Contrastive Coding for Active Learning under Class Distribution Mismatch

(ICCV, 2021)

Class Distribution Mismatch

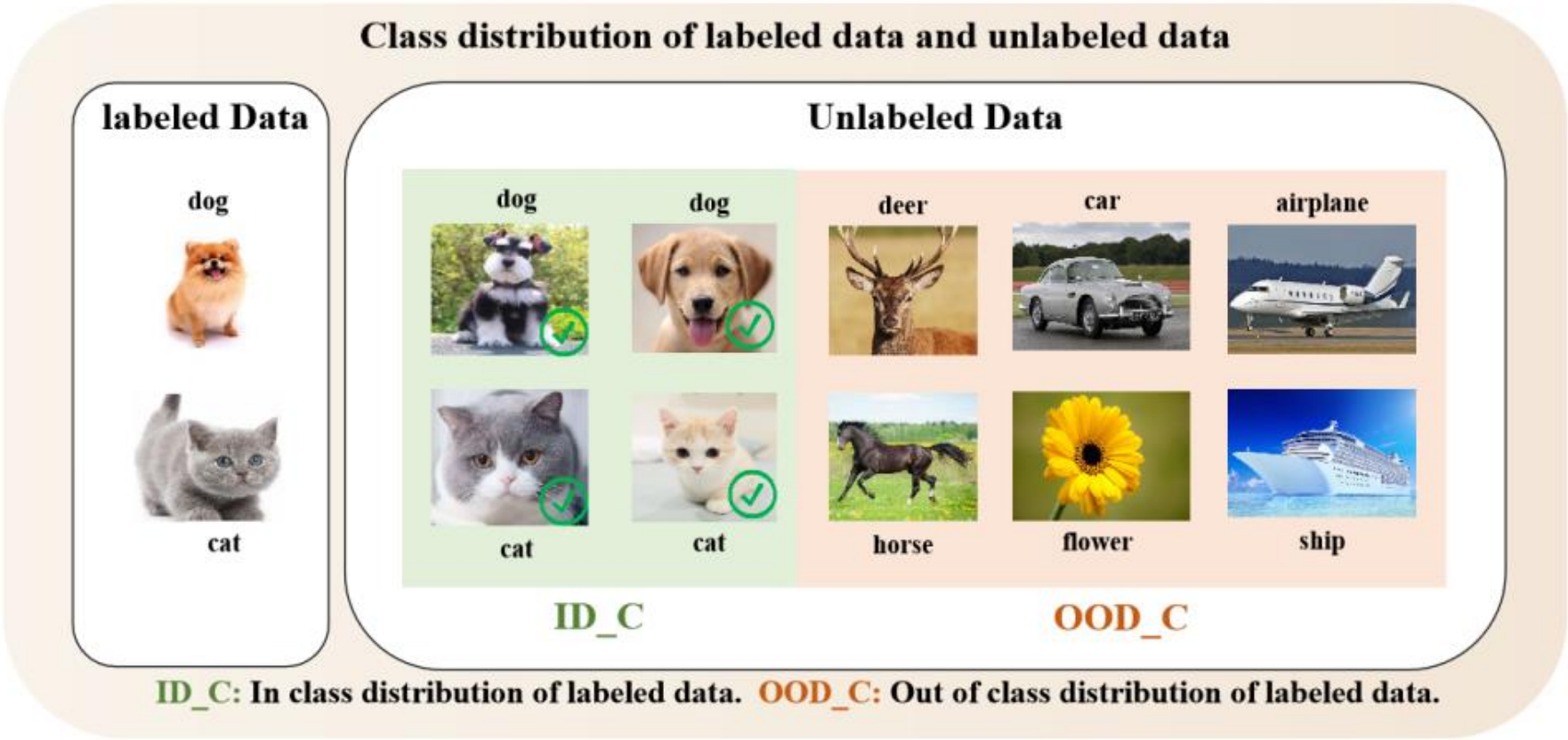


Figure 1: An instance of class distribution mismatch. Unlabeled data contains some samples that are out of the class distribution of labeled data.

Motivation

- Tradition AL methods generally assumed that labeled and unlabeled data are drawn from the same class distribution.

- Existing information-based active learning methods tend to query the unknown samples with mismatched categories.

