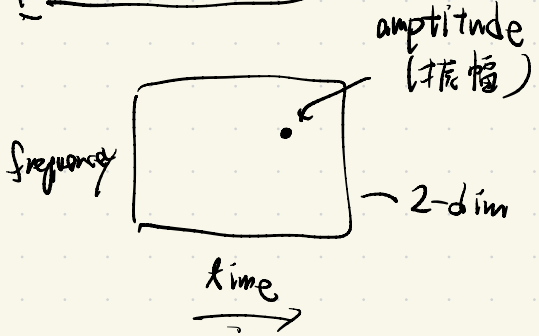
x : I am [mask] pizza

y : I am eating pizza.

Audio

Audio



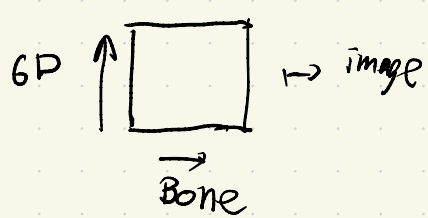


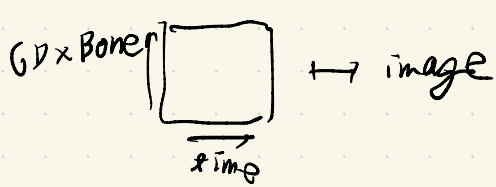
画像とにデータを作る

→

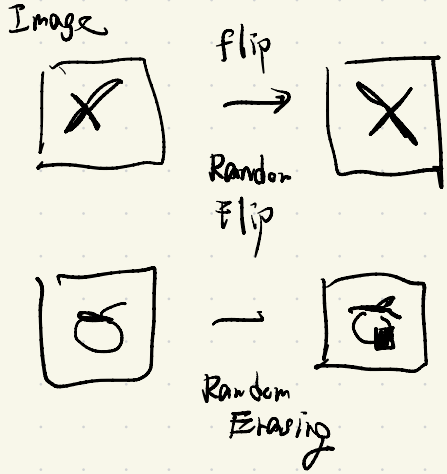
3D Bone Data

3D(+3D) × Bone





Augmentation



data に依存に増やす.

Model

$f : \mathbb{R}^n \longrightarrow \mathbb{R}^{n_0}$

$f = f_\theta$.

$f_\theta(x) = Wx + b$

$\theta = (W, b)$

$\begin{cases} W \in \mathbb{R}^{n_0 \times n} \\ b \in \mathbb{R}^{n_0} \end{cases}$

Multilayer-Perceptron

Aug 2

(MLP).

$x_0 = x$

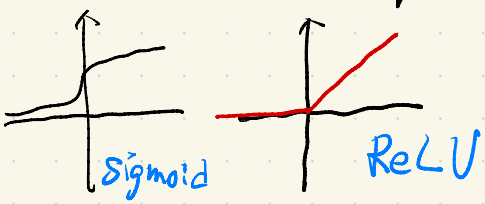
$$\begin{cases} h_\ell = W_\ell x_{\ell-1} + b_\ell \\ x_\ell = \varphi(h_\ell) \end{cases} \quad (\ell = 1, \cdots, L)$$

$L \in \mathbb{N}$

φ : activation ftn.

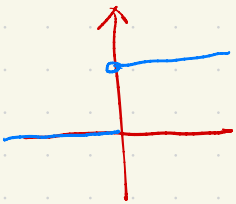
$\varphi : \mathbb{R} \to \mathbb{R}$

- conti.
- diff'! except for finite pts.



$f_\theta(x) = h_L$

$\theta = (W_1, b_1, \cdots W_L, b_L)$



✡ L : Loss ftn

↳ L2-ftn

↳ $L(f(x), y) = \|f(x) - y\|^2$

↳ Softmax Cross Entropy ...

↳ L1- ...

