質問のパターンに着目した医療分野における Visual QA の検討

馬田　英雄† 　青野　雅樹††

† 豊橋技術科学大学 博士前期課程情報・知能工学専攻　〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
†† 豊橋技術科学大学 情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
E-mail: †umada@kde.cs.tut.ac.jp, ††aono@tut.jp

あらまし　近年，人工知能の研究が発展しており，特にコンピュータビジョン，自然言語処理，および知識表現＆推論を組み合わせた画像およびビデオキャプションの研究がここ数年の間で劇的に増加している．そのなかでも Visual QA (Visual Question Answering: VQA) は研究が非常に盛んな分野であり，コンピュータビジョンと自然言語の関連性を学習する必要がある．本研究では ImageCLEF VQA-Med タスクである医療画像に基づく質問応答システムについて実験を行った結果の分析を述べる．VQA-Med タスクでは，アノテーションとして 4 カテゴリに分類しているが，本研究では質問のパターンに着目したより詳細なカテゴリ分類，強力な分類器の問題を複数の弱い分類器の問題として解く手法と，分類器のアーキテクチャに関する検討を述べる．

キーワード　Visual QA, 深層学習, 質問応答, CLEF, VQA-Med

1 はじめに

Visual QA とは，図 1 のようなある画像とその画像に関する質問を提示されたときに正しい答えを導き出すタスクである．VQA ではコンピュータビジョンと自然言語処理の両方の問題に焦点を当てる必要があるが，ディープニューラルネットワーク (Deep Neural Networks: DNN) の研究の進歩により，マルチモーダルなデータに由来する困難な問題をある程度解決が可能になってきている．また医療分野における AI の活用として，自動化された医療画像解釈のアルゴリズムを生成および活用する機会が現在検討されている．医療画像に基づく VQA システムは臨床士の判断の確信度を高めるセカンドオピニオンとしての役割が期待されている．さらに患者自身が自分の疾患状況に関心があるが，多額のコストをかけずに病院に訪問する意思がない場合，多くはインターネットの検索エンジンを頼ると考えられる．しかし，検索エンジンからの結果は膨大かつ，誤解を招きやすい，あるいは誤ったものが多く，このシステムに代わるものとして VQA システムが期待される．しかし，医療分野において専門性の高い画像や用語に対するデータセットや学習済みモデルは比較的小規模，非汎用なものが多く，医療分野を正しく解釈できるモデルに対しては研究があまり進んでいないのが現状である．

本研究は ImageCLEF VQA-Med タスク [5] である医療画像に基づく質問を解答するシステムについての検討を述べる．VQA-Med タスクでは，Modality，Plane，Organ，Abnormality の 4 つのカテゴリを定義しているが，本研究ではカテゴリの細分化を行いより詳細なカテゴリ分けを行い，強力な分類器の問題を複数の弱い分類器の問題として解く手を検討する．また，学習済みの DNN から得られた特徴量を効果的に解釈を行う弱い分類器のアーキテクチャに関する検討を示す．