Idea

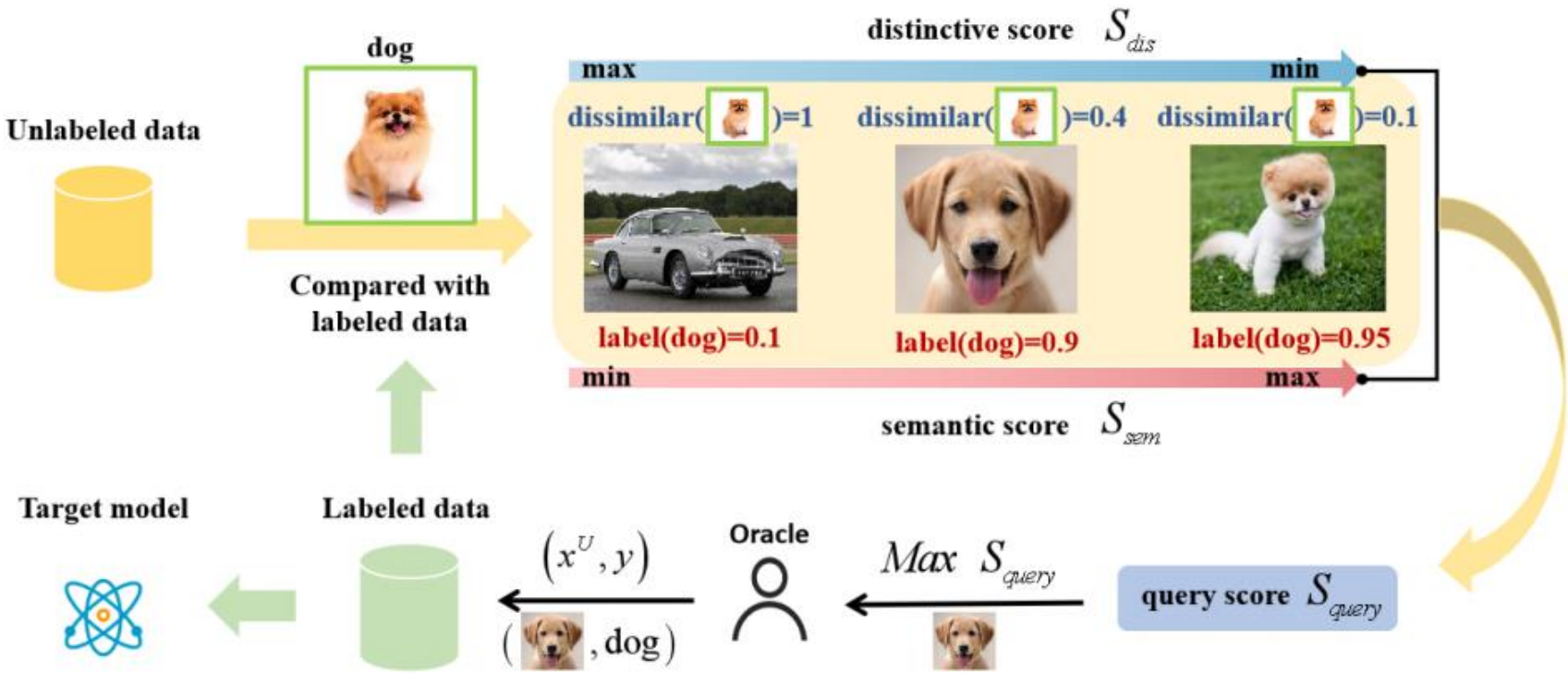


Figure 2: CCAL combining the semantic score S_{sem} and distinctive score S_{dis} to select samples to annotate.

1. Learn distinctive and semantic features.
2. Query samples with high semantic consistency and strong distinguishability

Methods

Learning semantic features

Loss

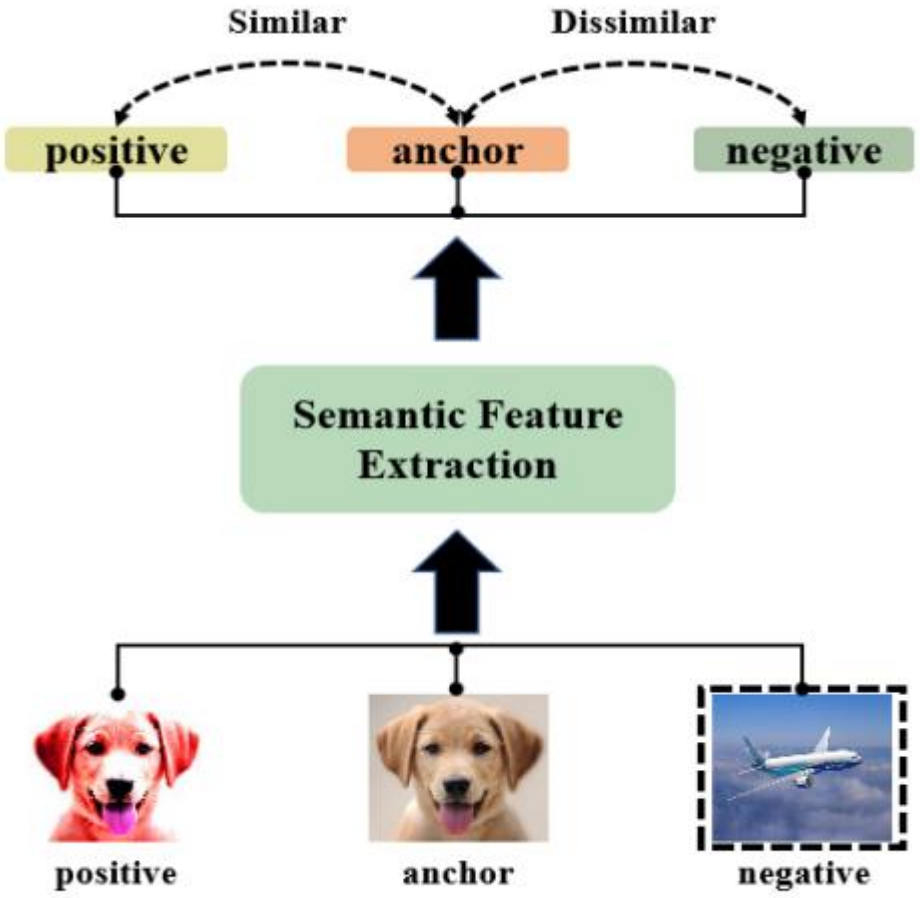
$$L_{sem}(B;\aleph) = \frac{1}{2\,|B|}\sum_{i=1}^{|B|} L_{con}(\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)}, \tilde{B}_{-i}) + L_{con}(\tilde{x}_i^{(2)}, \tilde{x}_i^{(1)}, \tilde{B}_{-i}) \quad (2)$$

