2D-05

物理環境におけるヒト脳内での予測符号化を模倣した変化点予測モデル構築への取り組み

黒田彗莉 † 小林一郎 †

† お茶の水女子大学

1 はじめに

ヒトは視覚から環境を見たとき，場面や状況を瞬時に理解できる．この仕組みは観測内容の刺激から脳内で環境をモデル化し，シミュレーションすることで成り立つとされる [1]．また同時にヒトは脳内で構築した環境モデルを介して，環境にある物体を理解したり，物体の次のふるまいを予測する．しかし予測を扱う先行研究の多くは画像のピクセル値の変化から予測画像を生成しており，画像内の物体の詳細やふるまいをもとにした予測内容は生成していない．さらにヒトの予測機能を考えると，視覚から取り入れた全ての情報をもとに予測をしているのではなく，観測対象における重要なイベントに対して機能していると考えられる．

そこで本研究では，これまで独立に研究されてきた観測環境の物理特性理解と環境の予測機能を組み合わせ，よりヒトが行うに近い実世界予測モデルを提案する．モデルの構築は，物体の動作変化に対する変局点取得モデルである graph VTA[2] と，ヒト脳内の予測メカニズムを表したモデルの PredNet[3] を組み合わせた．また提案モデルの優位性について実験を通して検証した．

2 変化点予測モデル

ヒトは環境を観測したとき，観測した物体の重量感や速度といった物理特性を理解する．しかし物理特性だけを理解しているわけではなく，同時に，得た情報から次ステップの出来事も予測している．しかし現在の実世界理解の先行研究では，これら二つの機能を分けて考えている．そのため物理法則を理解し，それにもとづいた予測モデルの構築は必要と言える．

ヒトが環境の物理特性をとらえ，その変化点を捉えた研究として，graph Variational Temporal Abstraction（graph VTA）[2] がある．graph VTA は，系列データ内に潜在的に含まれる時間的階層構造を取得する先行研究 Variational Temporal Abstraction (VTA) [4] を，画像特徴量からではなく系列データ内の物体の速度や加速度といった物理特性から取得できるよう拡張させた研究である．

次にヒトの環境予測機能を表した研究として PredNet[3] がある．PredNet はヒト脳内の大脳皮質における予測符号化を模した機械学習モデルであり，動画像を与えたとき次の画像を予測するモデルとして提案されている．ヒトの予測機能は観測と脳内での推論の誤差を小さくすることで，実観測とのズレを小さくする．この仕組みは大脳皮質における予測符号化処理によりなし得ているとされる．

本研究では PredNet と graph VTA の機構を組み合わせて，ヒト脳内での予測符号化を模した変化点予測モデルを構築する．モデルの概要を図 1 に示す．提案モデルは PredNet の階層構造を並列にした形になっており，さらに graph VTA の機構である変化点判別フラグ m を取り入れて構築した．入力情報は，画像情報の CLEVRER dataset[5] と CLEVRER から作成した物理特性を含む学習データセット (図 1 の physical training data) の 2 つとした．出力情報は，画像について逐次的に予測した予測画像 (図 1 の img output) と物理特性を表した埋め込みベクトルの推論から算出した変化点 m(図 1 の m_a output) の 2 つである．変化点 m は物理データか画像データのどちらか一方が大きく変化したタイミングを表すフラグであり，0 か 1 の値をとる．提案モデルは図 1 のように物理データの処理機構と画像データの処理機構に分かれており，表現層 R から推論した予測 $\hat{A}$ と実観測 A の差分が小さくなるように上位層にエラーを伝播させることで学習を行う．また変化点 m の決定は，時刻 $t-1$ と時刻 t における表現層 R の差分 $diff$ を物理データと画像データのそれぞれで算出し，差分 $diff$ が閾値 α を超えたとき変化点 m が 1 となるようにした．

A Study on the Construction of an Inflection Point Prediction Model Imitating Predictive Coding in the Human Brain Under Physical Environments
†Eri Kuroda (kuroda.eri@is.ocha.ac.jp)
†Ichiro Kobayashi (koba@is.ocha.ac.jp)

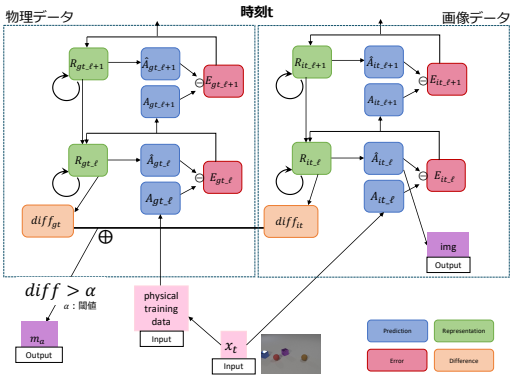

図 1: モデル概要図.

