

DCNN特徴を用いた Webからの質感画像の収集と分析

PRMU研究会

下田 和, 柳井 啓司

電気通信大学 情報理工学部 総合情報学科

背景

- 様々な画像認識

- 食事画像
- 動物画像



- 新しい特徴量

- これまで 局所特徴量 + コーディング
(SIFT, SURF → boww, vlad, fisher)
- 現在 DCNN特徴

(Deep Convolutional Neural Network activation feature)

目的

- 質感語に対応した画像の認識
 - 質感語（特にオノマトペ（擬音語））
 - 一般の物体認識より難しい
 - 新しい特徴量（DCNN特徴）で認識が行えるのか
- 意味的な認識に有効な特徴量
 - オノマトペ画像の多様性
 - 認識を通して、こういった画像に対してどのような特徴が有効であるか



ごつごつ画像

質感語としてオノマトペの利用

- オノマトペとは
 - 擬態語や擬音語
 - しわしわ、ふわふわ、ざらざらなど
 - 質感、音、食感、物体の状態



ふわふわ画像の例

関連研究

- 素材画像の認識
 - Flickr Material Database (FMD)
 - 一般の画像認識より意味的な認識
- 有効な特徴量
 - Improved fisher vector (IFV)
 - Deep Convolutional Neural Network activation feature (DCNN)