Backpropagation

L_Θ

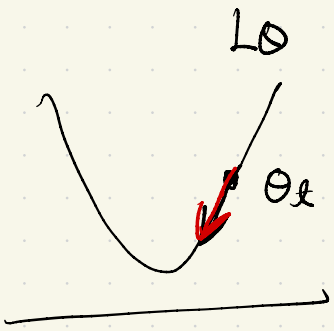
$$\text{minimize}_{\Theta \in \Theta} \sum_{(x,y) \in D} \|f_\Theta(x) - y\|^2 + \lambda \|\Theta\|^2$$

weight-Decay

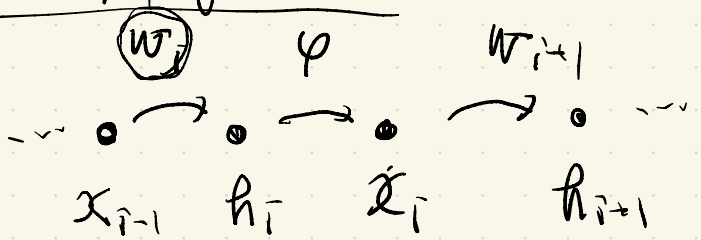
Gradient Descent

$$\Theta_{t+1} = \Theta_t - \eta_t \cdot \nabla_\Theta L_{\Theta_t}$$

$\eta_t > 0$: learning rate



Back propagation



$$\frac{\partial L_\theta}{\partial W_i} = \frac{\partial L_\theta}{\partial h_i} \cdot \frac{\partial h_i}{\partial W_i}$$

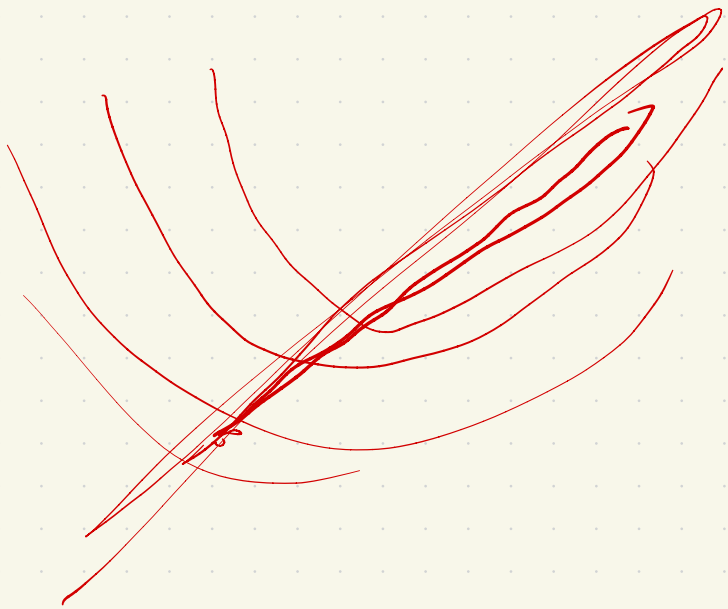
$\frac{\partial L_\theta}{\partial W_{i-1}}$ $\frac{\partial L_\theta}{\partial W_1}$

