物理環境を捉えたヒトの実世界認識解明への取り組み

理学専攻・情報科学コース　　黒田 彗莉

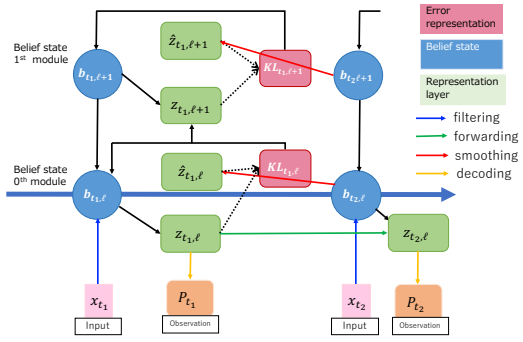
1　はじめに

近年，機械学習や深層学習を用いた研究が広く行われ，我々の生活にもその技術の一部が浸透している．またそれらの技術を用いた深層学習モデルを作業モデルと，ヒト脳の機能解明などの研究も進められている．一方で予測を対象にした先行研究の多くは画像ピクセル値の変化に着目し，画像内の物体の振る舞いには着目できていないという問題点もある．

本研究では，大脳皮質における予測メカニズムを模倣した深層学習モデルである PredNet [1] と，様々な時間間隔での将来の予測が可能である深層学習モデル TD-VAE [2] を用いて，様々な時間幅で予測を行えるヒト脳内の情報処理機構を模倣した予測を行う深層生成モデルを構築した．また構築したモデル [3] の有効性を実験を通して検証した．さらに予測を対象にした研究の問題点を解決するために，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）[4] 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．そして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　柔軟な時間幅での予測可能なモデル

PredNet [1] と Temporal Differential Variational Auto-Encoder （TD-VAE） [2] の機能を統合し，画像予測のための新たな深層学習モデルを構築した．構築したモデルの概要図を図 1 に示す．



$$L = KL(z_{t_1,\ell}, \hat{z}_{t_1,\ell}) + KL(z_{t_1,\ell+1}, \hat{z}_{t_1,\ell+1}) + \log P(z_{t_2,\ell}|b_{t_2,\ell}) - \log P(z_{t_2,\ell}|z_{t_1,\ell}) - \log P(P_{t_2}|z_{t_2,\ell})$$

図 1: 提案モデル 概要図.

図 2: 時間間隔 1 秒での予測.

2.1　実験設定

2.1.1　実験 1：柔軟な時間幅での予測画像生成

予測間隔の柔軟性を PredNet と提案モデルで比較した．予測時の時間幅は 1 秒とし，学習のパラメータは先行研究 [1][2] の設定に基づいた．

2.1.2　実験 2：脳活動情報との相関関係

構築した予測モデルを用いて fMRI データとモデルの隠れた状態の値との相関を観測し，モデルの予測に対する性能を調査した．各モデルの Representation 層における特徴表現と脳活動との対応関係を確認するためリッジ回帰による推定を行い，推定した特徴表現 {R0,R1,R2,R3} と各モデルに刺激動画像を適用して得られた特徴表現との相関係数を求めた．しかし R1 は内部状態が他層と比べても高次元であり，リソースの制約から推定を行わなかった．また TD-VAE は階層的な構造を持たないモデルのため，最下層 R0 のみリッジ回帰を行った．

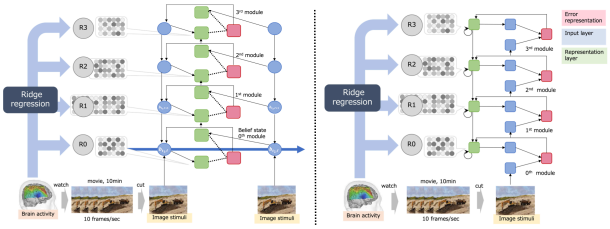


図 3: 脳活動との相関関係

表 1: 相関係数.