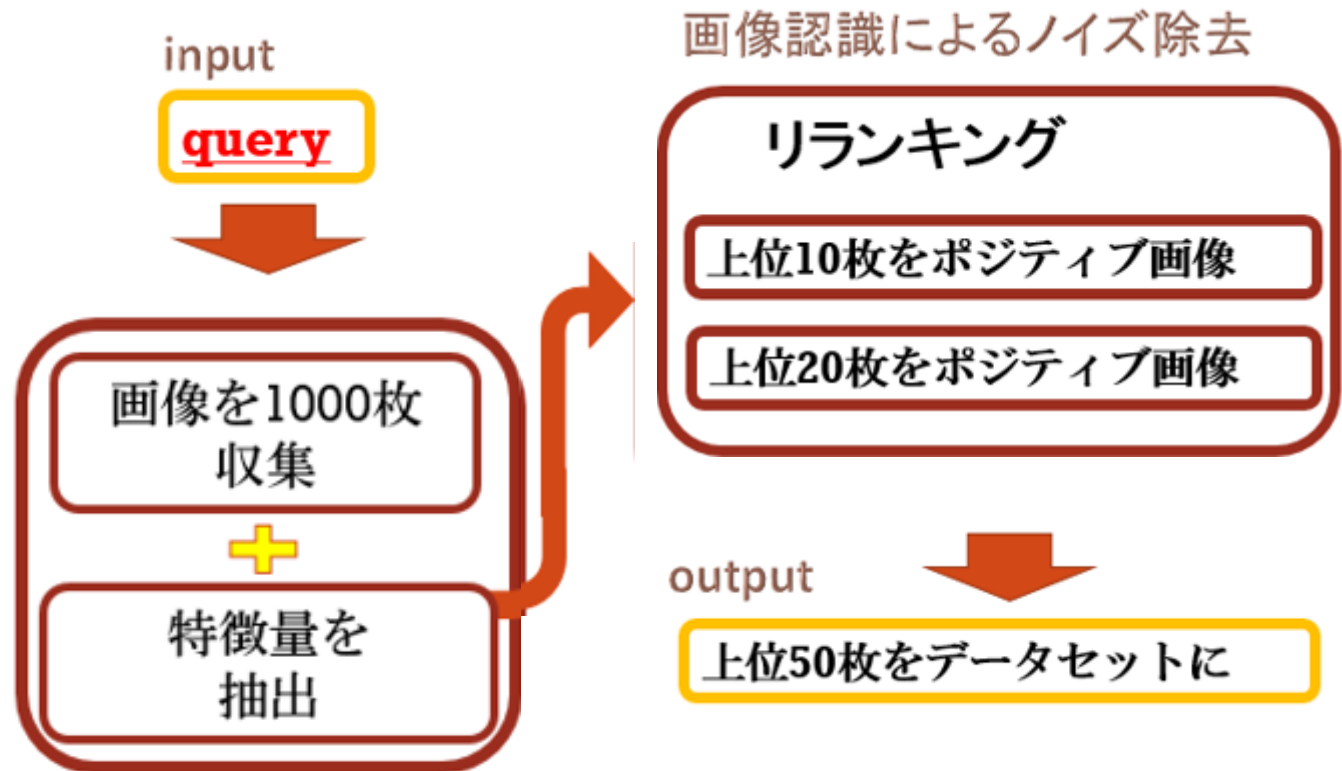


FMD image

実験の手順

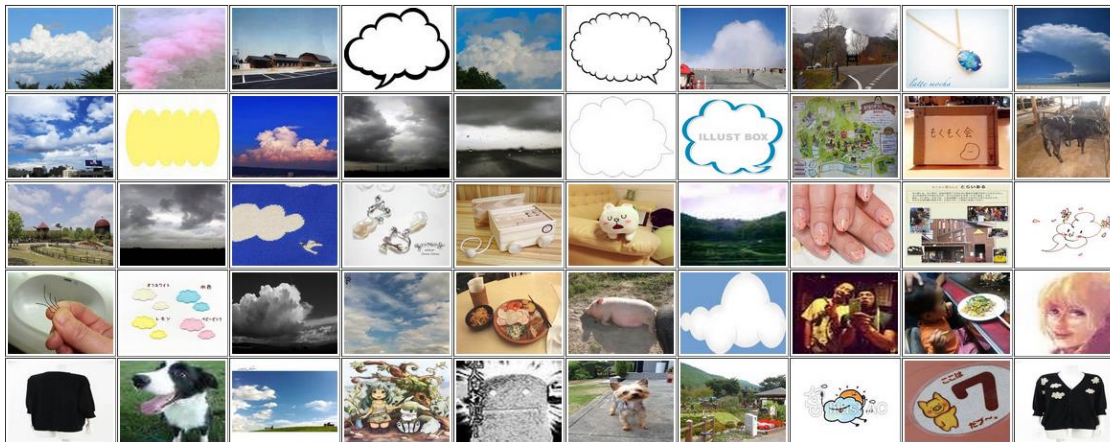
- データセットの構築
 - 代表的なオノマトペ画像データセットはない
 - オノマトペデータセットを作るところから始める
 - 画像認識を使って自動で構築
- データセットの視覚性の評価
 - ランダム画像との分離度で評価