where $\tilde{B} = \{\tilde{x}_i^{(1)}\}_{i=1}^{|B|} \cup \{\tilde{x}_i^{(2)}\}_{i=1}^{|B|}$, $\tilde{B}_{-i} = \{\tilde{x}_j^{(1)}\}_{j\neq i} \cup \{\tilde{x}_j^{(2)}\}_{j\neq i}$, $x_i \in X^L \cup X^U$, and L_{con} is a contrastive learning loss [3].

Semantic score

$$S_{sem}(x_i^U) = \sigma(\max_j cos(z_s(x_j^L), z_s(x_i^U))), \quad (3)$$

where $\sigma(s) = (s - min(S_{sem}(X^U)))/(max(S_{sem}(X^U)) - min(S_{sem}(X^U)))$, $x_i^U \in X^U$.



(a) Contrast of semantics.

Methods

Learning distinctive features

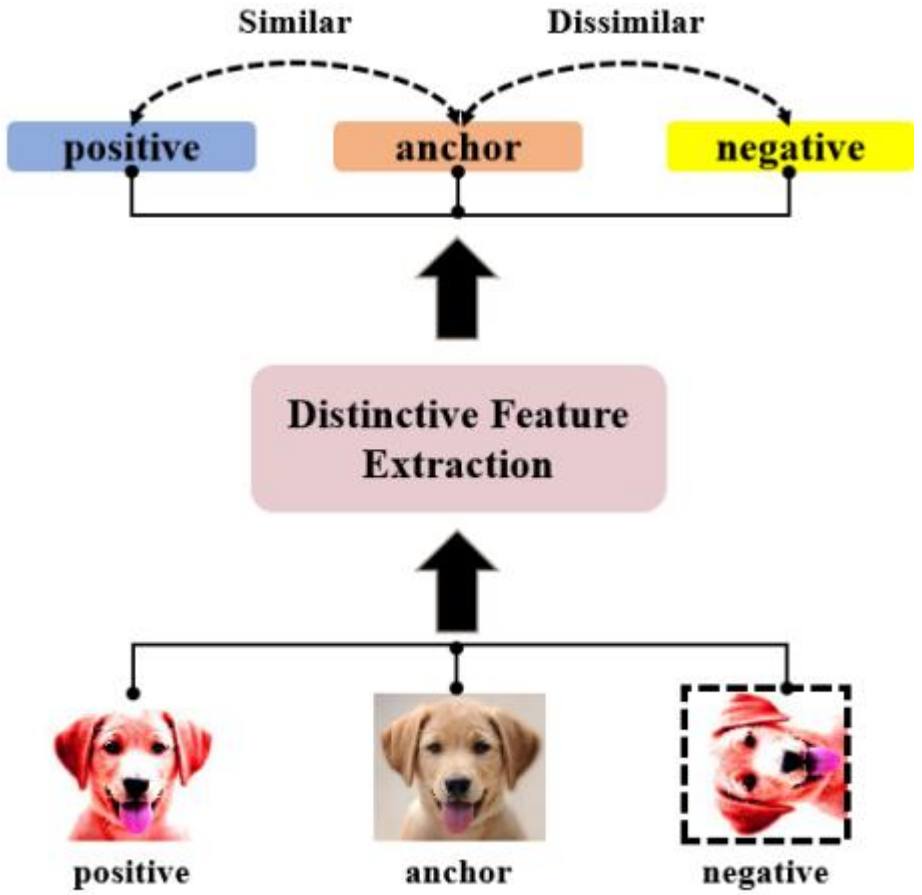
Loss

$$L_{dis}(B;\Re;\aleph) = \frac{1}{2|B|}\frac{1}{|r|}\sum_{i=1}^{|B|}\sum_{k\in r}[L_{con}(\tilde{x}^{(1)}_{\Re_k(i)}, \tilde{x}^{(2)}_{\Re_k(i)}, \tilde{B}^{\Re}_{-i}) + L_{con}(\tilde{x}^{(2)}_{\Re_k(i)}, \tilde{x}^{(1)}_{\Re_k(i)}, \tilde{B}^{\Re}_{-i})] \quad (4)$$

