深層学習を用いた系列データからの画像生成

理学専攻・情報科学コース　藤山 千紘（指導教員：小林 一郎）

1　はじめに

近年，汎用人工知能の構築を目指し，神経科学分野でのヒトの知能に関する知見を取り入れた人工神経回路網の研究が盛んに行われている．また，深層学習技術の著しい発展に伴い，生成モデルが数多く提案されているが，それらの多くは経験的な研究成果の報告に終始している．本研究では，ヒトの知能のメカニズムを模倣して構築された 2 つの人工神経回路網，予測画像生成モデルおよび自然言語を入力とする画像生成モデルを対象に，生成に加えて，モデルの特徴表現や内部計算機構の分析を行った．

2　予測画像生成モデル

2.1　予測符号化

神経科学分野では，大脳皮質で予測符号化と呼ばれるメカニズムが機能しているとする理論仮説が一定の支持を得ている．PredNet[1] は，このメカニズムを模倣した予測画像生成モデルとして提案されている．PredNet は同一構造のモジュールの積層になっており，各モジュールは，脳内の予測モデルに相当する Representation モジュール，入力処理を行う Input モジュール，予測を行う Prediction モジュール，予測と入力の差分を生成する Error ユニットから構成される．

2.2　PredNet の特徴表現と脳活動の相関関係

本提案手法では，PredNet の特徴表現と動画視聴時のヒトの脳活動データの相関を評価する．はじめに PredNet の学習を行い，得られたモデルに対して脳活動測定時の刺激動画像を入力し，その際の Representation モジュールにおける特徴表現と脳活動との対応関係を Ridge 回帰を用いて学習する．続いて，学習した Ridge 回帰を用いて脳活動から特徴表現を推定し，推定特徴表現と PredNet 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算出する．PredNet の一部および脳活動との対応関係を図 1 に示す．

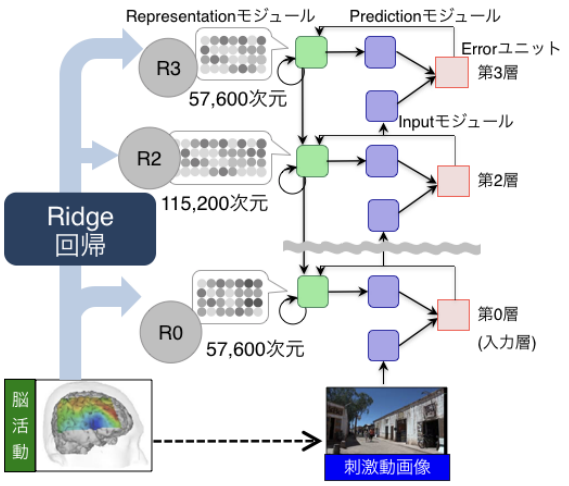

図 1: PredNet の特徴表現と脳活動の対応関係

2.3　脳活動データを用いた予測画像生成

予測画像生成時には，脳活動から Ridge 回帰を用いて推定した特徴表現で PredNet の Representation モジュールの出力値を置換した上で画像生成を行う．

3　画像生成モデル

3.1　自然言語を入力とする画像生成モデル

自然言語で記述されたキャプションを入力として画像を生成する深層学習モデルの一つに alignDRAW[2] がある．alignDRAW は，ヒトが絵を描く際の，「特定の言語表現に着目してそれに対応した部分を描く」というプロセスの反復を実現するべく，言語エンコーダ，注意機構，DRAW デコーダを用いて構成されている．

3.2　単語分割タスクを含む画像生成

既存手法ではすでに単語分割されているキャプションから画像を生成しているが，本研究では単語分割されていないキャプションと画像のデータセットを新たに作成し，単語の境界情報が欠落した場合の alignDRAW の言語エンコード能力および画像生成能力を評価する．具体的には，単語分割されていないキャプションに対して，妥当な画像を生成し得るか，その際の注意機構の挙動が言語の意味の単位を表現しているかを考察する．

3.3　埋め込み空間における言語の意味の構成的特性の分析

本分析では，既存モデルの言語エンコーダに埋め込み層を追加する形でモデルの拡張を行なった上で，単語分割済みのキャプションを用いて学習を行い，埋め込み層において単語の意味の構成的特性が表現されるかを調べる．具体的には，「左」,「上」,「左上」など空間を意味する単語について，埋め込み空間での加法構成性が見られるかを評価する．

4　実験

4.1　PredNet における脳活動との相関性調査

PredNet の学習に関するハイパーパラメータは先行研究 [1] に従い，学習データセットには，参考文献 [3] で用いられている自然動画像データセットを用いた．PredNet の特徴表現と脳活動の対応関係の学習に際しては，動画視聴時の被験者の血中酸素濃度に依存する信号 (BOLD 信号) を functional magnetic resonance imaging(fMRI) を用いて記録した脳神経活動データのうち皮質に相当する 65,665 ボクセルを説明変数とし，PredNet 各層の Representation モジュールを最下層から順に R0，R1，R2，R3 として取り出したものを目的変数として，Ridge 回帰を学習した．

Ridge 回帰を用いて推定された特徴表現 R0 と刺激画像から得られた最下層の特徴表現の相関係数は，0.249 となり，これはノイズの多い脳活動を扱う脳神経科学分野においては相関を認めるに値するとの知見がある．一方，より深層部に相当する R2 および R3 はほとんど相関を認められない結果となった．なお第 1 層の特徴表現 R1 については特徴表現が非常に高次元であるため Ridge 回帰の学習を行っていない．