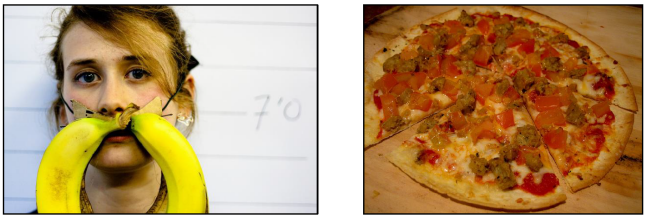


図 1: VQA の例（[4] より引用）

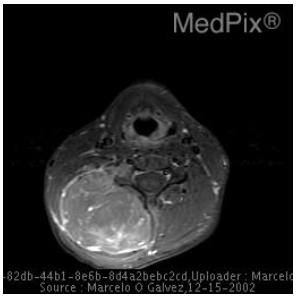
2 関連研究

VQA に関する研究は VQAv2.0 [9] を用いた VQA Challenge を通して盛んに行われている．QA で提案される DNN の多くは，質問文特徴量と画像特徴量をどのように合成するかという点に焦点が当てられており，特に近年は Attention を基にした DNN が多く提案されている．例えば P.Anderson et al. [3] による物体検出に用いられる畳み込みニューラルネット (Convolutional Neural Networks: CNN) である Faster R-CNN [8] を用いて得られる Bottom-up Attention を使用した DNN や Zhou Yu et al. [15] による 2 入力の特徴量の DNN に落とし込み合成を行う DNN である MFB などが提案されている．また，VQA-Med タスクとして ImageCLEF2018，2019 でコンペティションが行われており，VQA-Med2019 で提案された X.Yan et al. [13] の学習モデルや Y.Zhou et al. [14] の学習モデルなどで問題をカテゴリに分類したのち，それぞれを学習器で解く手法が提案されている．また，VQA-Med タスクは ImageCLEF2020 も開催されることが告知されており，さらなる画期的な手法が期待されている．

(a) Q: what is the modality?
A: be - barium enema

(b) Q: what imaging modality is used to acquire this picture?
A: xr - plain film

(c) Q: what organ system is primarily present in this image?
A: skull and contents

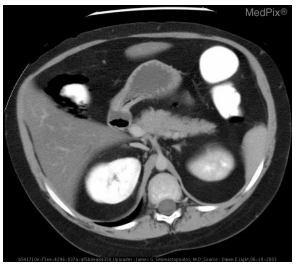


(d) Q: which organ is captured by this mri?
A: face, sinuses, and neck

図 2: (a), (b) Modality の例　(c), (d) Organ の例

(a) Q: what plane is demonstrated?
A: axial

(b) Q: what plane is the image acquired in?
A: sagittal



(c) Q: what abnormality is seen in the image?
A: juvenile dermatomyositis

(d) Q: what abnormality is seen in the image?
A: takayasu arteritis

図 3: (a), (b) Plane の例　(c), (d) Abnormality の例

3　VQA-Med データセット

データセットとして，ImageCLEF の VQA-Med-2019 Dataset [5] を用いる．本データセットは，Train，Validation，Test から成る．各データは医療画像と質問文，カテゴリ，解答の 4 つで構成される．カテゴリは Modality，Organ Systen，Plane，Abnormality の 4 つからなり，各カテゴリの詳細は表 1 に示す．各カテゴリでのデータについて具体的なデータとしてそれぞれ 2 組ずつ図 2〜 3 に示す．各カテゴリでの典型的な質問，解答の例として，各カテゴリでよく出現する質問，解答の上位 10 件をそれぞれ図 4，5 に示す．ただし，パターン数が 10 件に満たない場合はすべてのパターンを表示している．また本論文は検討段階であるため Test データに関しては実験・分析は行わない．

表 1: カテゴリ及びパターン数

	Modality	Organ	Plane	Abnormality	All
Training	3,200	3,200	3,200	3,192	12,792
Validation	500	500	500	500	2,000
Testing					500
質問パターン	44	78	84	45	251
解答パターン	45	10	16	1,632	1,701

4　VQA-Med における詳細カテゴリ

本研究では予めデータについているカテゴリとは異なり，質問と解答に着目した 7 つの詳細カテゴリを提案する．詳細カテゴリは modality，contrast，weighting，yesno，organ，plane，abnormality の 7 つで構成される．各カテゴリの詳細を表 2 に示す．質問カテゴリと詳細カテゴリの関係を図 8 に示す．この図は各カテゴリとその抱合関係を示している．上部にあるカテゴリは下部にあるカテゴリを抱合している．ただし 7 カテゴリ中の yesno カテゴリは 4 カテゴリ中の Abnormality 及び Modality 双方に含まれる．カテゴリの下の数字は Training，Validation でのデータセットの数を表している．各カテゴリについてはルールベースで識別しており，表 3 に示す正規表現で分類を行う．yesno カテゴリはその他となっているが，Training，Valdation の範囲では分類漏れがないことを確認している．また詳細カテゴリの各カテゴリでよく出現する質問，解答の上位 10 件をそれぞれ図 6，7 に示す．ただし，Plane，Organ カテゴリに関しては図 4，5 と同様であるため省略する．

この詳細カテゴリを導入することによって，膨大な解答パターンを持つ Abnormality の一部を低次元な分類問題に落とし込むことができること，及び比較的解答パターンが多い Modality カテゴリにおいて問題の整理ができることが期待できる．

