MuSCM: Mutual Spatial Correlation Mapping for Class Incremental Detection Transformer

Supplementary Material

1. Generalize to DN-DETR

1.1. DN-DETR

The denoising mechanism proposed by DN-DETR can improve convergence speed while retaining most of the original DETR network design [1]. This plug-and-play characteristic has made it widely adopted in subsequent works [2–4]. Considering the reason mentioned above, we selected DN-DETR to validate the generalizability of the proposed MuSCM.

DN-DETR comprises four key components: a backbone $\mathcal{B}$, a transformer encoder $\mathcal{E}$, a transformer decoder $\mathcal{D}$, and a prediction head $\mathcal{H}$. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the backbone $\mathcal{B}$ extracts the image features $F^{\mathcal{B}}$ from I. These features are then optimized by the encoder $\mathcal{E}$, which employs attention mechanisms, resulting in $F^{\mathcal{E}} = \mathcal{E}(F^{\mathcal{B}})$.

The decoder is a cascade of L transformer layers which processes $F^{\mathcal{E}}$ alongside a set of learnable object queries $Q = \{q_i\}_{i=1}^{N_q}$ ($q_i \in \mathbb{R}^d$) as inputs, where N_q is the number of queries and d is the embedding dimension. For a intermediate decoder layer index as l, it takes over the output queries $Q^{l-1} = \{q_i^{l-1}\}_{i=1}^{N_q}$ of last layer, and generates refined queries $Q^l = \{q_i^l\}_{i=1}^{N_q}$ responsible for predicting a potential object. Specifically, each layer primarily consists of self-attention, cross-attention, and a feed-forward network in sequence.

Each cross-attention mechanism processes three inputs: a set of queries $X = \{x_i^l\}_{i=1}^{N_q}$, a set of keys $Y = \{y_i^l\}_{i=1}^{N_{kv}}$, and a set of values $Z = \{z_i^l\}_{i=1}^{N_{kv}}$, where N_{kv} denote the number of keys or values. The cross-attention maps are computed based on the softmax of dot-products between queries and keys:

$$n_{ij}^l = \frac{e^{\frac{1}{\sqrt{d}} x_i^{l\top} y_j^l}}{\sum_{j=1}^{N_{kv}} e^{\frac{1}{\sqrt{d}} x_i^{l\top} y_j^l}}, \tag{1}$$

where i is a query index, and j is a key index. The attention output for each query is the aggregation of values weighted by attention maps:

$$\mathcal{K}_i^l = \boldsymbol{W}_{out}^l \sum_{j=1}^{N_{kv}} n_{ij}^l z_j^l, \tag{2}$$

where $\boldsymbol{W}_{out}^l$ denote the output weights.

Cross-attention is responsible for aggregating object-related local features in the image, with keys and values derived from $F^{\mathcal{E}}$ and queries from the output of self-attention. The output of the cross-attention is then processed by the FFN to generate object queries Q^l, and fed them into head $\mathcal{H}$ to generate the predictions $\hat{Y}^l = \{(\hat{c}_i^l, \hat{b}_i^l)\}_{i=1}^{N_q}$, where $\hat{c}_i^l$ is the class prediction and $\hat{b}_i^l$ is the bounding box.

Then, like MuSCM for Deformable DETR, we compute $\mathcal{L}_{\text{gt}}(Y, \hat{Y})$.

1.2. Spatially Consistent Distillation

Similar to Spatially Consistent Distillation (SCD) for Deformable- DETR, we map the student or teacher model's spatial correlations to the opposite model and obtain two pairs of spatially consistent predictions.

Formally, in the l^{th} decoder layer of student $\mathcal{M}_S$, we extract the cross attention $\hat{n}_{S,i}^l$ corresponding to query $q_{S,i}^l$, and map them into the l^{th} decoder layer of teacher $\mathcal{M}_T$:

$$\mathcal{K}_{S2T,i}^l = \boldsymbol{W}_{T,out}^l \sum_{j=1}^{N_{kv}} \hat{n}_{S,ij}^l z_{T,j}^l, \tag{3}$$

where $\mathcal{K}_{S2T,i}^l$ denotes aggregated features based on student's spatial correlations, and $S2T$ denotes the mapping direction.

We also map the self-attention map to the self-attention module to ensure the consistency of the $\mathrm{Add\&Norm}$ operations. Ultimately, the prediction head $\mathcal{H}_T$ generates consistent targets $\hat{Y}_{S2T}^l = \{(\hat{c}_{S2T,i}^l, \hat{b}_{S2T,i}^l)\}_{i=1}^{N_q}$ to supervise $\hat{Y}_S^l = \{(\hat{c}_{S,i}^l, \hat{b}_{S,i}^l)\}_{i=1}^{N_q}$.

Next, we compute the $\mathcal{L}_{DP}^{S2T}$, $\mathcal{L}_{DP}^{T2S}$, $\mathcal{L}_{DR}^{S2T}$, and $\mathcal{L}_{DR}^{T2S}$, as in Deformable DETR.

1.3. Spatial Correlation Matching

Next, we propose Spatial Correlation Matching (SCM) module to transfer teacher's spatial correlations for old class objects to student. As in Deformable DETR, we get a set of confident foregrounds $\tilde{Y}_T^l = \{(\tilde{c}_{T,i}^l, \tilde{b}_{T,i}^l)\}_{i=1}^{N_p}$, where N_p is the number of selected foregrounds.

To transfer the teacher's spatial correlation to student, we incorporate the similarity between $\tilde{n}_T^l$ and $\hat{n}_S^l$ in bipartite matching, which represents cross-attention maps of object queries corresponding to $\tilde{Y}_T^l$ and $\hat{Y}_S^l$, respectively.

The bipartite matching process turns to:

$$\hat{\epsilon}^l = \arg\min_{\epsilon} \sum_{i=1}^{N_p} \mathcal{C}_{\text{cls}}(\tilde{c}_{T,i}^l, \hat{c}_{S,\epsilon(i)}^l) + \mathcal{C}_{\text{box}}(\tilde{b}_{T,i}^l, \hat{b}_{S,\epsilon(i)}^l) + \mathcal{C}_{KL}(\tilde{n}_{T,i}^l, \hat{n}_{S,\epsilon(i)}^l), \tag{4}$$

where $\hat{\epsilon}^l$ is the optimal permutation between $\tilde{Y}_T^l$ and $\hat{Y}_S^l$, $\mathcal{C}_{KL}$ represents the KL divergence.

Next, we use the KL divergence to enhance the consistency of matched cross-attention maps, together with $\mathcal{L}_{\text{gt}}(\tilde{Y}_T^l, \hat{Y}_S^l)$ which adopts $\hat{\epsilon}^l$, forming the Spatial Correlation Matching loss $\mathcal{L}_{SCM}$:

$$\mathcal{L}_{SCM} = \sum_{l=1}^{L} \mathcal{L}_{\text{gt}}(\tilde{Y}_T^l, \hat{Y}_S^l) + \mathcal{L}_{KL}(\tilde{n}_{T,i}^l, \hat{n}_{S,\epsilon(i)}^l). \quad (5)$$

References

[1] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018. 1

[2] Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, et al. Detection transformer with stable matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6491–6500, 2023. 1

[3] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[4] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 1