脳活動データから推定した特徴表現を用いて予測画像の生

成を行った例を図 2 に示す．

図 2: 刺激画像と脳活動から推定した特徴表現を用いた生成画像，および PredNet による予測画像の例

R2 や R3 は脳活動との相関がほぼ認められないにも関わらず R0 を推定した場合の生成画像と比較して視認性が高い画像を生成できているが，これは Error ユニットの作用に因る．また参考として学習用脳活動データから推定した特徴表現 R0 を用いて予測画像を生成した例を図 3 に示す．

図 3: 学習用脳活動から推定した特徴表現 R0 を用いた生成画像例

学習用脳活動データを用いた場合には，評価用データの場合と比べて視認性の高い画像を生成できている．これは対応関係の学習に用いるモデルとして Ridge 回帰がある程度の妥当性をもつことを示唆している．

4.2 alignDRAW における言語と画像の対応関係分析

alignDRAW に関する実験では，データセットを，表 1 に示すテンプレートを用いて作成したキャプションと，手書き数字画像のデータセット MNIST[1]をキャプションに適合するように配置した画像で構成した．各実験は表 2 の設定で行った．なおモデルの実装には深層学習のフレームワーク TensorFlow[2]を利用した．

表 1: キャプション作成時のテンプレート

(A) 単語分割を含む画像生成	(B) 構成的特性の分析
すうじ＿がすうじ＿のひだりにある．	数字＿が画像の左にある．
すうじ＿がすうじ＿のみぎにある．	数字＿が画像の右にある．
すうじ＿がすうじ＿のうえにある．	数字＿が画像の上にある．
すうじ＿がすうじ＿のしたにある．	数字＿が画像の下にある．
すうじ＿ががぞうのひだりうえにある．	数字＿が画像の左上にある．
すうじ＿ががぞうのみぎしたにある．	数字＿が画像の右下にある．
すうじ＿ががぞうのみぎうえにある．	数字＿が画像の右上にある．
すうじ＿ががぞうのひだりしたにある．	数字＿が画像の左下にある．

学習の結果，実験 (A)，(B) ともに，図 4 のように，キャプションに対して妥当な生成画像を得た．

単語分割を含む画像生成については，注意機構の挙動が意味の単位を表現していることが期待されたが，本設定下ではそのような傾向は見られなかった．これは言語エンコーダの表現能力が高いことが一因であると考えられ，言語エンコーダの簡素化を行ってさらなる考察を加えたが，紙面の都合上本要旨では割愛する．

埋め込み空間における言語の意味の構成的特性の分析については，「左上」，「左下」，「右上」，「右下」の 4 単語について，それぞれ「左」，「右」，「上」，「下」に対応する分散表現を使って推定したベクトルと，学習により得られた各分散表現の cos 類似度を評価した．その結果を表 3 に示す．

表 2: alignDRAW 学習時のハイパーパラメータ

	(A) 単語分割を含む画像生成	(B) 構成的特性の分析
言語エンコーダ	1hot → 128 ユニット双方向 LSTM	32 次元分散表現 → 128 ユニット双方向 LSTM
注意機構	512 ユニット Bahdanau Attention	256 ユニット Bahdanau Attention
デコーダ	300 ユニット DRAW LSTM	300 ユニット DRAW LSTM
潜在変数 z	150 次元	150 次元
最適化アルゴリズム	RMSProp	RMSProp

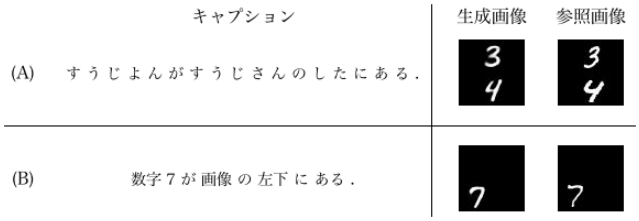

図 4: 生成画像例

表 3: 推定分散表現と実分散表現の cos 類似度

	左上	左下	右上	右下
cos 類似度	0.89	0.80	0.29	0.92

本設定下では，「右上」以外の 3 単語については比較的高い cos 類似度が得られ，埋め込み空間において言語の意味の構成的特性が表現される可能性を示唆する結果を得た．

5 おわりに

本研究では，2 つの生成モデルを対象に，画像の生成および，ヒトの知能のメカニズムとの親和性の観点からの内部計算機構や特徴表現の分析を行った．予測画像生成モデルについては，モデルの特徴表現の一部に動画視聴時の脳活動データとの相関を認めた．自然言語を入力とする画像生成モデルについては，入力の粒度変更時の内部計算機構の挙動を観察し，また埋め込み空間での言語の意味の構成的特性の評価を行った．今後の課題としては，より精緻な分析を行うことや，ヒトの知能のメカニズムを反映して動作する生成モデルの構築が挙げられる．

参考文献

[1] Lotter, W., Kreiman, G., Cox, D., “Deep predictive coding networks for video prediction and unsupervised learning.” in ICLR, 2017.

[2] Mansimov, E., Parisotto, E., Ba, J. L., Salakhutdinov, R., “Generating images from captions with attention.” in ICLR, 2016.

[3] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J. L., “Reconstructing visual experiences from brain activity evoked by natural movies.” Current Biology, 21(19) (pp.1641-1646), 2011.

[1] http://yann.lecun.com/exdb/mnist/
[2] https://www.tensorflow.org/