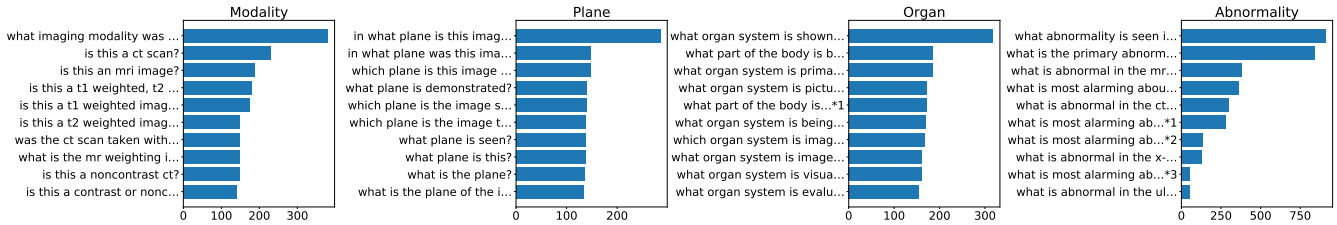


図 4: 各カテゴリにおける質問文の上位 10 件

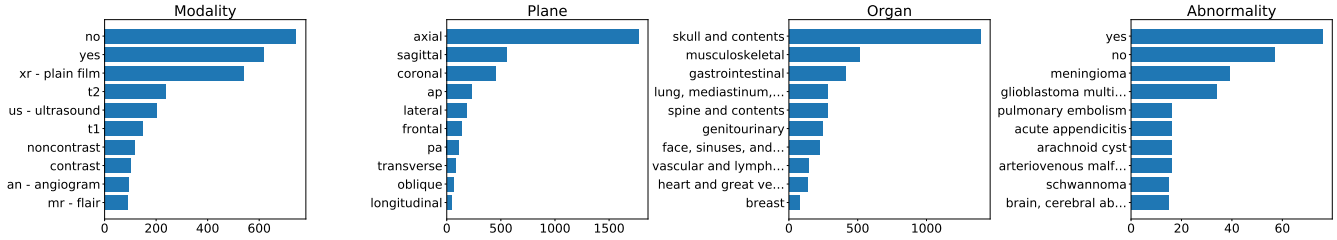


図 5: 各カテゴリにおける解答の上位 10 件

表 2: カテゴリ及びパターン数

	moda.	yesno	weig.	cont.	organ	plane	abno.
Training	1,435	1,294	394	187	3,200	3,200	3,082
Validation	235	202	54	32	500	500	477
質問パターン	29	35	3	2	78	84	20
解答パターン	38	2	3	2	10	16	1,630

