1X-06

物理環境における変化点抽出の取り組み

黒田彗莉 † 小林一郎 †

† お茶の水女子大学

1 はじめに

機械学習を用いた実世界理解は近年の重要な課題の一つと言われている．その中でも特に予測という機能は実世界を正確に理解することで行うことができるヒト特有の能力である．しかし予測を対象にした先行研究は多数あるが，多くの研究は画像ピクセルの値の変化に着目し，画像内の物体の振る舞いまでは着目できていない．それと同時にヒトによる予測機能は視覚から取り入れた全ての系列情報を認識するのではなく，観測から抽出した重要なイベント対して働くものと考える．

本研究ではヒトの予測機能を模倣したモデルを構築するために，環境が変化したタイミングを抽出するモデルである Variational Temporal Abstraction （VTA）[1] を改良し，物理的特性や位置関係などを理解したうえで重要なイベントの抽出を可能にした．データセットとして物体の物理特性を表現した CLEVRER [2] を用い，物体の関係性をグラフとして構築した．そして観測した環境で大きな変化が起きている場面をグラフ構造の潜在状態の変化から推測し，そこで抽出されたイベントが物体の衝突などを正しく判定しているかの精度を検証した．

2 物理イベントの変化点抽出手法

本研究の概要図を図 1 に示す．先行研究 VTA は動画像などの系列情報から階層的な抽象度を見つける状態空間モデル （HRSSM）を提案する研究である．HRSSM は潜在状態の変化から系列情報において大きく環境が変化しているタイミングを抽出する．しかし VTA は入力情報として画像特徴量のみを対象とし，画像特徴量の変化を観測した環境の変化とみなしている．そのためヒトのように観測した物体の物理的な関係を理解をすることはできていない．そこで本研究は観測した出来事についてヒトのように状況を理解し，出来事が変化したタイミングを抽出できるように VTA の改良を行った．入力情報として物体の物理特性を情報にもつ CLEVRER [2] を用いた．CLEVRER は 20,000 個の動画とアノテーション情報から構成されており，本研究ではこれらの動画をそれぞれ 128 フレームに分割して訓練データを構築した．そして CLEVRER に写っている物体の種類や位置情報を考慮したグラフを構築した．グラフの構築には以下の 2 つの方法を用いた．

1. YOLOv3 [3] を用いて物体検知を施し，画像内の 2 次元位置情報と物体の種類（cube，cylinder，sphere）を取得
2. CLEVRER のアノテーション情報から空間内の 3 次元位置情報・物体の速度・加速度を取得

グラフを構築した後，それらのグラフを埋め込みベクトルに変換した．グラフの埋め込みベクトル作成においては node2vec [4] と graph2vec [5] の 2 種類の手法を用いた．node2vec を用いた埋め込みでは 1 フレーム内に写っている物体一つずつをノードとみなし，構築したグラフのノードを元に埋め込みベクトルを作成した．一方で graph2vec では，環境（CLEVRER 1 フレーム）を表現したグラフ全体を用いて埋め込みベクトルを作成した．

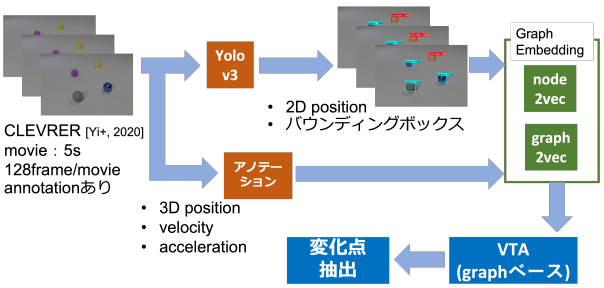

図 1: 変化点抽出手法の概要図

3 実験

3.1 設定

学習における設定は先行研究 [1] を参考に設定した．使用データセット数は 60 万，学習回数は 50 万，変化点のタイミングとして出力するフラグの個数を 80 個とした．精度の検証は表 1 にある 11 種類を対象にした．11 種類の訓練データの特徴を以下に記す．

A Study on Extracting the Inflection Point in the Physical Environment
†Eri KURODA (kuroda.eri@is.ocha.ac.jp)
†Ichiro KOBAYASHI (koba@is.ocha.ac.jp)