	PredNet			TD-VAE			提案モデル		
α	0.5	1K	25K	0.5	1K	25K	0.5	1K	25K
R0	0.2623	0.2971	0.3207	0.2636	0.2983	0.3285	0.2637	0.2983	0.3291
R2	0.0925	0.1459	0.1955	-	-	-	0.0003	0.0012	0.0016
R3	0.0254	0.1217	0.1871	-	-	-	0.0004	0.0009	0.0012

2.2　結果と考察

実験 1 における予測結果を図 2 に示す．実験結果から PredNet は時間幅を長くすると出力画像にズレが生じるが，提案モデルは比較的実画像に近い予測画像を生成することが確認できた．

次に実験 2 おける相関係数を表 1 に示す．先行研究 [1][2] を含む全てのモデルに対して，脳活動データから推定した特徴表現 R0 と実際の R0 の相関係数は α が 2.5×10^4 のとき約 0.32 であり、これは有意な相関関係を示していると言える．また PredNet と提案モデルを比較すると，両者ともに R2 と R3 については相関が見られなかった．しかし R0 の相関係数について，提案モデルは PredNet よりもわずかに高いことが確認できた．一方で TD-VAE と提案モデルを比較すると，相関係数の差は小さいことがわかる．原因として，提案モデルは TD-VAE と同様に R0 のみで推論を行うためであり，その影響により R2 および R3 の相関係数の値も小さいことがわかる．

表 2: 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 (node2vec)	① graph only	50	**100**	-	-	-	-
	② graph+image	14.3	25	9.1	37.5	14.3	28.6
annotation (node2vec)	③ graph only	20	**100**	20	**100**	**50**	33.3
	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 (graph2vec)	⑦ graph only	-	-	-	-	-	-
	⑧ graph+image	0	20	0	33.3	20	0
annotation (graph2vec)	⑨ graph only	-	-	-	-	-	-
	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．

3 物理イベントの変化点抽出手法

先行研究 [1][2] や構築したモデルは画像ピクセル値の変化から予測画像を生成しており，画像内の物体の振る舞いに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系列情報全て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と考える．本項目は Variational Temporal Abstraction (VTA)[4] を改良し，観測した環境について正しく認識したうえで，観測時に生じた重要なイベントや場面の選択を可能にする手法を提案する．

提案手法の概要図を図 4 に示す．YOLOv3 [5] 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) を取得する方法と，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・加速度を取得する方法でグラフを構築した．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の作成においては node2vec [6] と graph2vec [7] の 2 種類の手法を用いた．

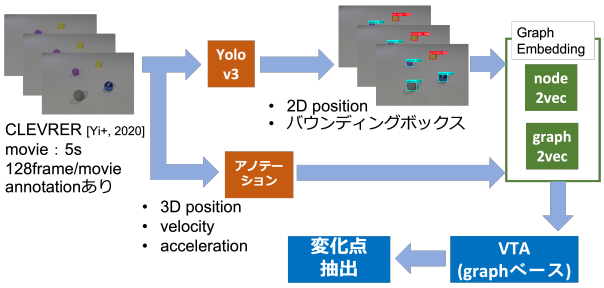


図 4: 提案手法の概要図

3.1 実験設定

学習における設定は先行研究 [4] 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 2 にある 11 種類を対象にした．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 128 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 を観測した場合と約 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 2 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 [4] 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，物体ごと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

本研究では柔軟な時間幅での予測が可能なヒト脳内の情報処理機構を模倣した深層生成学習モデルを構築した．また実際のヒト脳活動データと合わせることで，構築した予測モデルの有効性を示した．そして従来の予測モデルにあった問題点を解決するために，観測した環境において大きな変化が起きている場面を表現した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．結果としてグラフ構造とそれぞれの物体の位置関係を用いることで変局点抽出の精度が高くなることが判明した．

参考文献

[1] W. Lotter, G. Kreiman, and D. Cox. ”Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning.” arXiv preprint arXiv:1605.08104.

[2] K. Gregor, G. Papamakarios, F. Besse, L. Buesing, and T. Weber. ”Temporal Difference Variational Auto-Encoder.” In ICLR, 2019.

[3] E. Kuroda, S. Nishimoto, S. Nishida and I. Kobayashi. ”A Deep Generative Model imitating Predictive Coding in the Human Brain.” In ISIS, 2021.

[4] T. Kim, S. Ahn and Y. Bengio. ”Variational Temporal Abstraction.” arXiv:1910.00775.

[5] J. Redmon and A. Farhadi. ”YOLOv3: An Incremental Improvement.” arXiv:1804.02767.

[6] A. Grover and J. Leskovec. ”node2vec: Scalable Feature Learning for Networks.” arXiv:arXiv:1607.00653.

[7] A. Grover and J. Leskovec. A. Narayanan, M. Chandramohan, R. Venkatesan, L.Chen, Y. Liu and S. Jaiswal. ”graph2vec: Learning Distributed Representations of Graphs.” arXiv:arXiv:1707.05005.