- ActiveFT.

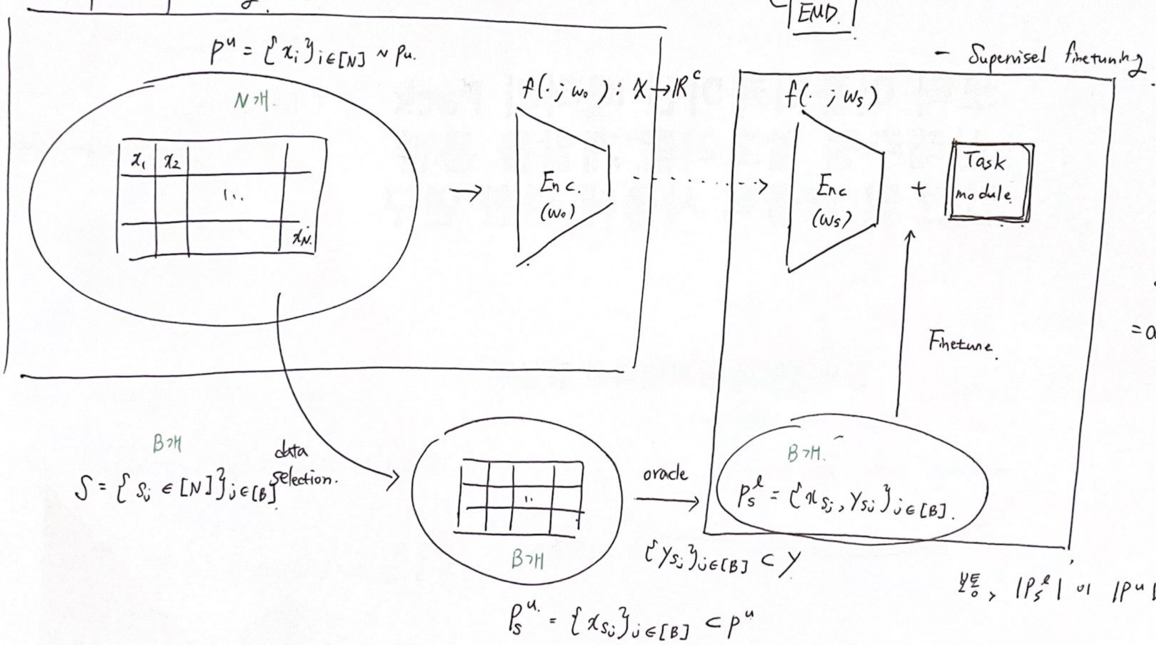
unsupervised pretraining + $\boxed{\text{supervised finetuning}}$

↑ $\boxed{\text{active finetuning}}$

↑ $\boxed{\text{parametric model optimization}}$ ≡ $\boxed{\text{continuous space optimization}}$

↑ $\boxed{\text{EMD}}$

- unsupervised pretraining.



- active finetuning의 목표

= S_{opt}.

$$S_{opt} = \underset{S}{\operatorname{argmin}} \; \underset{x,y \in \mathcal{X} \times \mathcal{Y}}{E} \left[\operatorname{error}(f(x; w_s), y) \right]$$

보통, $|P_s^l|$ 이 $|P^u|$ 의 10%. 보다 작음.

$\mathcal{X}$: data space

$\mathcal{Y}$: label space.

data space

$\mathcal{X}$. P^u

distribution : $p_u(x)$

data : x_i

$\boxed{P^u = \{x_i\}_{i \in [N]} \sim p_u}$

$\rightarrow$ $f(\cdot\,; w_0)$ Enc (w_0) $\rightarrow$

feature space

F^u

distribution : $p_{f_u}(f)$

feature : $f_i = f(x_i\,; w_0)$

$\boxed{F^u = \{f_i\}_{i \in [N]} \sim p_{f_u}}$

$\downarrow$ $S = \{s_j \in [N]\}_{j \in [B]}$

selected data space P_s^u

$\boxed{P_s^u = \{x_{s_j}\}_{j \in [B]} \sim p_u}$

$\rightarrow$ $f(\cdot\,; w_0)$ Enc. $\longrightarrow$

selected feature space F_s^u

$\boxed{F_s^u = \{f_{s_j}\}_{j \in [B]} \sim p_{f_s}}$; discrete

$\Downarrow$ $\theta_s = \{\theta_s^j\}_{j \in [B]}$; continuous parameters.