3 実験

提案モデルが正しく次ステップの状態の変化点を抽出できるか，モデルの有効性を検証する実験をした．学習における設定は先行研究 [3] を参考にした．使用データセットは CLEVRER データセット [5] と CLEVRER から作成した physical training data を用いた．作成手順は図 2 に示す．データセット数は 60 万，学習回数は 50 万，変化点を示すフラグの本数は 10 本とした．

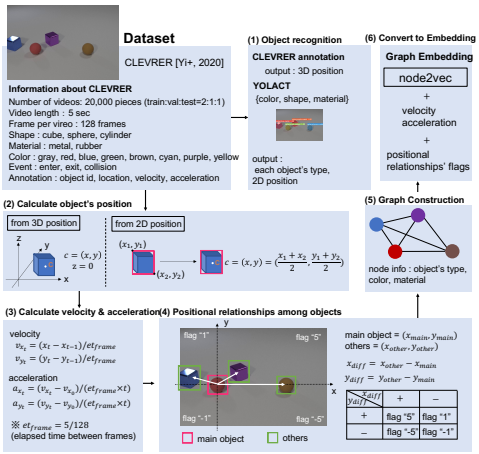

図 2: physical training data 作成手順.

3.1 結果と考察

精度の算出は 6 種類 (i~vi) の検証範囲で行った．iii と v では場面変化があり，それ以外では物体の衝突があった．正解のタイミングは CLEVRER のアノテーション情報の衝突データから取得した．また衝突のタイミングについて，アノテーション情報の衝突データは 3 次元空間において，物体同士が接触した瞬間を指す．一方で環境を 2 次元の画像で見たときは，物体の速さや移動方向の変化から衝突を認識する．これらのタイミングには約 2 フレームの差があったため正解には幅をもたせ，精度の算出は（正解のフラグ数）/（m=1 の本数)(%) とした．

予測変化点の抽出結果．提案モデルにおける変化点予測の精度結果を表 1 に示す．

結果から，physical data での精度は正解データであるアノテーションデータから作成した精度と同等の精度で変化点の予測ができているとわかる．また範囲 i での予測画像とフラグを図 3 に示す．

表 1: 提案モデルにおける予測精度比較.

検証範囲	i	ii	iii	iv	v	vi
physical data	33.3	**50**	50	33.3	**66.7**	**50**
annotation data	**66.7**	**50**	**66.7**	**40**	50	**50**

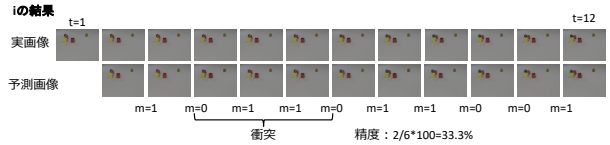

図 3: 範囲 i での予測変化点抽出結果.

graph VTA vs. 提案モデル．次に graph VTA[2] と提案モデルの優位性を検討するために，実データにおける変化点抽出の精度を比較した．

表 2: graph VTA との精度比較.

	検証範囲	i	ii	iii	iv	v	vi
graph VTA	physical	75	50	33.3	50	40	50
	annotation	**100**	**100**	33.3	**66.7**	25	**100**
Ours	physical	50	66.7	**50**	33.3	50	50
	annotation	75	75	**50**	**66.7**	**66.7**	**100**

physical training data を用いた場合，提案モデルの精度は graph VTA と同等かやや優位であることがわかる．また正解データのアノテーションデータでの結果を比べても，提案モデルの精度は同等と言える．ここから提案モデルは PredNet に変化点を示す機構を組み込んだ新たな変化点予測モデルであり，学習も正しく機能している．さらに本モデルはヒト脳内で行われているとする予測符号化を表した，観測の物理的な変化から次ステップの状態の変化点も予測することができるモデルといえる．

4 おわりに

本研究ではヒト脳内で行われているとされる予測符号化を模した変化点予測モデルの構築を行った．そして実験を通じて提案モデルの有効性についても検証した．

謝辞

本研究は特別研究員奨励費（22J21786）の助成を受けたものである．

参考文献

[1] David Ha and Jürgen Schmidhuber. World models. March 2018.

[2] 黒田彗莉, 小林一郎. 画像内の物体に着目した動きの変化点抽出への取り組み. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 2M1OS19a02–2M1OS19a02, 2022.

[3] Lotter, Kreiman, and Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv:1605. 08104 [cs, q-bio]*, February 2017.

[4] Taesup Kim, Sungjin Ahn, and Yoshua Bengio. Variational temporal abstraction. *CoRR*, Vol. abs/1910.00775, , 2019.

[5] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and B. Joshua Tenenbaum. Clevrer: Collision events for video representation and reasoning. *ICLR*, 2020.