データセットの構築（概要）



オノマトペ画像の収集

- bingAPIを利用
- 一つのクエリにつき1000枚画像を収集

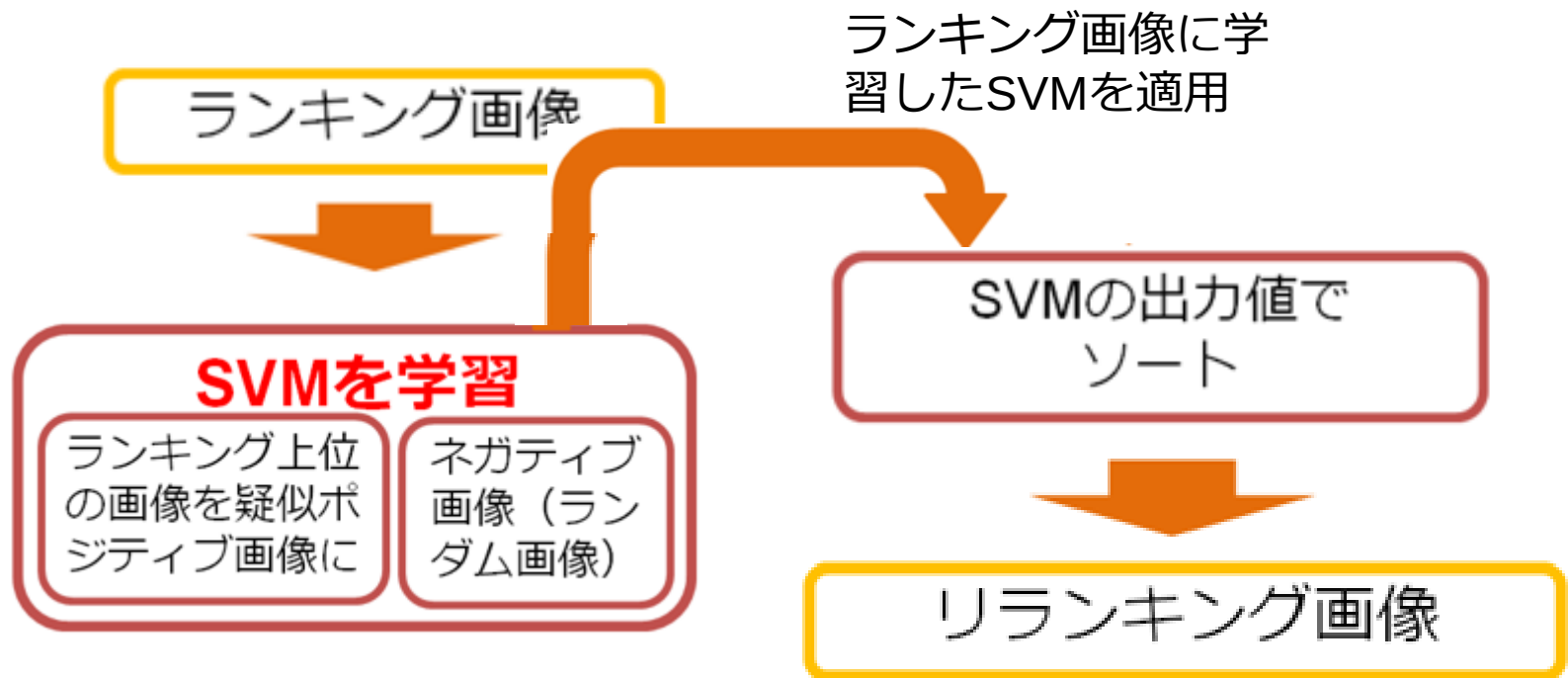


検索結果には上位でもノイズ画像が含まれてしまう
(モクモク検索結果 上位50枚)