$\boxed{F_s^u = \cancel{\{f_{s_j}\}_{j \in [B]}} \sim p_{\theta_s}}$ $\Rightarrow$ $\boxed{\theta_s = \{\theta_s^j\}_{j \in [B]} \sim p_{\theta_s}}$; continuous

$f_{s_j} \xleftrightarrow[\text{close}]{} \theta_s^j$ after optimization

θ_s

- active finetuning 의 목표 = S_{opt}.

$$S_{opt} = \underset{S}{\operatorname{argmin}}\ D(p_{f_u}, p_{f_s}) - \lambda \cdot R(F_s^u)$$

$$\Rightarrow \theta_{s,opt} = \underset{\theta_s}{\operatorname{argmin}}\ D(p_{f_u}, p_{\theta_s}) - \lambda R(\theta_s), \quad \|\theta_s^j\|_2 = 1$$

— parametric model p_{θ_s}

$$p_{\theta_s}(f) = \sum_{j=1}^{B} \phi_j \cdot p(f \mid \theta_s^j) \quad , \quad \sum_{j=1}^{B} \phi_j = 1 \; , \quad j \in [B]$$

$$= \phi_1 \cdot p(f \mid \theta_s^1) + \phi_2 \cdot p(f \mid \theta_s^2) + \cdots + \phi_B \cdot p(f \mid \theta_s^B)$$

$$p(f \mid \theta_s^j) = \frac{\exp(\mathrm{sim}(f, \theta_s^j)/\tau)}{Z_j} \quad , \quad \mathrm{sim}(f_1, f_2) = f_1^T f_2 \; , \quad \|f_1\|_2 = \|f_2\|_2 = 1.$$

각 $f_i \in F^u$ 에 대해, f_i 와 가장 가까운 $\theta_s^{c_i}$ 로, optimization 진행 중 c_i 계속 업데이트.

$$c_i = \underset{\substack{j \in [B] \\ \cancel{i \in [N]}}}{\mathrm{argmax}} \; \mathrm{sim}(\underset{F^u}{\tilde{f_i}}, \underset{\theta_s}{\tilde{\theta_s^j}}) \quad , \quad i \in [N]$$

N 개인데 종류는 B개.

F^u N개 θ_s B개

— Assumption 1.

$$\forall i \in [N], j \in [B] \; , \; \tau \text{ is small} \; , \quad \exp(\mathrm{sim}(f_i, \theta_s^{c_i})/\tau) \gg \exp(\mathrm{sim}(f_i, \theta_s^j)/\tau) \; , \quad j \neq c_i$$

$$p(f_i \mid \theta_s^{c_i}) \gg p(f_i \mid \theta_s^j) \; , \quad j \neq c_i, \; j \in [B].$$

- parametric model for $f_i \in F^u$.

$$p_{\theta_s}(f_i) \approx \phi_{c_i} \cdot p(f_i \mid \theta_s^{c_i})$$

$$= \frac{\exp(\mathrm{sim}(f_i \mid \theta_s^{c_i}) / \tau)}{Z_{c_i} / \phi_{c_i}}$$

$$= \frac{\exp(\mathrm{sim}(f_i, \theta_s^{c_i}) / \tau)}{\tilde{Z}_{c_i}}, \quad \tilde{Z}_{c_i} = Z_{c_i} / \phi_{c_i}$$

$$\therefore p_{\theta_s}(f_i) \propto \exp(\mathrm{sim}(f_i, \theta_s^{c_i}) / \tau)$$

p_{f_u} 와 p_{θ_s} 는 KL-divergence 를 최소화함으로써 가까워질수 있음.

