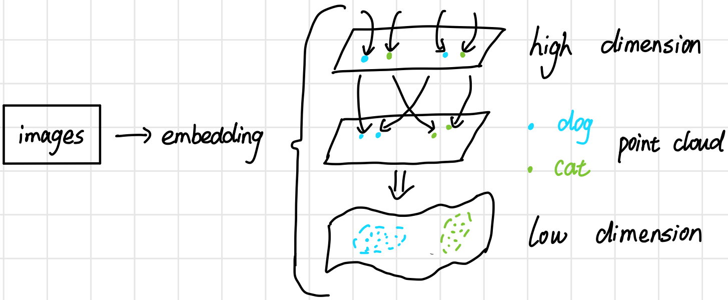
Information Amount-Guide Angular Margin Loss (IGAM Loss)

1: Category information amount

① Definition

Since Manifold Distribution Hypothesis :



Manifold Volume → Category information amount

② Measure

Category i → information amount I_i → $I_i = Vol(X_i)$

$X_i = [x_1, x_2, \cdots, x_m]$, m: instances number

embeddings set $X_i = [x_1, x_2, \cdots, x_j, \cdots x_m] \in R^{p\times m}$

$x_j \in R^p$ → embedding dimension

High dimension embedding space distribution characteristics → Covariance matrix $\Sigma(X_i)$

$\Sigma(X_i) = \frac{1}{m}\sum_{j=1}^{m}(x_j - \bar{x})(x_j - \bar{x})^T$, $\bar{x} = \frac{1}{m}\sum_{j=1}^{m} x_j$

for Covariance matrix estimation accuracy ↑

Employ Ledoit-péché nonlinear shrinkage

$\Sigma(X_i) = V diag(\lambda_1, \lambda_2, \cdots, \lambda_p) V^T$

V: eigenvectors matrix $\lambda_i = max(\lambda_i, \lambda_-)$

λ_-: nonlinearly transformed minimum eigenvalue

$\lambda_- = (1 - \sqrt{\frac{p}{m}})^2$

$\Sigma(X_i)$ determinant → Manifold Volume

$I_i = Vol(X_i) = \frac{1}{2}\log_2 det(\Sigma(X_i) + I)$

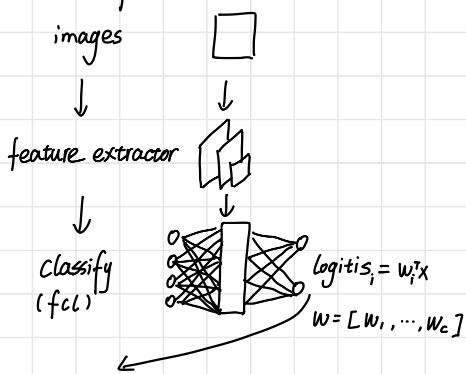
numerical stability

positive-definite

2: IGAM Loss

① Etiology (motivation)

classification:



feature vector : x

label : i

Cross-entropy loss: $L = -\log\left(\frac{e^{w_i^T x}}{\sum_{j=1}^{c} e^{w_j^T x}}\right)$ (softmax)

For example : Binary classification

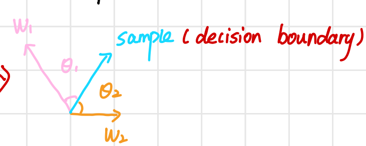
Core : for sample locate on decision boundary

we assume $w_1^T x = w_2^T x$

$w_1^T x = w_2^T x \Rightarrow \cancel{\|x\|_2} \cdot \|w_1\|_2 \cdot \cos\theta_1 = \cancel{\|x\|_2} \cdot \|w_2\|_2 \cdot \cos\theta_2$

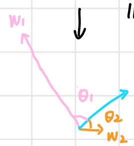
$\|(\,)\|_2$: L_2 norm, $0 \le \theta_1, \theta_2 \le \frac{\pi}{2}$

Decision space :



↓ $\|w_1\|_2 >> \|w_2\|_2$ [long-tail scenarios : weight vectors extremely unbalance]

$\theta_1 >> \theta_2$



Obviously! decision space for class 2 is pathologically compressed !

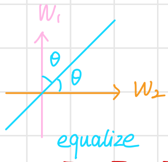
② Solution

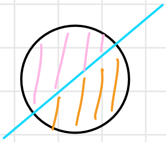
(1) Directly equalize decision space:

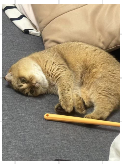
Ignore → weight vectors of norm

$$L = -\log\left(\frac{e^{s\cdot\cos\theta_i}}{\sum_{j=1}^{C} e^{s\cdot\cos\theta_j}}\right), \quad \cos\theta_i = \frac{W_i^T x}{\|W_i\|_2\cdot\|x\|_2}$$

optimization goal: Minimum θ between W_i, X







Information Amount: cat > plane

perceptual manifold volume: cat > plane

model learning focus: cat > plane

So Directly equalize is unreasonable!

(2) Dynamically adjust decision boundary:

Guided by category information amount

$$L = -\log\left(\frac{e^{s\cdot\cos\theta_i}}{e^{s\cdot\cos\theta_i} + \sum_{j=1, j\neq i}^{C} e^{s\cdot\cos(\theta_j + m_{ij})}}\right)$$

$$m_{ij} = \max\left(0, \frac{1}{\pi}\cdot\log\left(\frac{I_i'}{I_j'}\right)\right), \quad I_i' = \frac{e^{e^{I_i/(\bar{I}\cdot\sqrt{C})}}}{\sum_{j=1}^{C} e^{I_j/(\bar{I}\cdot\sqrt{C})}}\cdot C + 1, \quad \bar{I} = \sum_{i=1}^{C} I_i$$

I_i': normalized information amount of class i

m_{ij}: Ratio of information amount of class i and class j

$I_i' > I_j'$ decision space expanded

$I_i' < I_j'$ decision space compressed

IGAM Loss making model focus on complex classes, allocate more decision space.

3: End to end train frame

(1) Dynamic (2) low-cost

Engineering challenge: Dynamic update category information amount → All instances embedding Covariance matrix

feature slow shift:

sample a in epoch n embedding ≈ sample a in epoch n+1 embedding

① original strategy:

empty queue → epoch 1 train → embedding → epoch 2 train → information → embedding ...

defect: storage space ↑

② optimization strategy:

Information Amount ⇐ Global Cov Matrix

Global ← local 1, local 2, local 3

C categories, N instances, d queue length (d<N)

Batch → empty → d → local cov Matrix, mean

New Batch → empty → → ~

⋮

⇓

Local: Floor$\left(\frac{N}{d}\right)+1$ local cov matrices Σ_i^k, local means μ_i^k

$i = 1, \cdots, C$, $k = 1, \cdots, \text{floor}\left(\frac{N}{d}\right)+1$

⇓

global:

$$\mu_i = \frac{1}{N_i}\sum_{k=1}^{\lfloor N/d\rfloor+1} n_i^k \mu_i^k, \quad n_i^k: \text{local instances number}$$

N_i: total instances number

$$\Sigma_i = \frac{1}{N_i}\left(\sum_{k=1}^{\lfloor N/d\rfloor+1} n_i^k \Sigma_i^k + \sum_{k=1}^{\lfloor N/d\rfloor+1} n_i^k (\mu_i^k - \mu_i)(\mu_i^k - \mu_i)^T\right)$$

$$\Rightarrow \text{Vol}_i = \frac{1}{2}\log_2 \det(I + \Sigma_i)$$