画像刺激における脳活動状態推定への取り組み

田口　遥香　(指導教員：小林 一郎)

1　はじめに

近年，脳神経科学分野において，脳神経活動を定量的に理解する研究が盛んに行われている．また，Yaminsら [1] によって視覚情報を処理する深層学習モデルの層と脳内の視覚情報処理の階層的処理の層の間に相同性があることが示されている．このような背景から深層学習モデルを作業モデルとして援用することによりヒト脳内の情報処理機構の解明を目指す研究が盛んになっている．本研究では，画像刺激が与えられた際のヒト脳の状態を推定することを目的とし，特に作業モデルとして画像処理のための深層学習モデルによる推定精度について検証を行う．

2　画像識別深層学習モデル

ニューラルネットワークを用いた画像処理には，脳内の視覚野における情報処理機構を模倣して提案された畳み込みニューラルネットワーク (CNN) が使われる．CNN は，全結合層だけでなく畳み込み層とプーリング層から構成されるニューラルネットワークとなっており，ネットワークの層が入力から深くなるほど情報が抽象化され，観測された情報を識別可能になる．画像識別の深層学習モデルには，そのアーキテクチュアの工夫により様々なモデルが存在するが，本研究では，画像識別深層学習モデルのベースラインと見なされる VGG16 およびネットワークアーキテクチュアに「スキップ構造」という仕組みを取り入れることで層を深くすることを可能にし画像識別性能を向上させた代表的な CNN である ResNet-50 を使用する．

2.1　VGG16

VGG16 [2] は，畳み込み層が 13 層，全結合層が 3 層の合計 16 層からなる CNN である．小さいフィルターを持つ畳み込み層を 2〜4 回連続して重ね，それをプーリング層でサイズを半分にすることを繰り返し行う構造を特徴として持つ．大きいフィルターで画像を一気に畳み込むよりも小さいフィルターを何個も畳み込み層を深くすることで特徴をより良く抽出できる工夫をもつ．ネットワークのアーキテクチュアがシンプルであることから様々な用途で用いられている．

2.2　ResNet50

ResNet-50 [3] は，深さが 50 層の CNN である．スキップ構造を取り入れることによって，より深く層を重ねることを可能にしている．スキップ構造とは，手前の層の入力を後続の層に直接足し合わせることである層で求める最適な出力を学習するのではなく，層の入力を参照した残差関数を学習する手法である．出力から入力を引いた残差 $F(x) = H(x) - x$ を計算することで $H(x)$ が小さくなり，勾配損失問題が軽減される．

VGG16 および ResNet-50 において，ImageNet[1]の 1000 クラス分類問題を学習した際の性能比較を表 1 に示す．ここで，Top-1 エラー率は ImageNet データセットでの確率が一番高い予測ラベルが正解ラベルと一致していない割合を示し，Top-5 エラー率は ImageNet データセットでの確率が高い上位 5 個の予測ラベルに正解ラベルが含まれていない割合を示す．この表から，パラメータ数およびエラー率から VGG16 よりも ResNet-50 の性能が優れていることがわかる．

表 1: VGG16 と ResNet50 の性能

モデル名	パラメータ数	Top-1 エラー率	Top-5 エラー率
VGG16	138357544	28.41	9.62
ResNet-50	25557032	23.85	7.13

[1]http://www.image-net.org/

2.3　AutoEncoder

オートエンコーダ（AutoEncoder) とは，ニューラルネットワークの 1 つである．入力されたデータを一度低次元のデータに圧縮する．その際，元のデータを再構成できるように特徴量を抽出し，再度，元の次元にデータを復元処理をするものである．このように，オートエンコーダのエンコーダ部分は次元削減および特徴抽出の機能を有し，デコーダ部分は低次元の情報をソースとするデータ生成機能を有する．

3　画像特徴量に基づく脳活動状態の推定

脳活動測定時の刺激動画像を静止画像として切り出したものを，CNN(ここでは，VGG16 と ResNet を使用) を用いて画像特徴量を抽出したもの, および，オートエンコーダを用いて中間層の画像特徴量を抽出したものから Ridge 回帰を用いて脳活動データを推定する (図 3 参照).