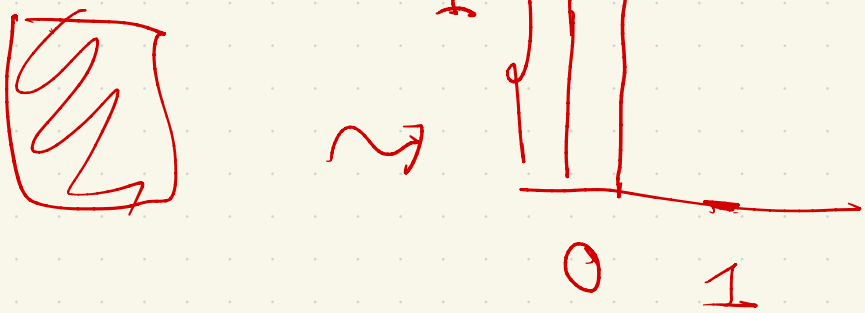
Summary

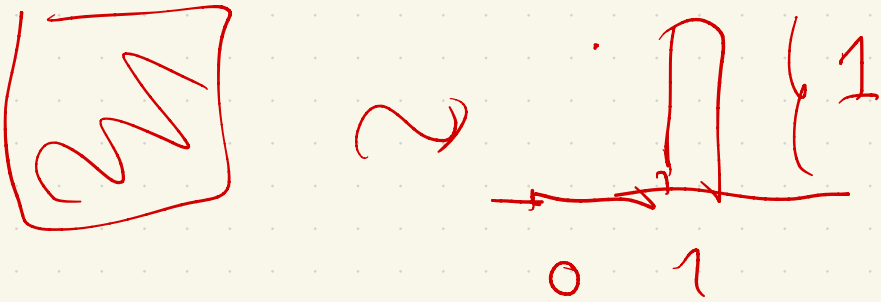
- Dataset
- Augmentation
- NN
- Forward
- Loss
- Back propagation

optimizer.

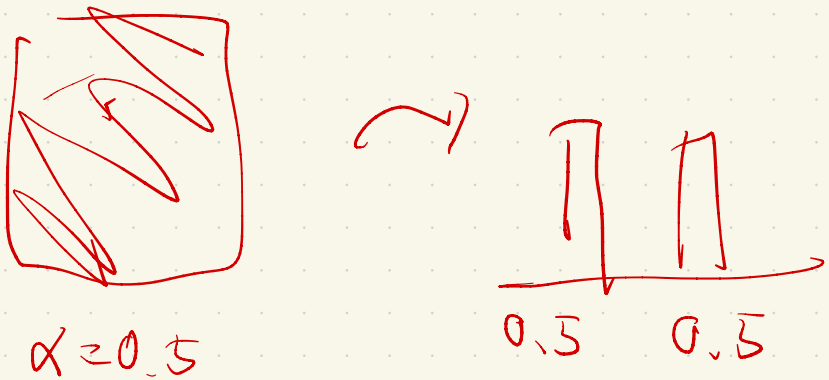
$$y = ABx$$







Mixup



Day2 Signal Propagation

くり返し アルゴリズム

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta L_\theta$$

$$L_\theta = \sum_{x,y} L(f_\theta(x), y) + \boxed{\quad}$$

$$\theta_t = (w_1^{(t)}, \cdots, w_L^{(t)}, b_1^{(t)}, \cdots, b_L^{(t)})$$

Initialization が大事.

$$W_{\ell,ij} \sim N\left(0, \frac{1}{n}\right) \quad \text{i.i.d.}$$

$(i,j = 1, \cdots, n)$

$(\ell = 1, \cdots, L)$

$$b_{\ell,i} \sim N\left(0, \frac{1}{n}\right) \quad \text{i.i.d}$$

$(i,j = 1, \cdots, n)$

$(\ell = 1, \cdots, L)$

$$\Rightarrow \quad \mathbb{E}\left[\|h_\ell\|^2\right] = \mathbb{E}\left[\|x_{\ell-1}\|^2\right] + 1.$$

$x_i = \varphi(h_i)$ で 2倍を打ち消す.

ランダム行列

: 行列値 (R-valued)

- 確率変数.

$W_{ij} \sim N(0, \sigma^2)$ $(i,j=1,\cdots,n)$

i.i.d

Ginibre Random Matrix.

$W \sim$ uniformly sample from $O(n)$

Haar Orthogonal Random Matrix

W : Ginibre

$$W = U D V$$

$U, V \in O(n)$ (Haar Orthogonal)

$$D = \mathrm{diag}\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

$\lambda_1 \geq \cdots \geq \lambda_n$: singular value of W

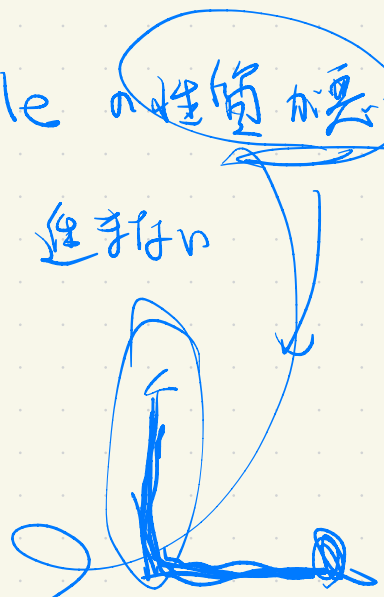
($W^T W$ の固有値の square root)

