Paper: The details matter: preventing class collapse in supcon

Notations:

Labeled input data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

$(x, y) \in P$

$x \in \mathcal{X} \in \mathbb{R}^{\text{data dim.}}$

$y \in \mathcal{Y} = \{1, 2, \dots K\}$

Data label $h(x) \in \mathcal{Y}$ // $P(y \mid x)$

$P(y = i) = 1/K$

Target: to learn a model $\hat{P}(y \mid x)$

Data point belongs to categories beyond labels

STRATA

strata as latent labels $z \in Z = \{1, 2, \dots c\}$

Z divided into disjoint subset: $S_1, S_2 \dots S_K$

$z \in S_k, \quad y = k$

$S(c)$ denotes determinis label c.

1st strata is sampled $P(z)$

x is sampled $P_z = P(\cdot \mid z)$

label is $y = S(z)$

SupCon and collapse embedding

Similarity

$$\sigma(x, x') = f^T(x) f(x') / \tau$$

$B \rightarrow$ set of batches of labeled dataset on D

positive

$$P(i, B) = \{ p \in B \setminus i : h(p) = h(i) \}$$

SupCon loss:

$$\hat{L}_{sc}(f, x_i, B) = \frac{-1}{|P(i,B)|} \sum_{p \in P(i,B)} \log \frac{\exp(\sigma(x_i, x_p))}{\sum_{a \in B \setminus i} \exp(\sigma(x_i, x_a))}$$

positive pair.

Negative pair.

Class collapse: simplex embedding scenario.

if $h(x) = i$ then $f(x) = v_i \quad \forall x \in \mathbb{B}$

$\{v_i\}_{i=1}^{K} \rightarrow$ forms regular simplex.

with

properties of v_i

i) $\sum_{i=1}^{K} v_i = 0$

ii) $\|v_i\|_2 = 1$

in collapse of

(iii) $\exists\, c_k \in \mathbb{R}$ s.t. $v_i^T v_j = c_k$ for $i \neq j$

End model

Linear Classifier: $W \in \mathbb{R}^{k \times d}$

$$\|w_y\|_2 \leq 1 \quad ; \quad y \in \mathcal{Y}$$

Empirical loss:

$$\hat{\mathcal{L}}(w, \mathcal{D}) = \sum_{x_i \in \mathcal{D}} -\log \frac{\exp\left(f(x_i)\, w_{h(x_i)}\right)}{\sum_{j=1}^{k} \exp\left(f(x_i)^T w_j\right)}$$

Prediction:

$$\hat{p}(y \mid x) = \hat{p}(y \mid f(x))$$

generalized error:

$$\mathcal{L}(x, y, f) = \mathbb{E}_{x,y}\left[-\log \hat{p}(y \mid f(x))\right]$$

Methodologies

1. class collapse minimize $\mathcal{L}(x, z, f)\ \forall x$

losses strata:

when i) $p(y = h(x) \mid x) = 1$

ii) $p(z \mid x) = 1/m$; $z \in S_{h(x)}$

iii) $p(x \mid z) = p(x \mid y)$ // no strata distinction

Modified Contrastive loss

$$\mathcal{L}_{spread} = \alpha L_{attract} + (1-\alpha) L_{repel}.$$

Negative examples : $N(i, B) = \{ a \in B \setminus i : h(a) = h(i) \}$

$$\hat{L}_{att}(f, x_i, B) = \frac{-1}{|P(i,B)|} \sum_{p \in P(i,B)} \log \frac{\exp(\sigma(x_i, x_p))}{\exp(\sigma(x_i, x_p)) + \sum_{a \in N(i,B)} \exp(\sigma(x_i, x_a))}$$

$$\hat{L}_{rep} = -\log \frac{\exp(\sigma(x_i, x_i^{aug}))}{\sum_{p \in P(i,B)} \exp(\sigma(x_i, x_p))}$$

[typical solution]

spread the positive around.