2Q-08

深層モデルを用いた顕微授精支援に向けた分析 -深層モデルは精子の動画像をどのように捉えるのか-

藤井巧朗† 濱上知樹† 竹島徹平‡ 山本みずき‡ 上野寛枝‡ 湯村寧‡
† 横浜国立大学 ‡ 横浜市立大学附属市民総合医療センター

1 はじめに

近年，生殖医療による出生児の増加[1]や，厚生労働省の不妊治療の保険適用範囲拡充[2]など，生殖医療への期待が高まっている．顕微授精は生殖医療の1つで，胚培養士が選別した精子により受精を行う．しかし，精子の選別には高度な技術が必要であったり，胚培養士の人材不足などの課題がある．そこで，本研究では胚培養士の負担軽減や育成を目的とし，深層学習を用いた精子の選別を行う．具体的には，図1のように，精子の動画像から精子に対するグレード分布[3]を推定する．

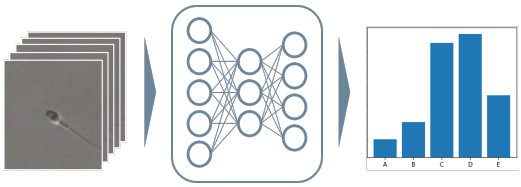

図1 グレード分布推定のフロー図

精子の選別に深層学習を導入する研究に関して，画像を入力とする研究 [1] は存在するが，動画像を入力とする研究は未だ行われていない．精子の選別には，形状だけでなく動き情報を捉える必要がある．そのため，本研究では動画像を用いて精子の選別を行う．

本研究の目的は，精子の動画像に特化した深層モデルを構築することである．その中で，既存モデルの中でどれが精子に適するのか，それらのモデルが精子をどのように捉えるのかを明らかにする必要がある．そのため，本稿では精子に対する画像・動画像モデルの網羅的な調査と，深層モデルが精子をどのように捉えるかの調査を行い，精子に適したモデルを選定する．

本研究の貢献は，深層モデルを用いた動画像からの精子の選別，特定ドメインにおける深層モデルの網羅的な調査，深層モデルの精子の捉え方の分析の3つである．

Analysis of deep neural net for supporting ART – How does the deep neural net recognize sperm image and video-

† {Takuro Fujii, Tomoki Hamagami}, Yokohama National University
‡ {Teppei Takeshima, Mizuki Yamamoto, Hiroe Ueno, Yasushi Yumura}, Yokohama City University

[1] https://www.jsog.or.jp/activity/art/2020_ARTdata.pdf
[2] https://www.mhlw.go.jp/content/12404000/000718601.pdf
[3] 本研究で用いられるグレード分布は，胚培養士40名が精子に対して5段階グレード (A-E) を付与した分布である．

2 関連技術

画像認識モデルには， [2, 3] などの CNN モデルと [4, 5, 6, 7] などの Transformer モデルがある．動画像認識モデルにも， [8, 9, 10, 11] などの CNN モデルと [12, 13] などの Transformer モデルがある．CNN は局所的な特徴を捉え，テクスチャに反応するのに対し，Transformer は大域的な特徴も捉えることができ，形状に反応する [14]．本研究で用いるデータセットは，精子が画像に占める割合は小さく，同じ動きを繰り返すのではなく長期に渡って異なる動きをする場合が多い．従って，空間次元は局所特徴を捉え，時間次元は大域特徴を捉える必要があると考える．

3 実験

■実験データ 本実験で用いるデータは615件あり，画像サイズ150×150，フレームレート16fpsの1秒間の動画像である．画像サイズは224×224にリサイズした．ただし，画像認識モデルによる分析には，初期フレームを用いる．

■タスク グレード分布を回帰タスクとし，損失関数にMSEを用いる．

■実験設定 バッチサイズ32，エポック数200とする．また，5分割交差検証により性能評価を行う．

■実験内容 3つの実験を行う．1つ目は種々の画像認識モデルでの性能評価，2つ目はVgg16とViTの説明性の可視化による精子の捉え方の分析，3つ目は種々の動画像モデルでの性能評価である．画像認識モデルの比較には，Vgg16[2]，Resnet[3]，VisionTransoformer[4]，SwinTransformer[5]，Convit[6]，MLPMixer[7]を，動画像認識モデルの比較には，R3D[9]，R(2+1)[9]，I3D[8]，SlowFast[10]，X3D[11]，TimeSformer[12]，ViViT[13]を用いる．用いるモデルは全て事前学習済みモデルである．

4 実験結果

■画像認識モデルの分析 種々の画像認識モデルの性能を表1に示す．Vgg16とSwinTransformerの性能が高い結果となった．SwinTransformerは局所的なAttentionを行い，層ごとに局所窓をずらすことで階層的に大域的な処理を行っている点でVgg16と類似している．従って，精子の空間次元には局所的な処理と階層的な構造が有効だと考えられる．