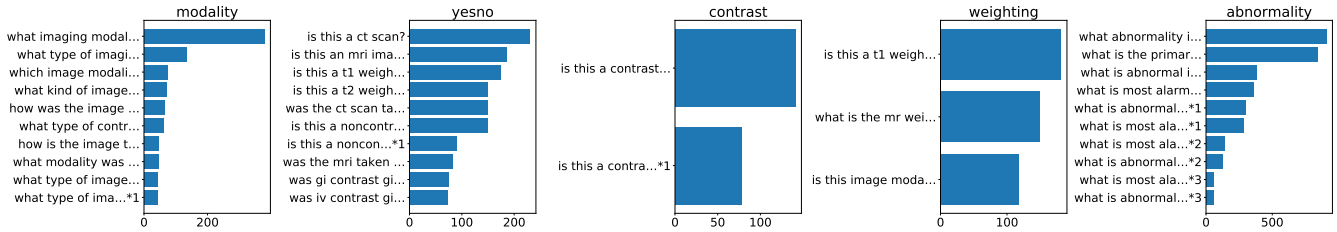


図 6: 各詳細カテゴリにおける質問文の上位 10 件

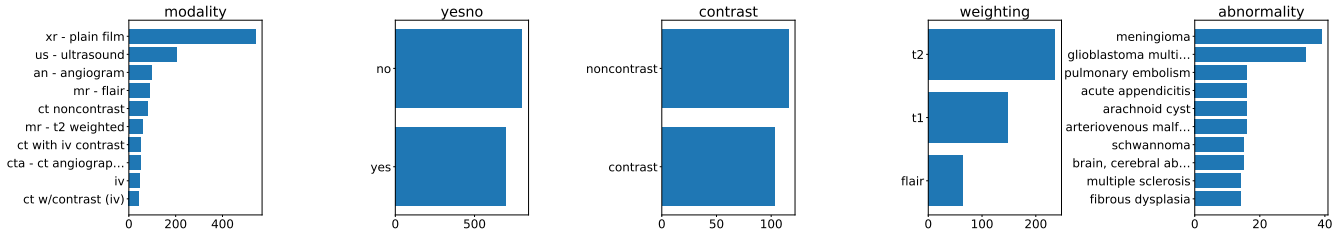


図 7: 各詳細カテゴリにおける解答の上位 10 件

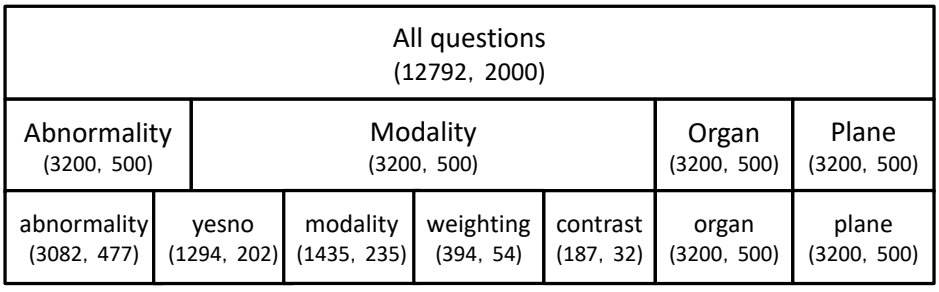


図 8: カテゴリと詳細カテゴリの関係

表 3: カテゴリ分類ルール

modality	what (type|kind) of (what|which) .*(modality|method) (what|how) (was|is) (this|the) image taken
organ	(what|which) .*organ what part of the body
plane	(what|which) .*plane
abnormality	what .*(abnormal|most alarming)
contrast	a contrast or noncontrast
weighting	t1.* t2.* flair what is the mr weighting
yesno	–others–

5 ベースラインモデル

ベースライン手法のシステム概要図を図 9 に示す．ここでは特徴量抽出とカテゴリ分類を行う．質問文の特徴量抽出器として Wikipedia と BookCorpus で事前訓練された BERT-Large [7] の最後から 2 番目のレイヤーを用いる．また画像特徴量抽出器としては Imagenet [6] で学習済みの VGG モデル [10] を用いる．特徴量抽出前の前処理は [1] に準ずる．その後，各分類器に特徴量を入力する．ベースラインでは単純な多層パーセプトロンを用いて，質問応答を分類問題として解く．今回の分類器では最終出力層を除いて同じ構造となっている．ここで FC 層の活性化関数は ReLU [2] を，正則化として dropout [11] を用いる．ただし最終 FC 層では活性化関数として Softmax を使用する．

6 検討モデル

6.1 検討 1 ：FC 共有化モデル

このモデルはベースライン手法に対して過学習を抑える目的で，モデルの制約を強める構造を実験する．今回のベースラインモデルと比較して，画像特徴量に対する FC 層を全体で共有化するモデルを検討する．この FC 共有化モデルの概要図を図 10 に示す．基本的な流れ及び活性化関数などの条件はベースラインモデルに従う．

6.2 検討 2 ：VGG GAP 特徴量抽出

このモデルは [13] に影響を受け，画像特徴量抽出として VGG に対して Global Average Pooling (GAP) で得られる特徴を使用するモデルである．本モデルでは [13] と同様の方法 (図 11) で特徴量抽出を行う．その後，図 12 のように分類を行う．ここで MFB [15] の Sum Pooling のカーネル幅は 5 とする．

6.3 検討 3 :VGG GAP 特徴量抽出＋ Channel Attention

このモデルは VGG に GAP を加えたモデルは意味情報が大きく欠損していると仮説を立て，画像特徴量抽出にチャネル情報を Attention としてモデルである．Attention には Channel Attention Module (CAM) と呼んでいるアーキテクチャを使用した．このアーキテクチャは [12] の CAM にインスパイアを受けたものであるが一部構造を変更した．この CAM は VGG の CNN 最終層及び Max Pooling 最終層に適用する．また，各分類器の構造は検討 2 の分類器に準ずる．