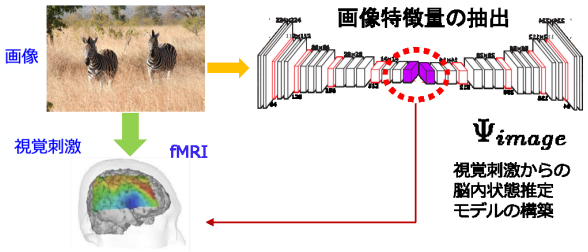

図 1: 画像特徴量から脳内状態の推定

4　実験

4.1　実験設定

ResNet-50 および VGG16 は ImageNet [4] と呼ばれる大規模な画像データセットを使って事前に学習したモデルを使用する．また，事前学習済み ResNet-50 をエンコーダおよびデコーダとして，オートエンコーダを構築し，動画像刺激として脳に与えられた動画を切り出した静止画を学習データ（train data）として学習したモデル（ResNet-50+train data）を使用する．

脳活動データとしては，情報通信研究機構脳情報通信融合研究センター（NICT CiNet）より提供された

表 2: 画像特徴量から脳活動データを推定した際の相関係数

	正則化項 α											
model	0.5	1.0	5.0	10.0	10^2	10^3	10^4	2.5×10^4	5.0×10^4	10^5	10^6	10^7
(i) ResNet-50+4 sec.	0.0258	0.0271	0.0329	0.0364	**0.0536**	**0.0778**	**0.1056**	**0.1139**	**0.1166**	**0.1160**	0.0716	0.0168
(ii) RestNet-50+train data	**0.0260**	**0.0303**	**0.0382**	**0.0403**	0.0437	0.0408	0.0352	0.0340	0.0337	0.0335	0.0283	0.0178
(iii)RestNet-50+4,5,6,7 sec.	0.0104	0.0115	0.0160	0.0189	0.0318	0.0523	0.0835	0.0955	0.1024	0.1068	0.0969	0.0287
(iv)RestNet-50+4,5,6 sec.	0.0212	0.0223	0.0267	0.0288	0.0362	0.0520	0.0800	0.0913	0.0979	0.1021	0.0922	0.0275
(v)RestNet-50+4,4.5,5,5.5 sec.	0.0097	0.0110	0.0148	0.0169	0.0292	0.0528	0.0897	0.1023	0.1083	0.1111	0.0929	0.0248
(vi)VGG16+4 sec.	0.0052	0.0065	0.0118	0.0153	0.0309	0.0537	0.0841	0.0953	0.1018	0.1060	0.0908	0.0237
(vii)VGG16+4,5,6,7 sec.	0.0138	0.0140	0.0149	0.0155	0.0202	0.0356	0.0634	0.0746	0.0823	0.0890	0.0962	**0.0439**
(viii)VGG16+4,5,6 sec.	0.0168	0.0170	0.0178	0.0186	0.0246	0.0404	0.0686	0.0808	0.0890	0.0954	**0.0972**	0.0379
(ix)VGG16+4,4.5,5,5.5 sec.	0.0221	0.0222	0.0226	0.0230	0.0259	0.0374	0.0639	0.0747	0.0820	0.0881	0.0936	0.0423

動画像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 96 × 96 × 72 ボクセルのうち皮質 に相当する 65,665 次元のデータを使用した.

画像特徴量を抽出する際の画像は，脳活動測定時の刺激動画像を 静止画像として 1 秒間に 20 枚ずつ切り出し，画像のサイズは，VGG16，ResNet-50 の入力次元に揃え，224 × 224 とした．また，画像特徴量のサイズは VGG16 は 4,096 次元， ResNet-50 は 2,048 次元のデータを使用した.