Distinctive score

$$S_{dis}(x^U_i) = 1 - \sigma(cos(z_d(x^U_i), z_d(x^L_{i,st})) - cos(z_d(x^U_i), z_d(x^L_{i,nd})) + cos(z_d(x^L_{i,st}), z_d(x^L_{i,nd}))) \quad (5)$$

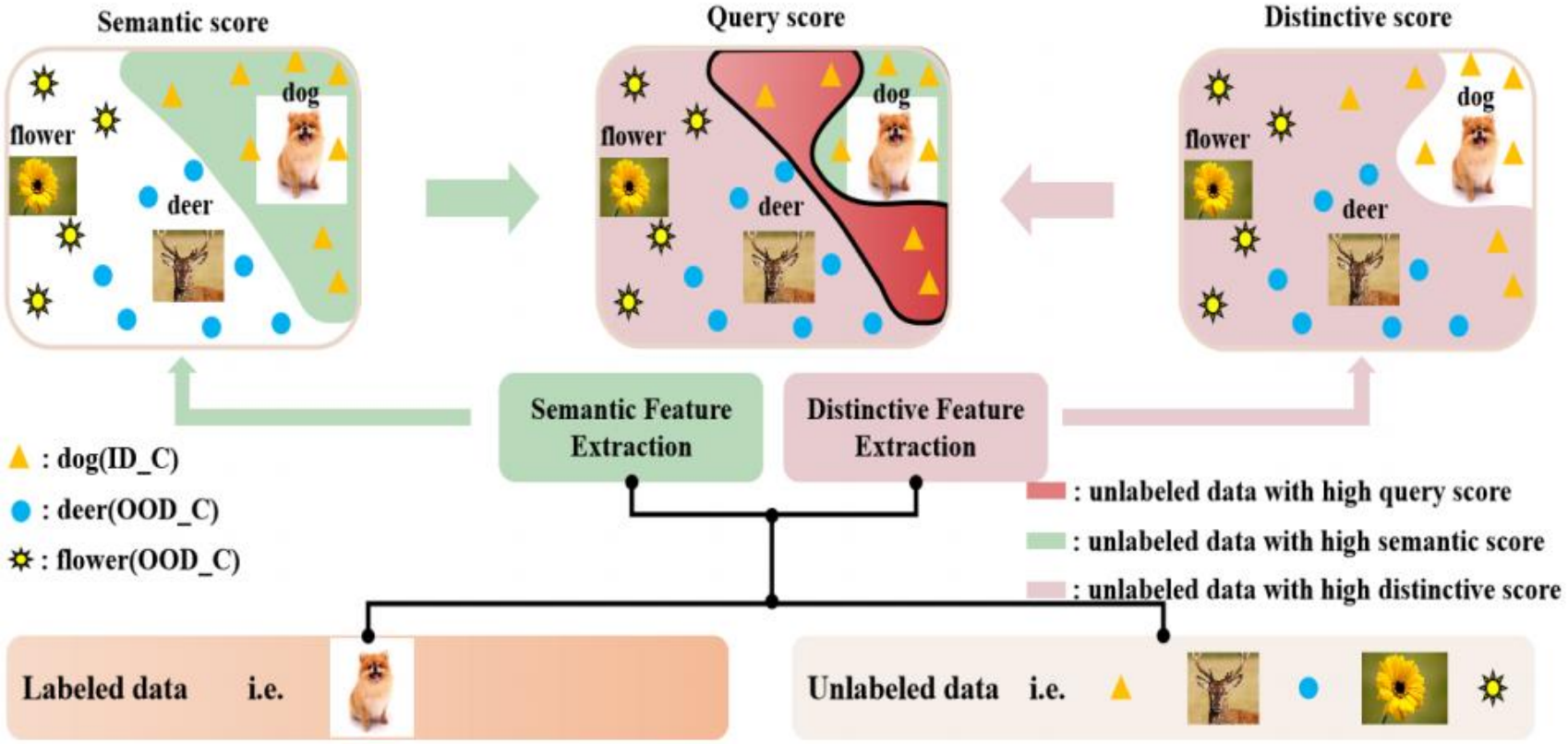
In Eq.5, $cos(z_d(x^U_i), z_d(x^L_{i,st})) - cos(z_d(x^U_i), z_d(x^L_{i,nd}))$ measures the difference of x^U_i to labeled samples.



(b) Contrast of distinctiveness.