7 実験・結果

ベースラインモデル及び FC 共有化モデルでは [1] の VGG11_bn モデルを使用した．また VGG GAP 特徴量抽出モデルでは [13] に則り，VGG16 を使用した．評価手法には Accuracy を用いる．各検討モデルでの実験条件は表 4 に実験結果は図 5 に示す．総合的にみると検討 2 が良い結果であるが部分的に他検討及びベースラインが勝る部分があるといった結果となった．さらにあらゆるカテゴリで検討モデルがベースラインに対して精度が勝っているということがわかる．また，図 14 に示す学習曲線をみると検討 1 モデルは Accuracy = 0.45 で極端な過学習傾向に陥るのに対し検討 2，3 モデルでは 200epoch 学習時でも過学習に陥らないということがわかる．

表 4: 実験条件

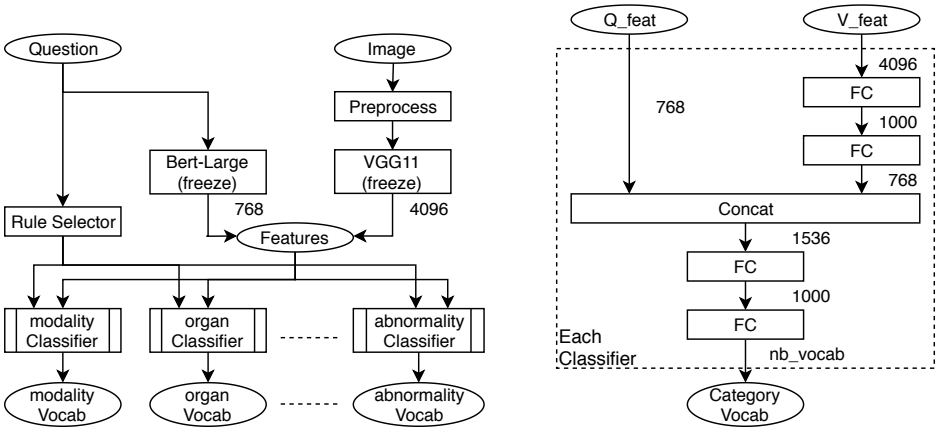
Optimizer	Adam lr $= 1e-3$
Epoch	200
Batchsize	64
Dropout	0.5

8 おわりに

本研究では ImageCLEF VQA-Med タスクである医療画像に基づく質問応答システムについて，より詳細なカテゴリ分類によるアプローチの検討を行った．また，各カテゴリについて適切な特徴量が異なる可能性を示した．また，VGG GAP を用いた学習モデルは VQA-Med のタスクにおいて過学習傾向を抑えられる可能性が高いことを示した．各カテゴリの差異に着目して，有効な画像特徴量を抽出するといった改良が期待される．