被験者が動画を見てから fMRI で観測される脳活動に影響が出るまでに約 4〜6 秒の時間がかかる．そこで，ResNet-50 および VGG16 から抽出される画像特徴量と脳活動データの対応を 4 秒ずらして構築した対データを用いたそれぞれのモデルを表 2 中の (i)(vi) として表す．同様に，4，5，6，7 秒のずれがあることを想定してそれらの秒数で得られる画像特徴量を一つにして脳活動データと対データをを用いたモデルを表 2 中の (iii)(vii) として表す．また，4，5，6 秒ずらしたものを表 2 中の (iv)(viii)，4,4.5,5,5.5 秒ずらしたものを表 2 中の (v)(ix) として表す．さらに，ResNet-50 によって構成されるオートエンコーダに train data を学習させたものを用意し，4 秒ずらしたものの画像特徴量を用いたモデルを表 2 中の (ii) として表す.

これらのモデルに対して，画像特徴量と脳活動の対データ（学習データ 4,497 対，評価データ 300 対）を用いて Ridge 回帰を行い，推定された脳状態をピアソン係数を用いて相関係数を調べた.

4.2 実験結果

表 2 に Ridge 回帰の正則化項 α を変えた時の相関係数を示す．これから，ResNet-50 に 4 秒ずらした静止画と回帰したもの, オートエンコーダに train data を学習して回帰したものの相関係数の値が良いことがわかる．また，正規化項が大きくなると ResNet-50 よりも VGG16 の方がよくなった.

4.3 考察

画像識別性能を反映して VGG16 よりも ResNet-50 の方が総じて良い結果が出たと考えられる．このことから，さらに画像識別能力が良いモデルを使うことにより，精度向上の可能性が期待できる．一方で，正規化項が 10^6 以上になると，VGG16 の方が精度が良くなった．また，1 番高い相関が見られたのは，正則化項が 5.0×10^4，10^5 の際の ResNet-50 の時であり，他のモデルでもこの正則化項の下，良い値が確認できる.

表 2 中 (ii) の,ResNet-50 によるオートエンコーダを train data を用いて学習させたものは，正規化項が大きくなるにつれ，それほど高い相関性が見れなかったのは，ResNet-50 の事前学習に使われている ImageNet のデータ約 1,400 万枚に対し，train data 数が 4,497 枚と少なかったことにより，学習が十分でなかったことが原因ではないかと考えている.

また，ImageNet は静止画であるのに対し，train data は動画を静止画として切り出したものであるため，動画としての影響が残り ImageNet のような鮮明な画像とは異なっていたことも原因として考えられる．より多くの画像を使ってモデルを再学習することにより，現在より精度の高いモデルを作ることができるのではないかと考えている.

画像特徴量と脳内状態活動との対データを作成する際に，脳の状態は緩やかに変化することを考慮して，複数時点の画像特徴量と脳活動データを対応させたデータに基づく回帰を試行したが，今回は 4 秒遅れを採用したものより，よい結果は得られなかった.

5 おわりに

本研究では，動画像視聴時に脳のどの部分が活性化しているか調べるために，Ridge 回帰を用いて脳活動データと ResNet-50 および VGG16 によって抽出した画像特徴量との相関関係を調べた．結果として，高い相関性を得ることはできなかったが，画像識別モデルの性能が回帰の性能に影響を与えることを確認した.

今後は，より高精度なモデルをつくるために対応データの構築を見直し動画のフレームレートを変更した画像を使用することで結果が変化するなどを確認したい．また，脳内の解剖学的な部位（ROI）を対象に脳状態を推定する回帰モデルを構築し検証したい.

参考文献

[1] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, pp. 8619 – 8624, 10/2014 2014.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, Vol. abs/1409.1556, , 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. cite arxiv:1512.03385Comment: Tech report.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE Computer Society, 2009.