表 1: グラフ構造を用いた変化点抽出の精度結果

検証フレーム範囲		40 ~ 56	101 ~ 117	120 ~ 127 0 ~ 8	53 ~ 69	113 ~ 127 0 ~ 1	5 ~ 25
衝突 or 場面変化のタイミング		54 ~ 56	102 ~ 104	127	59 ~ 63 68 ~ 70	127	18 ~ 21
YOLOv3	① graph only	50	**100**	-	-	-	-
(node2vec)	② graph+image	14.3	25	9.1	37.5	14.3	28.6
annotation	③ graph only	20	**100**	20	**100**	**50**	33.3
(node2vec)	④ graph+image	22.2	22.2	20	50	12.5	25
	⑤ obj → graph only	**100**	**100**	**33.3**	66.7	25	**100**
	⑥ obj → graph+img	11.1	22.2	10	37.5	11.1	43.9
YOLOv3	⑦ graph only	-	-	-	-	-	-
(graph2vec)	⑧ graph+image	0	20	0	33.3	20	0
annotation	⑨ graph only	-	-	-	-	-	-
(graph2vec)	⑩ graph+image	0	20	20	50	0	0
VTA	⑪ image のみ	-	-	-	-	-	-

※精度は%で算出，-はフラグが立たなかったことを示している．

- YOLOv3-{graph only, graph+image}
  YOLOv3 で取得した 2 次元位置情報からグラフを構築し，node2vec および graph2vec で作成したグラフ埋め込み表現のみ（①，⑦）．
  ①，⑦に画像特徴量を追加（②，⑧）．
- annotation-node2vec-{graph only, graph+image}
  CLEVRER のアノテーション情報から取得した 3 次元位置情報からグラフ埋め込み表現を作成．その情報に物体の速度・加速度・フレーム間での物体の移動距離を追加（③）．③に画像特徴量を追加（④）．
- annotation-node2vec-{obj→graph only, obj→graph+image}
  ③，④に物体同士の位置関係の表現した 4 つのフラグを追加（⑤，⑥）．
- annotation-graph2vec-{graph only, graph+image}
  従来の graph2vec ではグラフのノードのポジションを考慮できていなかったため，本研究ではノードのポジションを考慮するように改良をし作成（⑨）．
  ⑨に画像特徴量を追加（⑩）．
- 画像特徴量のみ（⑪）．

また各検証フレーム範囲内において衝突および場面の切り替えが起きているので，そのタイミングで正しいフラグが立ったかを（正解のフラグ数）/（全フラグ数）として精度を算出した．タイミングが 127 となっている部分は場面変化が起き，それ以外は物体同士の衝突が発生している．正解とするタイミングはアノテーション情報のコリジョンデータから取得した．しかしヒトが CLEVRER を観測したときと比べると約 2 フレームの誤差があったため，正解のタイミングには幅をもたせた．

3.2 結果と考察

表 1 に精度結果を示す．全体の精度を比較すると，アノテーション情報に対して物体の位置関係のフラグを追加した場合（⑤）が最も高かった．これはグラフ構造だけでなくそれぞれの物体の位置関係を取得したことで，物体同士の細かな変化を扱えるようになったためと考えられる．先行研究 [1] のような画像特徴量のみの結果（⑪）と比較しても，物体の関係を表すグラフ構造を用いるとさらに詳細な場面の変化点を抽出可能になった．しかし画像特徴量を追加すると（⑥）精度は低下していた．CLEVRER における画像特徴量がノイズとみなされていることが原因だと考える．

また graph2vec での精度が低いことから，正しく変化点を抽出できていないことがわかった（⑦ ~⑩）．原因として node2vec と比べ，一つひとつの物体の速度や加速度といった細かい情報を保持していないためだと考える．

4 おわりに

本研究では画像内の物体の挙動に着目した新たな予測モデル構築のために，観測における重要なイベントを抽出する手法の提案を行った．環境が変化したタイミングを抽出するモデルである VTA を改良し，観測した環境に対して大きな変化が起きている場面をグラフ構造の潜在状態の変化から抽出することを可能にした．そして実験を通じて提案手法の有効性についても検証した．

参考文献

[1] Taesup Kim, Sungjin Ahn, and Yoshua Bengio. Variational temporal abstraction. *arXiv:1910.00775 [cs, stat]*, October 2019.

[2] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: CoLlision events for video REpresentation and reasoning. *arXiv:1910.01442 [cs]*, March 2020.

[3] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv:1804.02767 [cs]*, April 2018.

[4] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. *arXiv:1607.00653 [cs, stat]*, July 2016.

[5] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. July 2017.