Methods

Joint Query Strategy



(c) Joint query strategy.

$$S_{query}(x_i^U) = tanh(\psi(S_{sem}(x_i^U))) + S_{dis}(x_i^U) \quad (6)$$

where $\psi(S_{sem}(x_i^U)) = k \times (S_{sem}(x_i^U) - t)$

Experiments

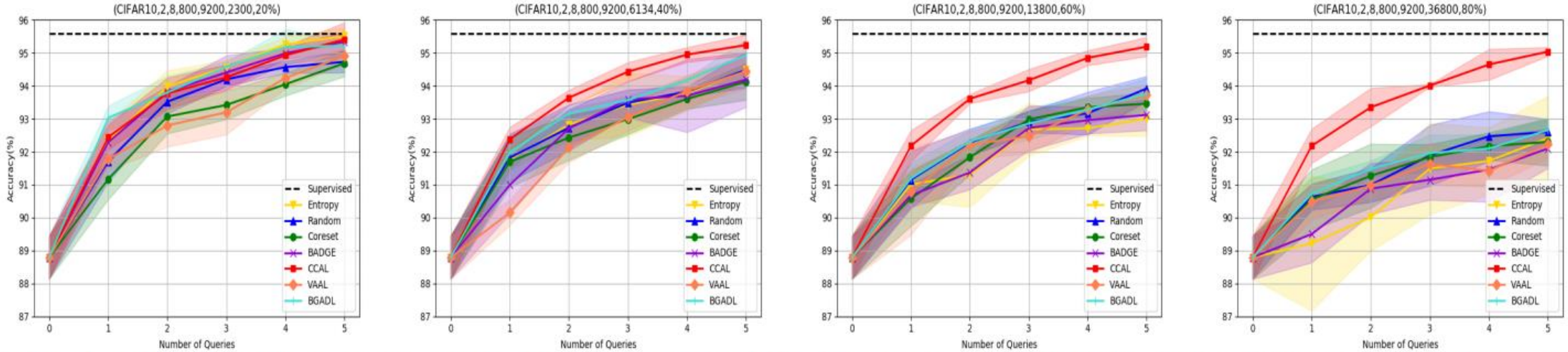


Figure 4: Classification accuracy of CCAL and compared AL algorithms on CIFAR10 under different mismatches. The shaded area represents the standard deviation of the five runs.

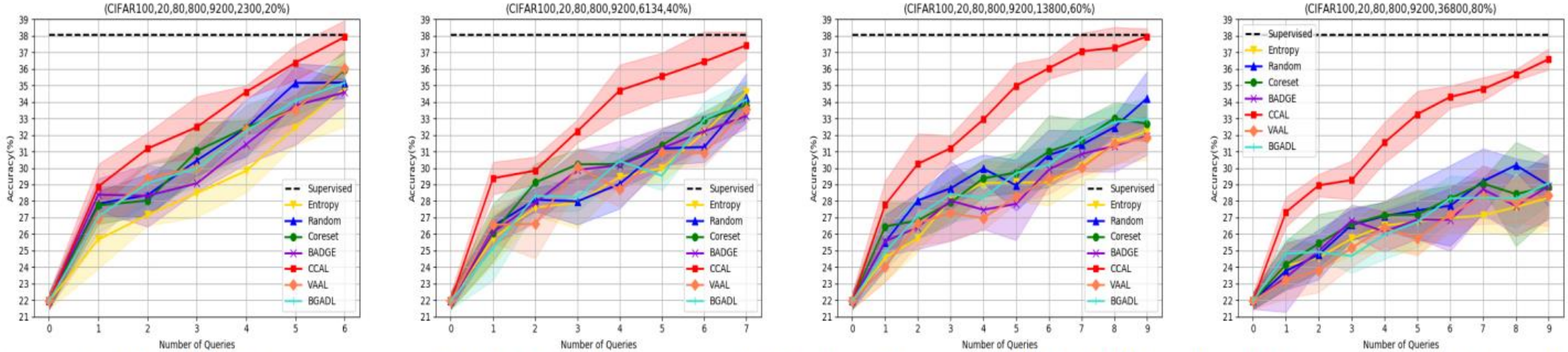
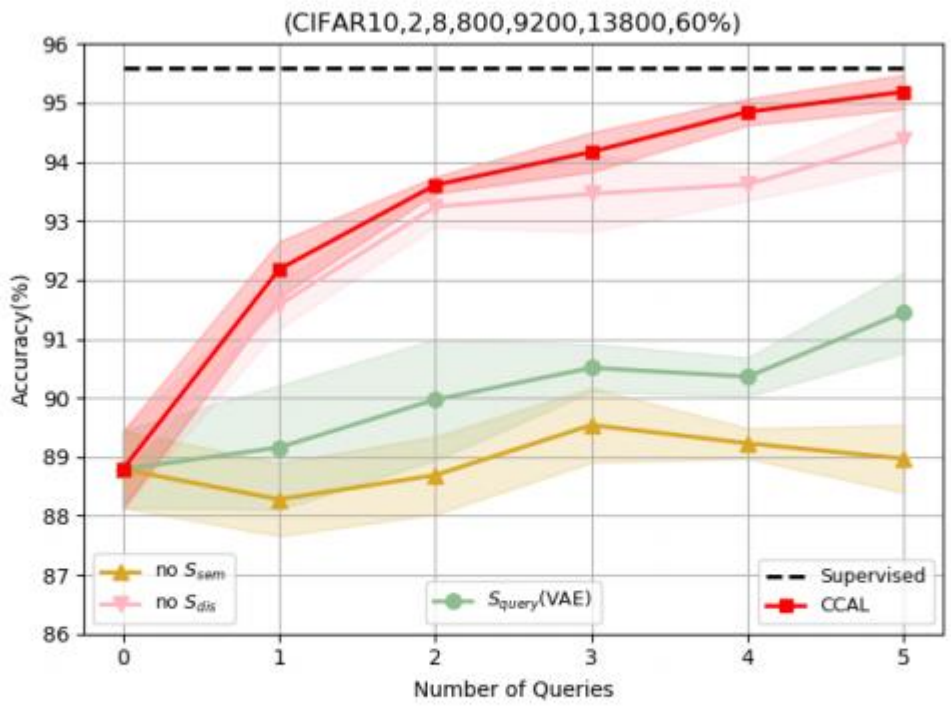


