Algorithm suggestion: Using test mAP for inference method under resource constraints

Dongwoo Goo

1 Variables

- K: number of the Exits
- C_i: Amount of $Layer_i$ and $EarlyExit_i$ computation sum
- C_{RD}: Amount of RPN, Detector computation sum
- $S_i = S_{i-1} + C_{RD}$: total valid computation of ith exit
- $u_d(t)$: Computation resource in time slot T (Hz)
- P_i: Test mAP value of the particular ith exit
- $\tau(t) = \frac{S_k + C_{RD}}{u_d(t)}$: Delay parameter

2 Objective

1. Delay Minimization

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\tau(t)\right]$$

2. Guarantee the minimum accuracy in mAP

$$(S \cdot T) \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[P_{exit}(t)\right] \ \geq \ P_{th}$$

3 Lyapunov Optimization

- Virtual Queue

$$Y(t+1) = \mathbf{max}(Y(t) + P_{th} - P_{\text{exit}}(t),\ 0)$$

$$L(t) = Y(t)^2 \quad \text{and} \quad \Delta L(t) = \mathbb{E}\left[L(t+1) - L(t)\right]$$

- DPP

$$\Delta L(t) \leq 2Y(t)(P_{th} - P_{\text{exit}}(t)) + (P_{th} - P_{\text{exit}}(t))^2$$

- Objective

$$\Delta L(t) + V\mathbb{E}\left[\tau(t)|Y(t)\right]$$
$$= 2Y(t)(P_{th} - P_{\text{exit}}(t)) + (P_{th} - P_{\text{exit}}(t))^2 + V(\frac{S_i}{u_d(t)})$$

Algorithm 1 : Finding the optimal Exit

```
Input: $Y(t), u_d(t)$
Output: exit
Initialization:
exit ← $K$
tmp ← $K$

Iteration:
for e= 1 to $K$ do
    objective ← $2Y(t)(P_{th} - P_{\text{exit}}(t)) + (P_{th} - P_{\text{exit}}(t))^2 + V(\frac{S_i}{u_d(t)})$
    if min(tmp,objective) = tmp then
        pass
    else
        $tmp$ = objective,
        exit = e
    end if
end for
```