表 1　画像認識モデルでの性能評価．数値は 5 分割交差検証の平均 (標準偏差) を表す．

		MSE($\times 10^{-2}$)
CNN	vgg16 [2]	1.52 (0.07)
	resnet50 [3]	1.58 (0.04)
Transformer	ViT [4]	1.67 (0.12)
	Swin [5]	1.53 (0.09)
	Convit [6]	1.59 (0.08)
	MLPMixer [7]	1.61 (0.10)

■深層モデルは精子をどのように捉えるか　Vgg16 の GradCam の可視化結果と ViT の AttentionRollout[15] と TransformerExplanability[16] の結果を図 2 に示す．これらを比較すると，Vgg16 は精子の頭部しか捉えていないのに対し，ViT は頭部だけでなく尾も捉えていることが確認できる．これは，CNN はテクスチャに反応し，Transformer は形状に反応するためだと考えられる．つまり，CNN では頭部のテクスチャがタスクに有効だと判断し，頭部に反応したのだと考える．

ViT は尾まで捉えるにもかかわらず，性能が低い結果となった．これは，動く精子の一瞬を切り出した 1 フレームから推定を行うため，様々な精子の形状が出現することに加え，データ数が少ないことによって，正解ラベルとの関係性の学習に失敗し，学習が収束しないためであると考えられる．動画像の場合，このような問題は生じないため，精子の空間特徴は Transformer モデルにより捉えることが有効であると言える．

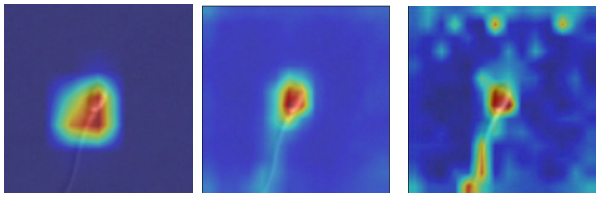

図 2　深層モデルの説明性の可視化．(左)vgg16 の GradCam, (中)ViT の AttentionRollout，(右)ViT の Explanability.

■動画像認識モデルの分析　種々の動画像認識モデルの性能を表 2 に示す．TimeSformer と SlowFast の性能が高い結果となった．SlowFast は時間幅の大きいカーネルを用いており，離れたフレームの関係性を獲得する点で，TimeSformer に類似している．従って，精子の時間次元には大域的な特徴を捉える構造が有効であると考えられる．

5　おわりに

本稿では，深層モデルにより精子のグレード分布推定を行う上で，どのようなモデルが適しているかの調査するため，種々モデルによる網羅的な分析を行った．また，CNN モデルと Transformer モデルが精子画像をどのように捉えるのかを分析した．その結果，空間次元では Vgg16 と SwinTransformer，時空間次元では SlowFast と TimeSformer が高い性能を発揮した．また，空間次元は局所的な処理と階層的な構造が，時間次元は大域的な特徴を捉える構造が有効であることが明らかになった．さらに，空間次元では，CNN モデルは精子の頭にのみ注目するのに対し，Transoformer モデルは頭だけでなく尾にも注目するため，Transoformer モデルの方が適していることが分かった．

表 2　動画像認識モデルでの性能評価．数値は 5 分割交差検証の平均 (標準偏差) を表す．

		MSE($\times 10^{-2}$)
CNN	R3D [9]	1.34 (0.08)
	R(2+1) [9]	1.70 (0.08)
	I3D [8]	1.45 (0.09)
	SlowFast [10]	1.30 (0.14)
	X3D [11]	1.71 (0.09)
Transformer	TimeSformer [12]	1.24 (0.08)
	ViViT [13]	1.35 (0.05)

今後は，性能の高かった SlowFast と TimeSformer に着目し、時間次元の捉え方の違いを分析する．さらに，そこで得た知見をもとに、両モデルをアーキテクチャレベルで組み合わせ，精子の動画像に特化したモデルを構築したい．

参考文献

[1] Jason Riordon, Christopher McCallum, and David Sinton. Deep learning for the classification of human sperm. *Computers in Biology and Medicine*, Vol. 111, p. 103342, 2019.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLM*, 2021.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[6] Stephaned' Ascoli, Hugo Touvron, Matthew L. Leavitt, Ari S. Morcos, Giulio Biroli, Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *ICML*, 2021.

[7] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *NeurIPS*, 2021.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[9] Du Tran, Heng Wang1, Lorenzo Torresani, Jamie Ray1, Yann LeCun1, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, 2019.

[11] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020.

[12] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[13] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021.

[14] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? In *CogSci*, 2021.

[15] Abnar Samira and Zuidema Willem. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. ACL, 2020.

[16] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021.