リランキング

リランキング手順

- 画像認識によるリランキング



このリランキングを二回繰り返す

リランキングの例

- Queryを入力し、bingAPIでWeb画像を収集「ざらざら」

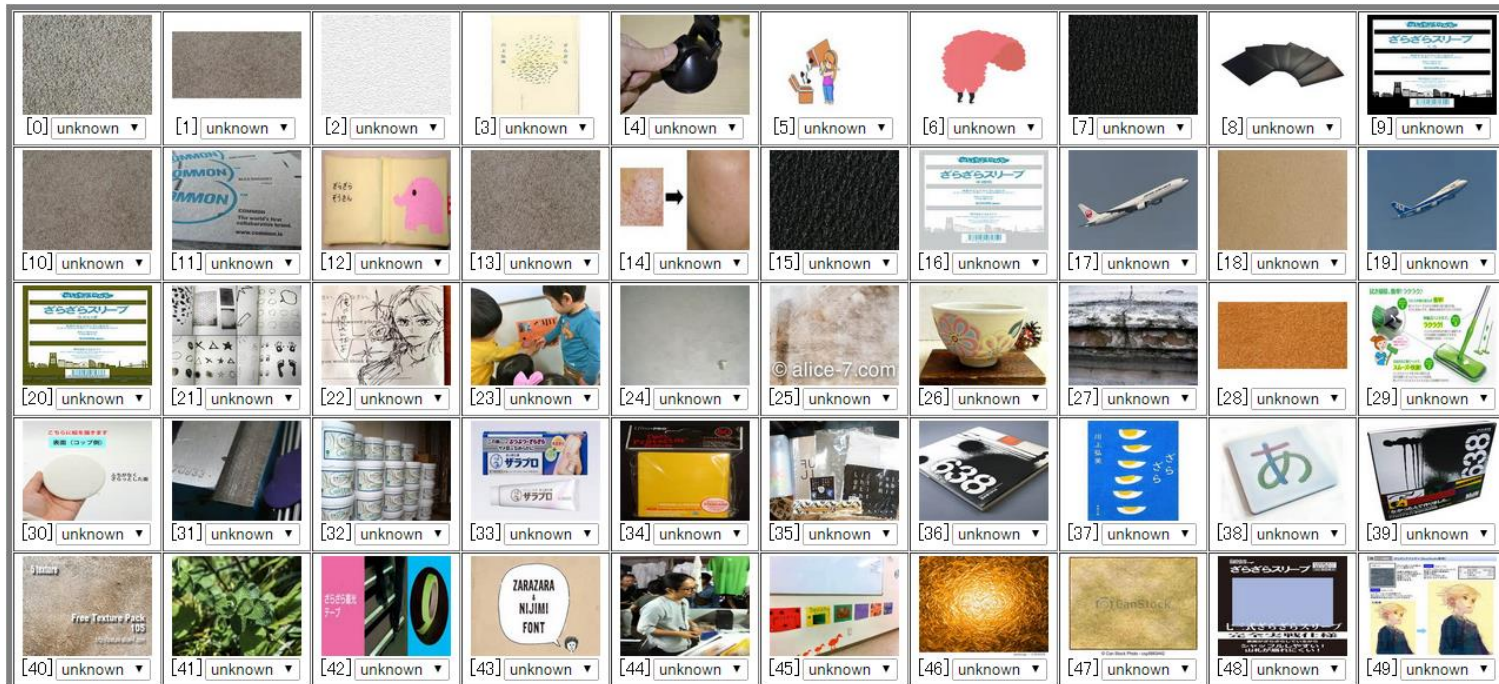


図1 検索結果 上位50まで (ざらざら)

リランキングの例

- 検索結果の上位10枚を使ってリランキング

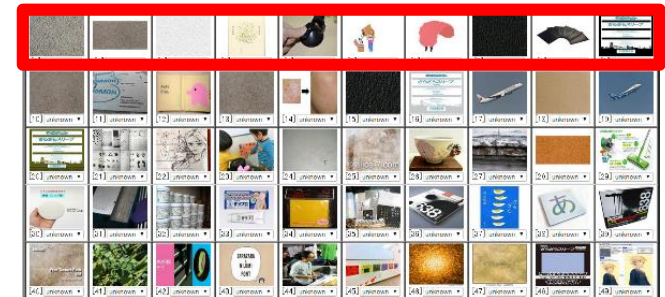


図2 リランキング結果その1（ざらざら）

リランキングの例

- リランキング結果の
上位20枚でリランキング

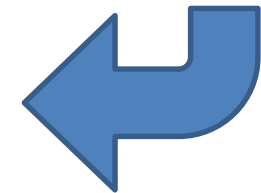


図3 リランキング結果その2 データセット (ざらざら)

認識可能性の評価

- 測定方法

- ランダム画像5000枚を
正例画像50枚に混ぜ、
学習モデルを使って、分離
(5 fold cross varidation)