Figure 5: Classification accuracy of CCAL and compared AL algorithms on CIFAR100 under different mismatches. The shaded area represents the standard deviation of the five runs.

Experiments



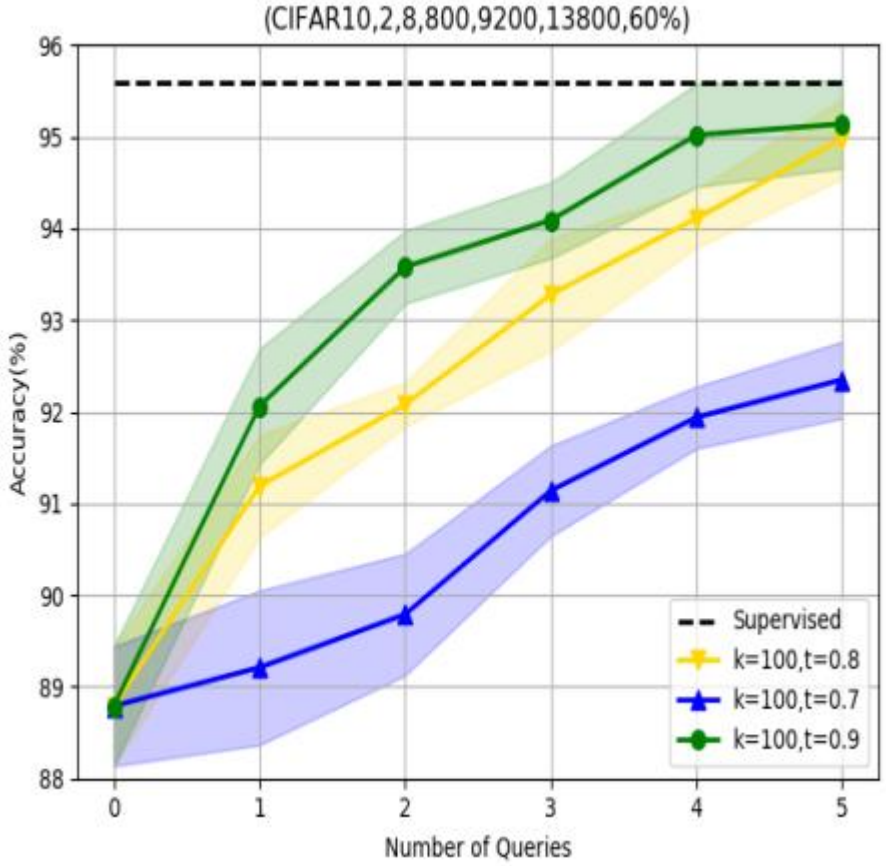


Figure 6: Ablation study to analyze the influence of each part of CCAL.

Upgoing policy update (UPGO)

(Nature, 2019)

TD(λ)

$$V_\pi(s) = E_\pi[G_t|S_t = t] = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots |S_t = t]$$

$$\Downarrow$$

$$V_\pi(s) \approx \text{average}(G_t),$$

$$\text{s.t.} \quad S_t = s$$

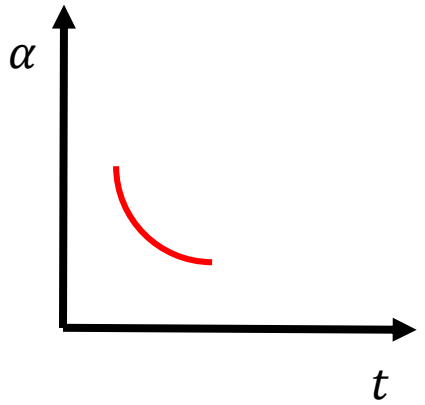
$$\Downarrow$$

$$N(s) = N(s) + 1$$

$$V(s) = V(s) + \frac{1}{N(s)}(G_t - V(s))$$

$$\Downarrow$$

$$V(s) = V(s) + a(G_t - V(s))$$



TD(λ)

$$V(s) = V(s) + a(G_t - V(s))$$

$$V_\pi(s) = E_\pi[G_t|S_t = t] = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = t]$$

