Class- incremental NCD

Problem Def. & Notation:

Labelled set, $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{n^l}$

Supervised task $\mathcal{T}^l$

$x^l \in \mathcal{X}^l$

$y^l \in \mathcal{Y}^l$

c^l Dimensional one hot encoding

New task $\mathcal{T}^u$

$\mathcal{D}^u = \{x_i^u\}_{i=1}^{n^u}$ $\in \mathcal{X}^u$

total class c^u from $\mathcal{Y}^u$ label.

NCD Assumption

$$\mathcal{Y}^l \cap \mathcal{Y}^u = \phi$$

Goal: cluster the $\mathcal{D}^u$ images be learning learnt information by

$$f^l : \mathcal{X}^l \rightarrow \mathcal{Y}^l$$

(maintaining previous task ($\mathcal{Y}^l$) performance)

Aim for single mapping function $f: \mathcal{X} \to \mathcal{Y}^l \cup \mathcal{Y}^u$

where $x \in \mathcal{X} = \mathcal{X}^u \cup \mathcal{X}^l$

$\boxed{\text{Overall framework}}$

Stage I :

parametric NN

Learn mapping function $f^l : x^l \to y^l$

compose into $h^l \circ g$ → encoder Net.

θ_h classification head

θ_g

Assumption: Base class learning is good

[what if not ??]

Prototype: μ_c

intermediate feature: $z^l = g(x^l)$

store the variance : v_c^2

→ generative approach. (Base class consistency)

Stage II : Data D^u

transfered f^l weights.

learn unique classifier $c^A = c^l + c^u$

h^A

Then $h \longrightarrow h$
becomes

$\boxed{f^u : h^u \circ g}$ → trained in clustering objective.

Preliminaries:

(i) Supervised Setting — Normal As heck.

$$\boxed{\mathcal{L}_{CE} = - \mathbb{E}_{p(x^l, y^l)} \frac{1}{C^l} \sum_{k=1}^{C^l} y_k^l \log \sigma_k\left(h^l(g(x^l))\right)}$$

(ii) KD to prevent foregetting:

$\boxed{\text{LwF loss}}$

$$\mathcal{L}_{KD}^{logit} = - \mathbb{E}_{p(x^{new})} \frac{1}{k^{old}} \sum_{k=1}^{k^{old}} \pi^k\left(h^{old}(g^{old}(x^{new}))\right)$$

$$\log \pi_k\left(h^{old}(g^{new}(x^{new}))\right)$$

vector

ith element

$$\pi_i(a) = \frac{\exp(a_i/\tau)}{\sum_j \exp(a_j/\tau)}$$

[weighted Softmax]

Class-incremental NCD:

Self-training:

D^u Dataset

[lack classification ability.]

Network $f^u = h^u \circ g$

Pairwise Similarity: (x_i^u, x_j^u) [weak supervision]

two Different instances.

Features $z_i^u = g(x_i^u)$

$z_j^u = g(x_j^u)$

Autonovel ranking stat.

[if top k ranked dimension of feature matches → they are from same classes.

Pairwise Pseudo label:

$$\tilde{y}_{ij}^u = \mathbb{1}\{\underbrace{top_k(z_i^u) = top_k(z_j^u)}\}$$

subset of top k activated

features position

features positives.

<u>Clustering Objective:</u>

$$\mathcal{L}_{BCE} = -\mathop{E}_{p(z^u)}\left[\tilde{y}^u_{ij} \log p_{ij} + (1 - \tilde{y}_{ij}) \log(1 - p_{ij})\right]$$

where, $p_{ij} = \underbrace{\langle \sigma(h^u(g(x_i^u))), \sigma(h^u(g(x_j^u)))\rangle}_{\text{Dot product of the predictions.}}$

<u>Self-training loss:</u>

$$\mathcal{L}_{self} = -\mathop{E}_{(x^u, \tilde{y}^u)} \frac{1}{|C^{|A|}|} \sum_{k=1}^{|C^{|A|}|} \tilde{y}^u_k \log \sigma_k(h^A(g(x^u)))$$

$\tilde{y}^u_k \downarrow$ $c^{|L|} + \mathop{\arg\max}_{k \in C^u} h^u(g(x^u))$

Augmented view consistency loss:

$$\mathcal{L}_{MSE} = \mathop{E}_{p(x^u, \bar{x}^u)} \frac{1}{|C^{|u|}|} \sum_{k=1}^{|C^{|u|}|} \left(\sigma_k(h^u(g(x^u))) - \sigma_k(h^u(g(\bar{x}^u)))\right)^2$$

<u>Overall</u>

$$\mathcal{L}_{novel} = \mathcal{L}_{BCE} + w_{self}(t)\,\mathcal{L}_{self} + w_{MSE}(t)\,\mathcal{L}_{MSE}$$

Feature replay and Distillation for class incremental learning

$$\mu_c^{l} = \frac{1}{n_c^{l}} \sum_{i=1}^{n_c^{l}} g(x_i^{l})$$ // mean → class c

$$v_c^{l\,2} = \frac{1}{n_c^{l}} \sum_{i=1}^{n_c^{l}} \left(g(x_i^{l}) - \mu_c\right)^2$$ // variance

= - - - - Replay loss (LwF)

$$\mathcal{L}_{replay} = -\mathop{\mathbb{E}}_{c \sim C^{l}} \mathop{\mathbb{E}}_{(z^{l}, y_c^{l}) \sim \mathcal{N}(\mu, v_c^2)} \sum_{k=1}^{|C^{A}|} y_{kc}^{l} \log\left(h^{A}(z^{l})\right)$$

sample from class

Doesn't train the complete Network

Extra regularization

$$\mathcal{L}_{KD}^{feat} = -\mathop{\mathbb{E}}_{p(x^{u})} \left\| g^{l}(x^{u}) - g(x^{u}) \right\|$$

$$\mathcal{L}_{past} = \mathcal{L}_{replay} + \lambda \mathcal{L}_{KD}^{feat}$$

Finally

$$\mathcal{L}_{\text{frost}} = \mathcal{L}_{\text{novel}} + \mathcal{L}_{\text{pasto}}$$