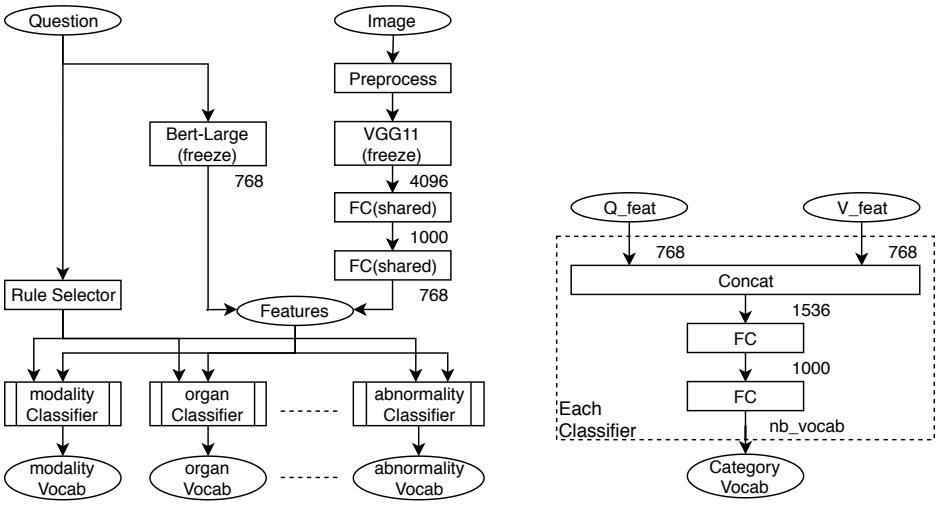
謝　辞

本研究の一部は，科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した．



(a) システム概要図　　(b) ベースライン 分類器モデル

図 **9:** ベースラインモデル概要



(a) システム概要図　　(b) 検討 1 分類器モデル

図 **10:** 検討 1 モデル概要

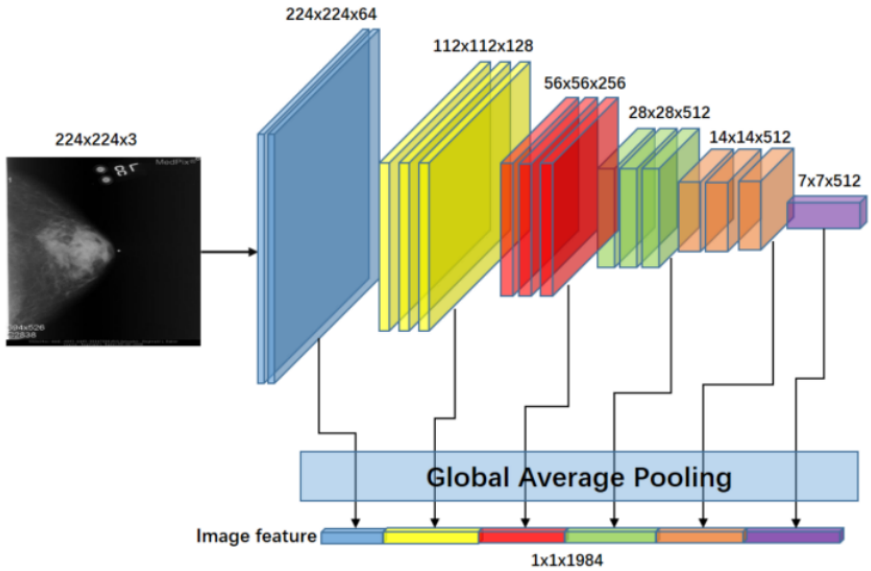
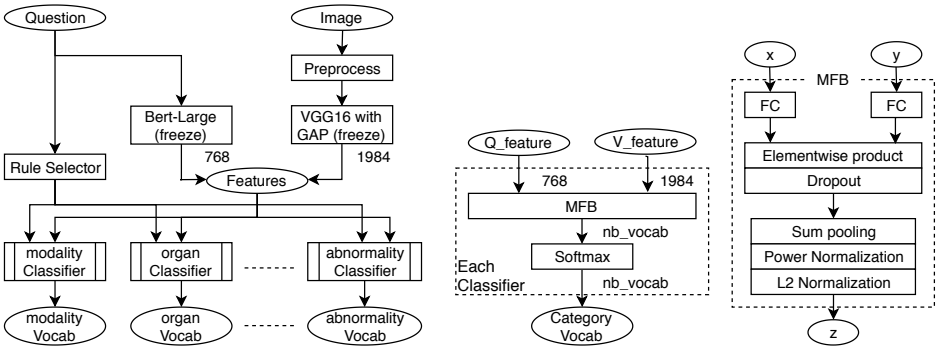
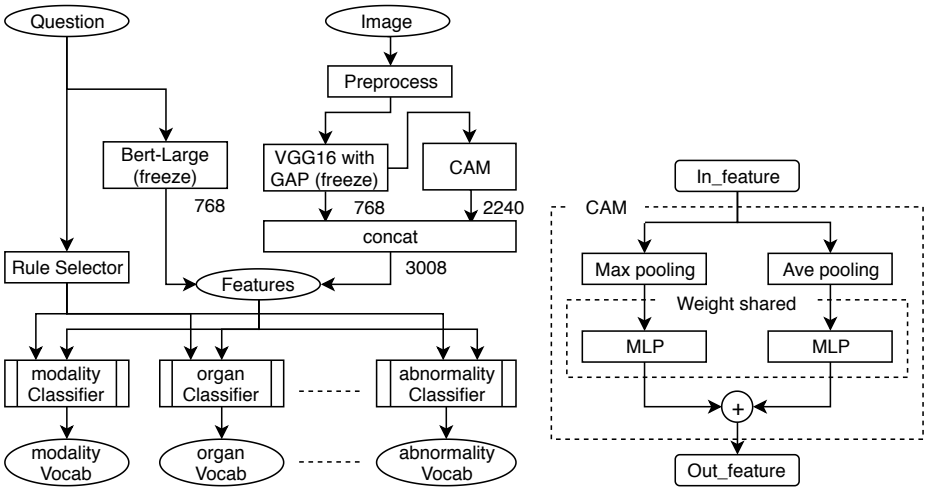