⇩ Bellman

$$G_t = R_{t+1} + \gamma V(S_{t+1})$$

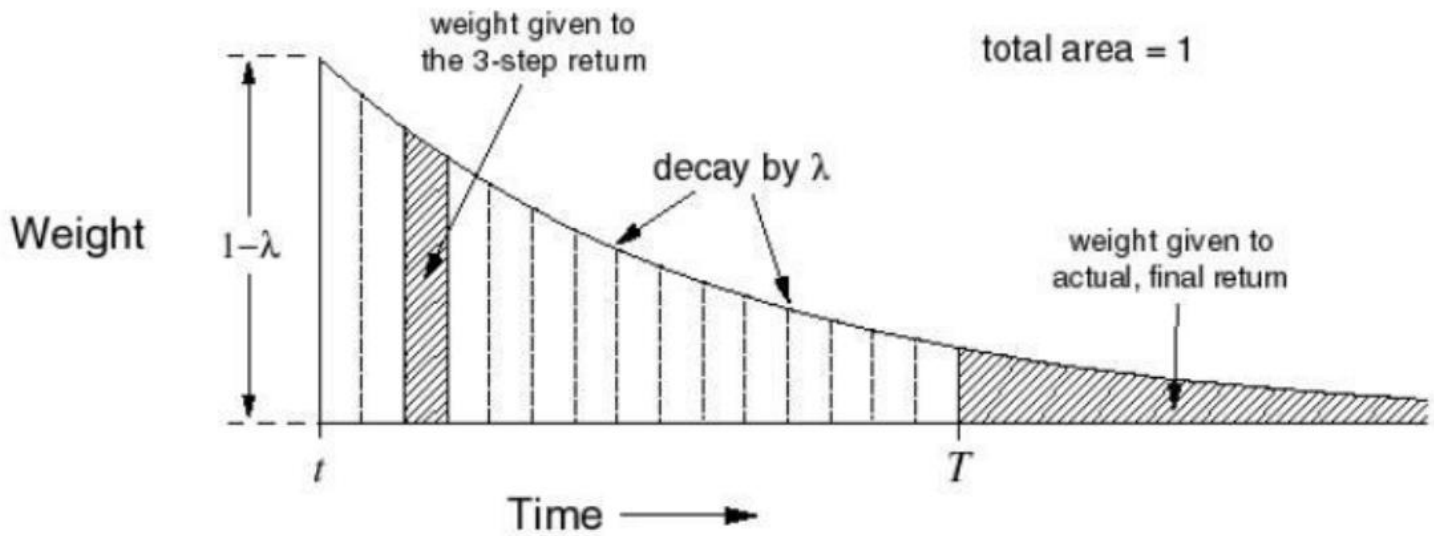
⇩

$$G_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$$

$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^n V(S_{t+n})$$

⇩

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$



权重的和加起来为 1，这样的设置让我们能够更关注步骤小的 G。

GAE

$$A_t = G_t - V(s)$$

$$\hat{A}_t^{(1)} := \delta_t^V = -V(s_t) + r_t + \gamma V(s_{t+1}) \tag{11}$$

$$\hat{A}_t^{(2)} := \delta_t^V + \gamma\delta_{t+1}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2}) \tag{12}$$

$$\hat{A}_t^{(3)} := \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 V(s_{t+3}) \tag{13}$$

$$\hat{A}_t^{(k)} := \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}^V = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{k-1} r_{t+k-1} + \gamma^k V(s_{t+k}) \tag{14}$$

$$\begin{aligned}
\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} &:= (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda\hat{A}_t^{(2)} + \lambda^2\hat{A}_t^{(3)} + \dots\right) \\
&= (1-\lambda)\left(\delta_t^V + \lambda(\delta_t^V + \gamma\delta_{t+1}^V) + \lambda^2(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V) + \dots\right) \\
&= (1-\lambda)(\delta_t^V(1 + \lambda + \lambda^2 + \dots) + \gamma\delta_{t+1}^V(\lambda + \lambda^2 + \lambda^3 + \dots) \\
&\qquad + \gamma^2\delta_{t+2}^V(\lambda^2 + \lambda^3 + \lambda^4 + \dots) + \dots) \\
&= (1-\lambda)\left(\delta_t^V\left(\frac{1}{1-\lambda}\right) + \gamma\delta_{t+1}^V\left(\frac{\lambda}{1-\lambda}\right) + \gamma^2\delta_{t+2}^V\left(\frac{\lambda^2}{1-\lambda}\right) + \dots\right) \\
&= \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V
\end{aligned} \tag{16}$$

