H-057

Open-Vocabulary 物体検出モデルの時空間動作検出への拡張

細谷優　堀史門　玉木徹
名古屋工業大学
E-mail: {y.hosoya.657,s.hori.110}@nitech.jp, tamaki.toru@nitech.ac.jp

1 はじめに

本研究では，画像用の Open-Vocabulary 物体検出モデルを Open-Vocabulary 時空間動作検出 (Spatio-Temporal Action Detection, STAD) モデルへと拡張する手法を提案する．本研究は open-vocabulary 物体検出モデルである GroundingDINO [5] をフレーム毎に適用し，時間情報のモデル化として特徴量シフト [4] を用いる．これは特徴量を前後のフレームにシフトすることで，計算コストやパラメータを追加することなく時間情報を簡易にモデル化する方法である．ただし，GroundingDINO の cross-modality query はフレーム毎に異なる物体に対応する可能性があり，異なる物体に対応するクエリ特徴量をシフトしてしまう問題がある．そこで，フレーム間でクエリマッチングを適用する手法 [7] を利用して，異なるフレームにおける同一物体のクエリ特徴量をシフトすることを保証する．

2 関連研究

2.1 Open-vocabulary 物体検出

テキストにマッチした画像内の領域を検出する open-vocabulary 物体検出タスクのモデルとして GroundingDINO [5] が提案されている．GroundingDINO の拡張として，本研究と類似する VideoGroundingDINO [6] が提案されている．VideoGroundingDINO では動画像のために新たな Transformer 層が追加されているが，学習コストが大きくなるという課題がある．これに対して本手法はシフトベースの手法であり，パラメータ数の増加なく効率的に動画へ適用することが可能である．

2.2 Spatio-Temporal Action Detection

STAD は動画の各フレームにおいて人物の動作に対する外接矩形 (bounding box, bbox) を取得し，そのbbox の系列を action tube [1] として出力するタスクである．本研究では GroundingDINO を拡張して，特徴量シフト [7] を導入し，open-vocabulary な STAD の実現を目指す．

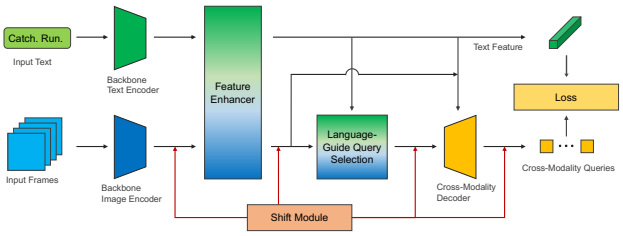

図 1: 提案手法の概略図．GroundingDINO を STAD へと拡張する．

3 提案手法

提案手法の全体像を図 1 に示す．まず，動画像の各フレームを open-vocabulary 物体検出モデルへ入力し，フレーム毎に人物動作を検出する．その際，動画像の時間方向の特徴量シフトによりフレーム間で時間情報を伝播する．

3.1 特徴量シフト

GroundingDINO を時間的に拡張するために，本研究では特徴量シフト [4] を導入する．また，GroundingDINO の層出力に対して特徴量シフトを適用する．T を時刻，D を特徴量次元として，以下のように入力 $z_{in} \in \mathbb{R}^{T\times D}$ を時間的に前後にシフトする．

$$z_{\text{out}}[1:,:D_{\text{f}}] = z_{in}[:-1,:D_{\text{f}}] \tag{1}$$

$$z_{\text{out}}[:-1,D_{\text{f}}:D_{\text{b}}] = z_{\text{in}}[1:,D_{\text{f}}:D_{\text{b}}] \tag{2}$$

$$z_{\text{out}}[:,D_{\text{b}}:] = z_{\text{in}}[:,D_{\text{b}}:] \tag{3}$$

ここで D の添え字の f は時間的に前方（forward）へのシフト，b は後方（backward）へのシフトを表す．つまり特徴量次元 D のうち，0 から $D_{\text{f}}-1$ までを前時刻 $t-1$ へのシフト，D_{f} から $D_{\text{b}}-1$ までを次時刻 $t+1$ へのシフト，D_{b} から D までをシフトなしとする．

3.2 Cross-modality query のシフト

GroundingDINO のようなクエリベースの手法に対して特徴量シフトを導入する方法が [7] で提案されている．本研究ではこれを用いて，各フレームにおいて同じ物体に対応する cross-modality query をマッチングしてシフトを行う方法を提案する．

まず，隣接するフレームにおける N 個の cross-modality query 同士の N^2 個の組み合わせについて cos

類似度を計算する．その後，二部マッチング問題をハンガリアンアルゴリズムで解くことにより，マッチする query を求める．具体的には，時刻 t における query に対応する $t+1$ の query の最適な順列 $\hat{\sigma}$ を次式で探索する．

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}}{\operatorname{argmin}} \sum_{i}^{N} \mathcal{L}_{\text{match}}(q_{\text{i}}^{t}, q_{\hat{\sigma}(i)}^{t+1}) \tag{4}$$

ここで $\mathcal{L}_{\text{match}}$ は，時刻 t におけるインデックス i の query q_{i}^{t} と，時刻 $t+1$ におけるインデックス $\hat{\sigma}(i)$ の query $q_{\hat{\sigma}(i)}^{\text{t+1}}$ との間のマッチングコスト

$$\mathcal{L}_{\text{match}}(q_{\text{i}}^{t}, q_{\hat{\sigma}(i)}^{t+1}) = \cos\left(q_{\text{i}}^{t}, q_{\hat{\sigma}(i)}^{t+1}\right) \tag{5}$$

である．この cross-modality query のマッチングを全ての隣接フレームで行い，マッチングした cross-modality query 同士の特徴量シフトを行う．

4　実験

4.1　実験設定

データセットとして JHMDB21 [3] と AVA [2] を使用する．また，STAD では人物の動作をフレームごとの bounding box の列である tube として表す．また，評価指標には Frame-mAP と，各フレームの予測と真値の bounging box の IoU を平均とした 3D IoU による Video-mAP の 2 つを用いる．実験においては，IoU のしきい値を 0.5，0.75，および 0.5 から 0.95 まで 0.05 ごとに上昇させた平均 (0.5:0.95) の 3 種類を評価に用いる．

4.2　実験

実験では GroundingDINO の各層出力に対して，それぞれ特徴量シフトを適用し，open-vocabulary 物体検出モデルの STAD 拡張の有効性を検証する．また，cross-modality query をシフトする際，クエリマッチングを適用する場合としない場合の影響を検証する．これらの実験を各データセットに対して実施する予定である．

5　おわりに

本稿では画像用の open-vocabulary 物体検出モデルを拡張し STAD へと対応させた．また拡張の中で，時間情報をモデル化するために特徴量シフトを利用した．この手法によってモデルにパラメータを追加することなく，各フレームの同じ物体に対応した cross-modality query の特徴量をシフトすることが可能である．

謝辞

本研究の一部は，JSPS 科研費 JP22K12090 の助成を受けた．

参考文献

[1] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 759–768, 2015. 1

[2] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[3] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pp. 3192–3199. IEEE Computer Society, 2013. 2

[4] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[6] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Video-groundingdino: Towards open-vocabulary spatio-temporal video grounding, 2024. 1

[7] 堀史門, 大見一樹, 玉木徹. フレーム間のクエリマッチングを用いた物体検出モデルの時空間動作検出への拡張. In *SSII2024*, 2024. 1