- 分離度合いで、
認識可能性を評価

分離度大⇒認識可能性が高い

分離度小⇒認識可能性が低い



認識率の評価

- 平均適合率

$$AP(n) = \frac{1}{m} \sum_{k=1}^m Precision_{true}(k)$$

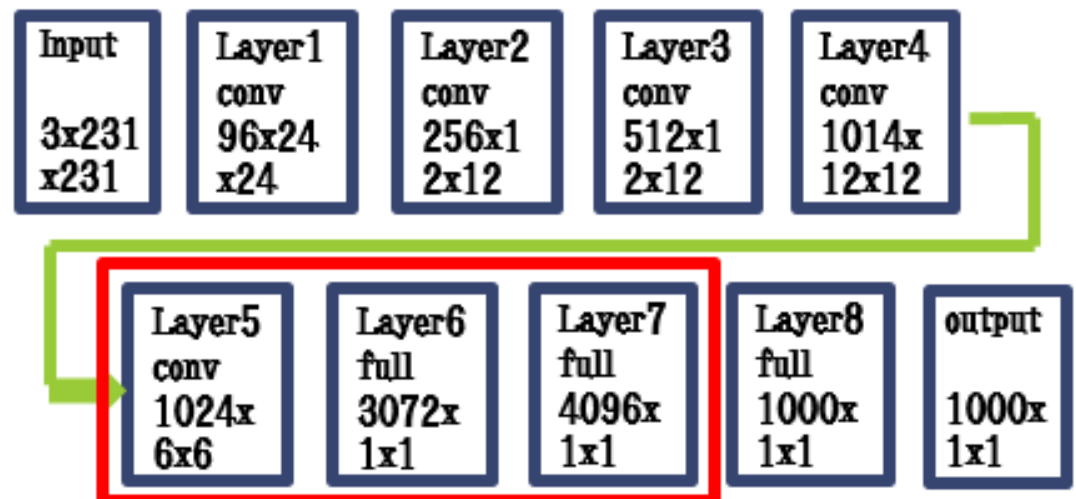
画像認識

- 画像特徴量（2種類を比較）
 - Improved fisher vector (IFV)
 - SURF 128次元、クラスタ数 256
 - Deep Convolutional Neural Network (DCNN)
- 分類器
 - Support vector machine (SVM)
 - 線形SVM

DCNN特徴の抽出

- Overfeat [Pierre Sermanet et al, 2013]
 - ImageNet Challenge の 1000 カテゴリ,
100万枚でpre-training
 - layer5, 6, 7 の出力結果を扱う

Layer5: **conv**
36864 dimension
Layer6: **full**
3072 dimension
Layer7: **full**
4096 dimension