図 **11:** VGG の GAP を用いた特徴量抽出の概要 ([13] より引用)



(a) システム概要図　(b) 検討 2 分類器モデル　(c) MFB概要

図 12: 検討 2 モデル概要

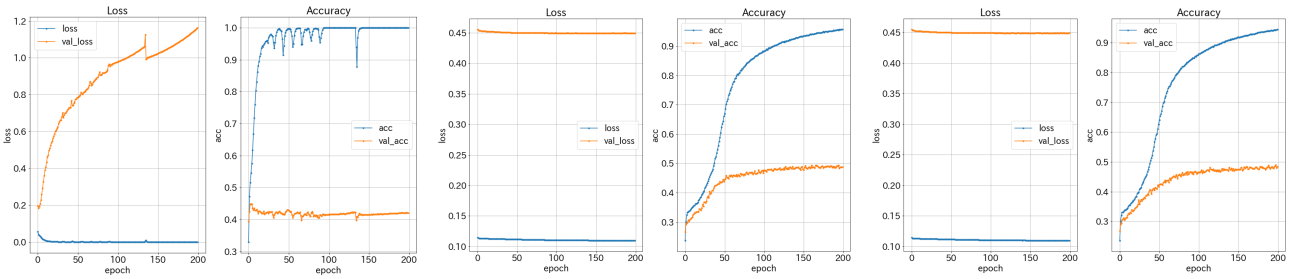


(a) システム概要図　(b) Channel Attention Module概要

図 13: 検討 3 モデル概要

表 5: 実験結果

	moda.	yesno	weig.	cont.	organ	plane	abno.	all
ベース	0.587	0.728	0.741	0.656	0.538	0.620	0.023	0.468
検討 1	0.579	0.767	0.796	**0.688**	0.528	**0.624**	0.015	0.470
検討 2	**0.642**	**0.827**	**0.833**	0.594	0.582	0.596	0.023	**0.491**
検討 3	0.634	0.807	**0.833**	0.563	**0.560**	0.594	**0.025**	0.482



(a) 検討モデル 1　**(b)** 検討モデル 2　**(c)** 検討モデル 3

図 14: 各モデルの学習曲線

文　　献

[1] *Torchvision.models*, https://pytorch.org/docs/master/torchvision/models.html.

[2] Abien Fred Agarap, *Deep learning using rectified linear units (relu)*, 2018, cite arxiv:1803.08375Comment: 7 pages, 11 figures, 9 tables.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, *Bottom-up and top-down attention for image captioning and visual question answering*, CVPR, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, *Vqa: Visual question answering*, International Conference on Computer Vision (ICCV), 2015.

[5] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller, *VQA-Med: Overview of the medical visual question answering task at imageclef 2019*, CLEF2019 Working Notes (Lugano, Switzerland), CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, September 09-12 2019.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database*, CVPR09, 2009.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2018, cite arxiv:1810.04805Comment: 13 pages.

[8] Ross Girshick, *Fast r-cnn*, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (Washington, DC, USA), ICCV '15, IEEE Computer Society, 2015, pp. 1440–1448.

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, *Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering*, Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

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

[11] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, J. Mach. Learn. Res. **15** (2014), no. 1, 1929–1958.

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, *Cbam: Convolutional block attention module*, The European Conference on Computer Vision (ECCV), September 2018.

[13] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu, *Zhejiang university at imageclef 2019 visual question answering in the medical domain*, CLEF, 2019.

[14] Xin Kang Yangyang Zhou and Fuji Ren, *Tua1 at imageclef 2019 vqa-med: A classification and generation model based on transfer learning*, CLEF, 2019.

[15] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, *Multi-modal factorized bilinear pooling with co-attention learning for visual question answering*, (2017).