$$KL(p_{f_u} \| p_{\theta_s}) = \sum_{f_i \in F^u} p_{f_u}(f_i) \cdot \log \frac{p_{f_u}(f_i)}{p_{\theta_s}(f_i)}$$

$$= \underbrace{\mathop{E}_{f_i \in F^u}[\log p_{f_u}(f_i)]}_{:\ \text{상수}} - \mathop{E}_{f_i \in F^u}[\log p_{\theta_s}(f_i)].$$

$\therefore$ Minimize KL-divergence $KL(p_{f_u} \| p_{\theta_s}) \doteqdot$ Maximize $\mathop{E}_{f_i \in F^u}[\log p_{\theta_s}(f_i)]$

$\doteqdot$ Maximize $\mathop{E}_{f_i \in F^u}[\log \exp(\mathrm{sim}(f_i, \theta_s^{c_i}) / \tau)] \doteqdot$ Maximize $\mathop{E}_{f_i \in F^u}[\mathrm{sim}(f_i, \theta_s^{c_i}) / \tau]$

7. ~~$\therefore D(p$~~

$$\therefore D(p_{f_u}, p_{\theta_s}) = -\mathop{E}_{f_i \in F^u}[\mathrm{sim}(f_i, \theta_s^{c_i}) / \tau]$$

- extra regularization term to ensure diversity of selected subset.

$$R(\theta_s) = -\mathop{E}_{j \in [B]}\left[\log \sum_{k \neq j, k \in [B]} \exp(\mathrm{sim}(\theta_s^j, \theta_s^k) / \tau)\right]$$

$$\therefore L = D(p_{f_u}, p_{\theta_s}) - \lambda \cdot R(\theta_s)$$

$$= -\mathop{E}_{f_i \in F^u}[\mathrm{sim}(f_i, \theta_s^{c_i}) / \tau]$$

$$+ \mathop{E}_{j \in [B]}\left[\log \sum_{k \neq j, k \in [B]} \exp(\mathrm{sim}(\theta_s^j, \theta_s^k) / \tau)\right]$$

- 결론.

optimization 끝나고,

θ_s^j 와 가장 비슷한 $\{f_{s_j}\}_{j \in [B]}$ 찾음.

$$f_{s_j} = \mathop{\arg\max}_{f_k \in F^u} \mathrm{sim}(f_k, \theta_s^j)$$

그리고 그것들의 상응하는 $\{x_{s_j}\}_{j \in [B]}$ 들로

P_s^u 구성. $S = \{s_j\}_{j \in [B]}$

optimization $\frac{1}{T}$, $\{f_{s_j}\}_{j \in [B]}$.

$$p_{f_s}(f_{s_j}) = \frac{|C_j|}{N}, \quad C_j = \{f_i \mid c_i = j\}, \quad f_{s_j} \in F_s^u$$

$$p_{f_u}(f_i) = \frac{1}{N}, \quad f_i \in F^u.$$

$$EMD(p_{f_u}, p_{f_s}) = \inf_{\gamma \in \Pi(p_{f_u}, p_{f_s})} \mathbb{E}_{(f_i, f_{s_j}) \sim \gamma} \left[\| f_i - f_{s_j} \|_2 \right]$$

$$\gamma_{f_u, f_s}(f_i, f_{s_j}) = \begin{cases} \frac{1}{N}, & f_i \in F^u, \ f_{s_j} \in F_s^u, \quad c_i = j \\ 0, & o/w \end{cases}$$

$$\therefore EMD(p_{f_u}, p_{f_s}) = \frac{1}{N} \sum_{i=1}^{N} \left[\sqrt{2 - 2\,\mathrm{sim}(f_i, f_{s_{c_i}})} \right]$$