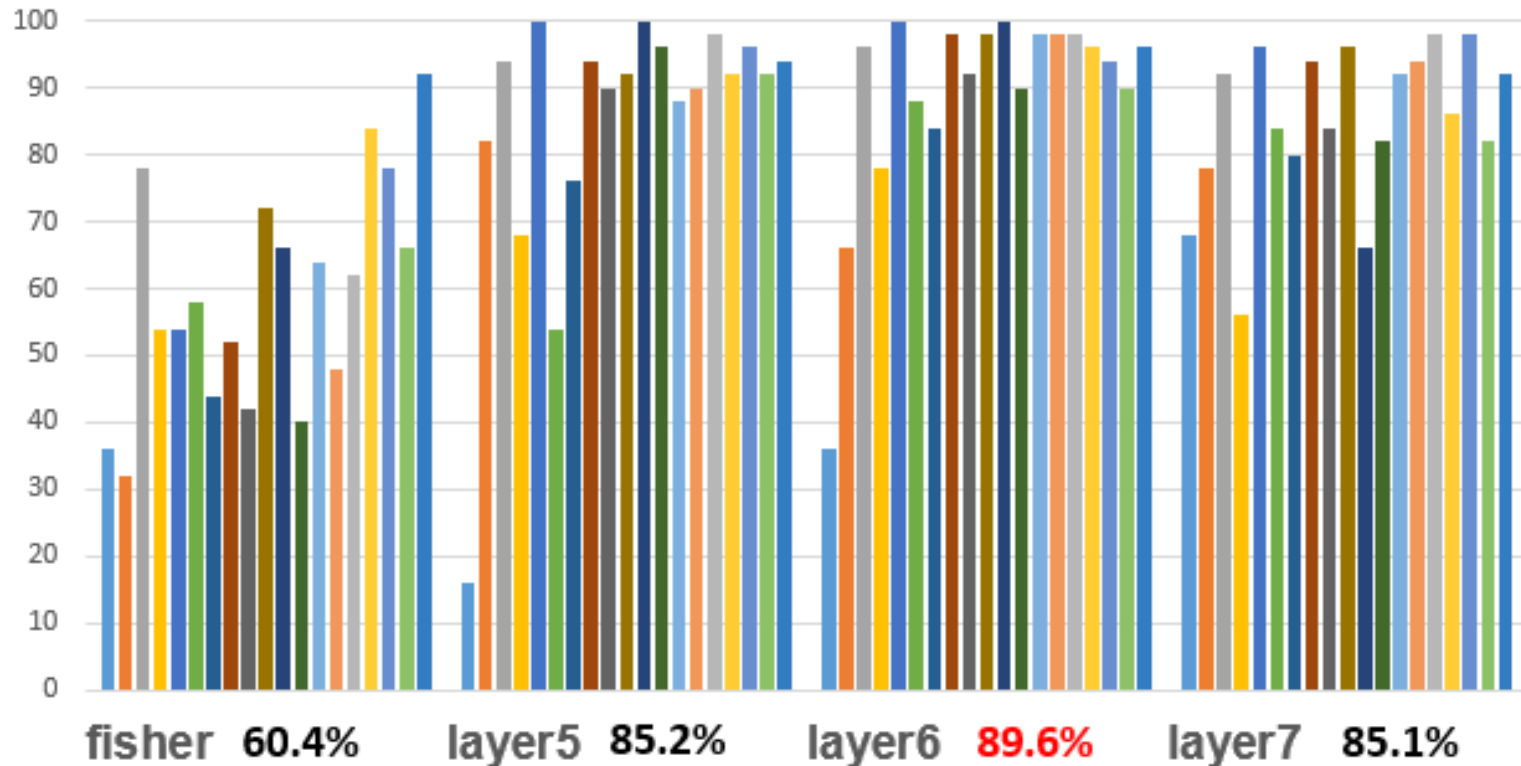


実験

• オノマトペ画像の認識 (20種類)

ぎざぎざ		ごっごっ	
ぐるぐる		プルプル	
かりかり		ブツブツ	
ざらざら		とげとげ	
しわしわ		グネグネ	
じゅわじゅわ		きらきら	
ぼうぼう		ニョキニョキ	
じゃらじゃら		ふわふわ	
モクモク		ばしゃばしゃ	
モフモフ		ぴかぴか	