Back propagation

$$\underline{\underline{J}} = \frac{\partial h_L}{\partial x_0} = W_L D_{L-1} W_{L-1} D_{L-2} \cdots W_1$$

J の singlular vale の性質が悪いと

↝ 学習が進まない



Expolsion / Vanishy gradent

→ J が orthogonal に近ければよい

→ Jの特異値が1周辺に集中しているような状況をつくれませんか?

↳ Dynamical Isotometry
(Saxe, Pennington)

$$J = W_L D_{L-1} W_{L-1} \cdots W_1$$

$$D_\ell = \varphi'(h_\ell)$$

D_ℓ ← Random Matrix.

$\varphi'(h_\ell)$ = Random vector.

RMの積のsigular valueをcontroleできますか?

→ free probability theory

---

prob. theory

random variable X

prob. distribution μ_X

expectation / variance

$\mathbb{E}[X^n]$

Noncommutative Probability

$a \in A$ (element of algebra)

$n \mapsto \mathrm{tr}(a^n)$

$\mathrm{tr}(a^n)$

Independence

" $\mu_{X,Y} = \mu_X \otimes \mu_Y$ "

free product

free probability theory

RM eigenvale $X = W^T W$

~~singular~~ value of

$\{ n \mapsto tr(X^n) \}$

$\rightsquigarrow$ eigenvalue

固有空間 $U \sim O(n)$

uniform

$\Leftrightarrow$ non-commutative with orther RM

i.e. $U, V \sim$ unifon on $O(n)$

U, V : non-commutative

$\rightsquigarrow$ "asymtotically" free

ふつうの prob.

$X \perp Y \Rightarrow \mu_{X+Y} = \mu_X * \mu_Y$

$$\mathbb{E}[e^{it(X+Y)}]$$

$$= \mathbb{E}[e^{itX}]\,\mathbb{E}[e^{itY}]$$

W_1, W_2 : asym. free

$$\mu_{W_1+W_2} \simeq \mu_{W_1} \boxplus \mu_{W_2}$$

free additive convolution

$$\mu_{W_1 W_2} \simeq \mu_{W_1} \boxtimes \mu_{W_2}$$

$$J = D_{L-1} W_{L-1} \cdots D_1 W_1$$

$$\Rightarrow \mu_{J^T J} \cong \mu_{W_\ell^T W_\ell} \boxtimes^{(L-1)} \mu_{D_\ell^2} \boxtimes^{(L-1)}$$

asym. freeness

B. Collins

Comm. Math. Phys.

$\int$ trivial or

Marchenko-Pastur

AISTATS

Pennington, Emergence universarity in Neural Network

arXiv 2018 2019

$\partial_\theta f$

$$\theta_{t+1} = \theta_t - \nabla_\theta L_\theta$$

$$(\nabla_\theta L_\theta)^T (\nabla_\theta L_\theta) \; ?$$

$$H := \sum_m \frac{\partial f(x(m))^T}{\partial \theta} \frac{\partial f(x(m))}{\partial \theta}$$

$$D = \{(x_m, y_m) \mid m = 1, \cdots, N\}$$

jacot 2018 → Neural Tangent kernel

$$\frac{df_{\theta t}}{dt} = \eta H (Y - f_\theta(X))$$

$$\begin{cases} Y = (y(1) \cdots y(N)) \\ X = (x(1) \cdots x(N)) \end{cases}$$

arXiv.
Conjugate Kernerel
2022 E. Wang.