UPGO

$$L^{PG}(\theta) = \hat{\mathbb{E}}_t\left[\log \pi_\theta(a_t \mid s_t)\hat{A}_t\right].$$

$$A_t = \mathrm{G_t} - \mathrm{V(s)}$$

$$G_t = \begin{cases} R_t + \gamma G_{t+1}, & if\ Q(s_{t+1}, a_{t+1}) \geq V_\theta(s_{t+1}) \\ R_t + \gamma V_\theta(s_{t+1}), & otherwise \end{cases}$$

$$Q(s_{t+1}, a_{t+1}) \geq V_\theta(s_{t+1}) \Rightarrow R_t + \gamma \mathrm{V}_\theta(s_{t+1}) - V_\theta(s_t) \geq 0 \Rightarrow TD(1) \geq 0$$

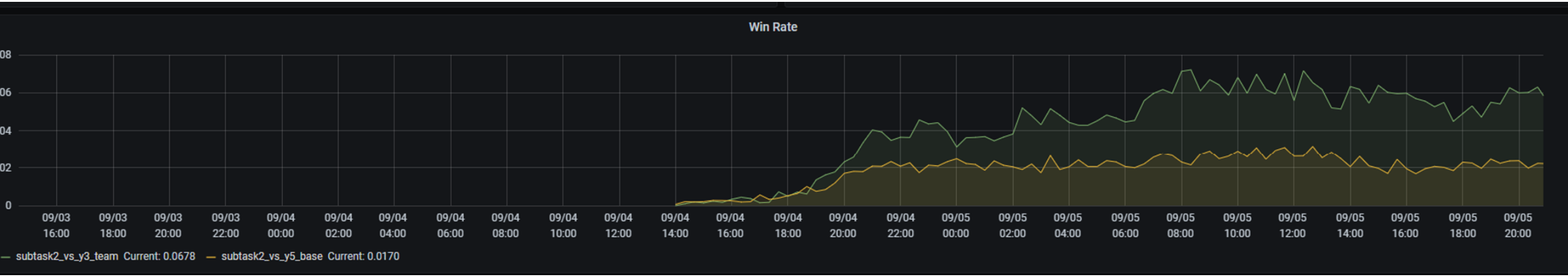
⇧

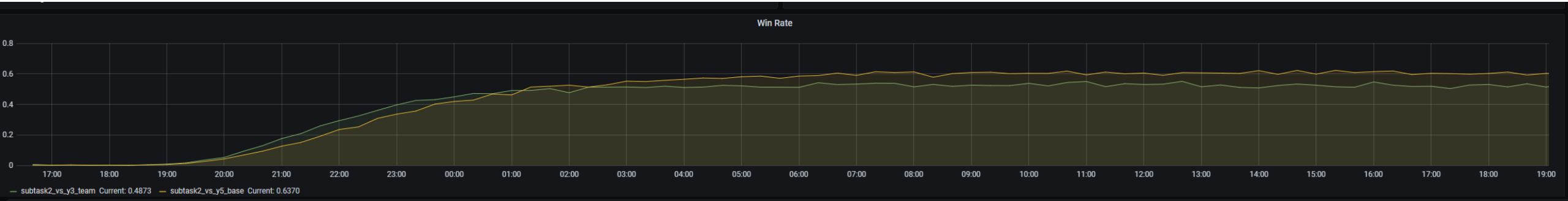
等价于Self-Training. 贪心的放大最确定的经验。

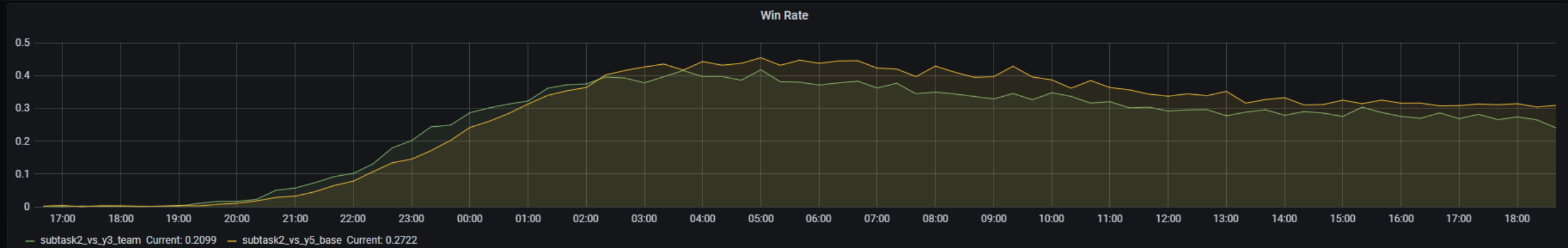
⇩

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots + \gamma^n R_{t+n}$$

Experiments







Thanks