自動収集されたオノマトペ画像の精度評価

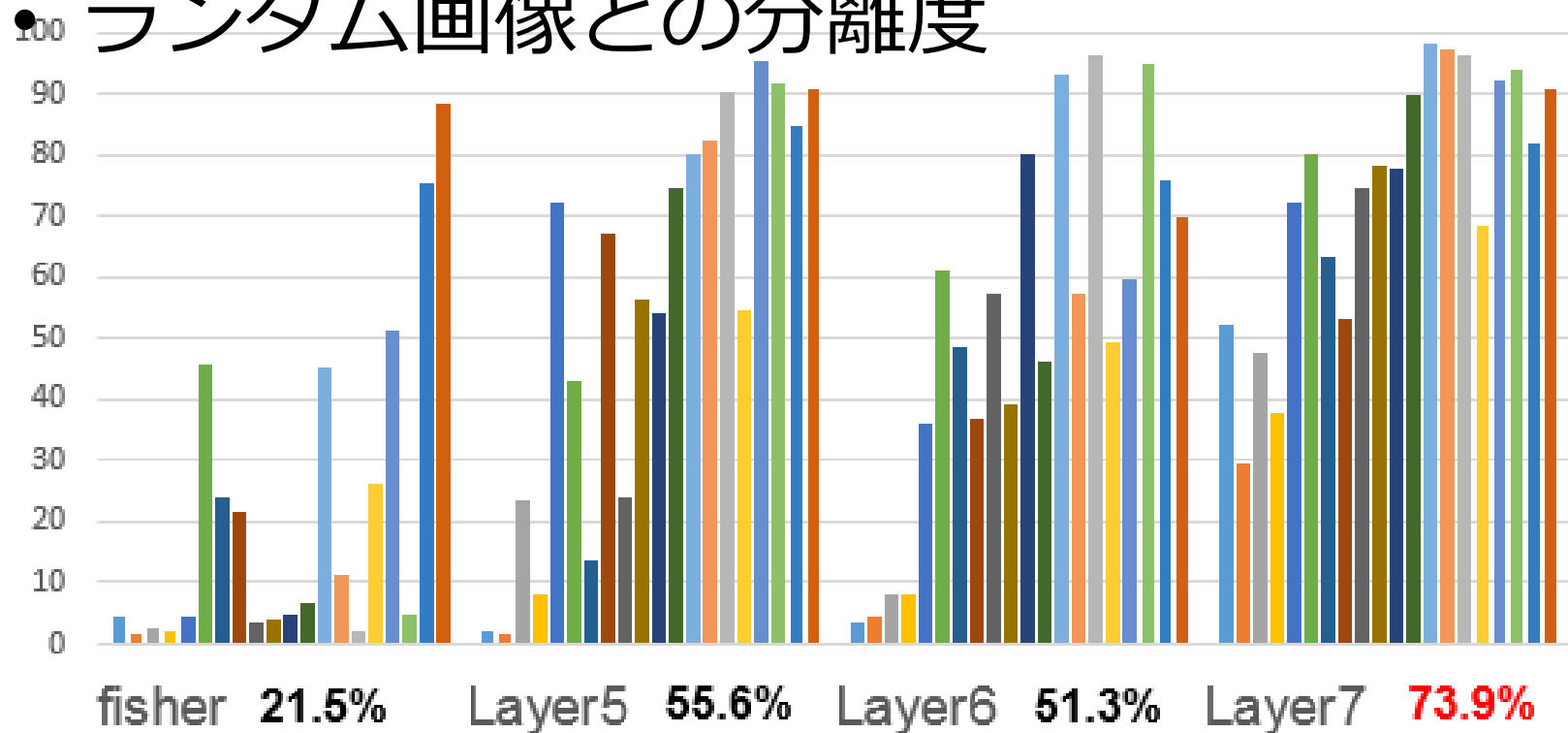


Feature	IFV	Layer5 _(DCNN)	Layer6 _(DCNN)	Layer7 _(DCNN)
平均値(%)	60.4	85.2	89.6	85.1

リランキングにはLayer6が有効

認識可能性の評価（Layer6データセット）

ランダム画像との分離度



Feature	IFV	Layer5(DCNN)	Layer6(DCNN)	Layer7(DCNN)
平均値(%)	21.5	55.6	51.3	73.9

ランダム画像との分離度の可視化

「ふわふわ」 (赤枠はノイズ画像)

Fisher

2.8%



Layer7_(DCNN)

47.4%



「ごっごっ」

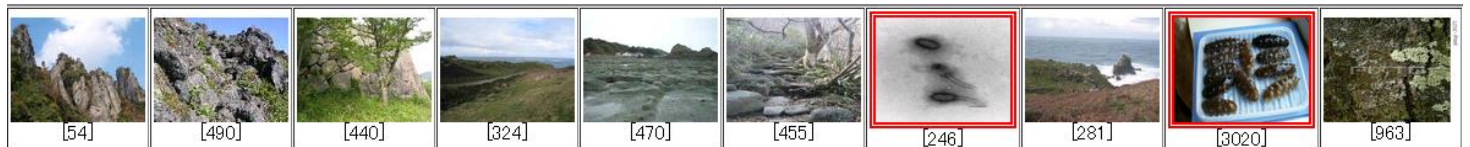
Fisher

3.8%



Layer7_(DCNN)

72.8%



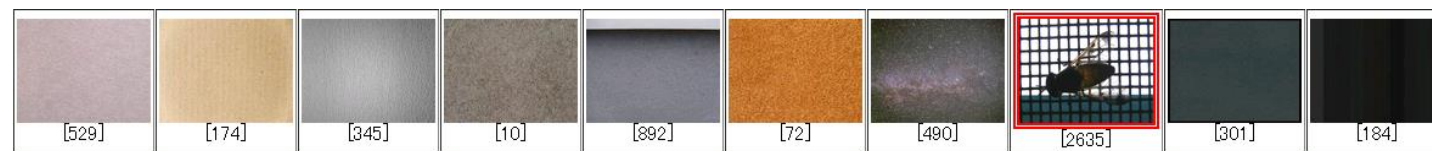
ランダム画像との分離の可視化

「ざらざら」 (赤枠はノイズ画像)

Fisher
51.4%



Layer7_(DCNN)
92.4%



「じゅわじゅわ」

Fisher
2.1%



Layer7_(DCNN)
96.6%



【追加実験】 名詞＋オノマトペ画像

オノマトペ画像のみだと特定の物体に偏る場合がある

- ・ 例) ごつごつ→山の画像が多い
- ランダム画像との分離→山画像の認識？



名詞＋オノマトペ画像の認識

- ・ 名詞内でオノマトペのマルチクラス分類
- ・ オノマトペ以外の要素が認識に関わる可能性を抑える

(例 ふわふわケーキ、パサパサケーキ、サクサクケーキ)

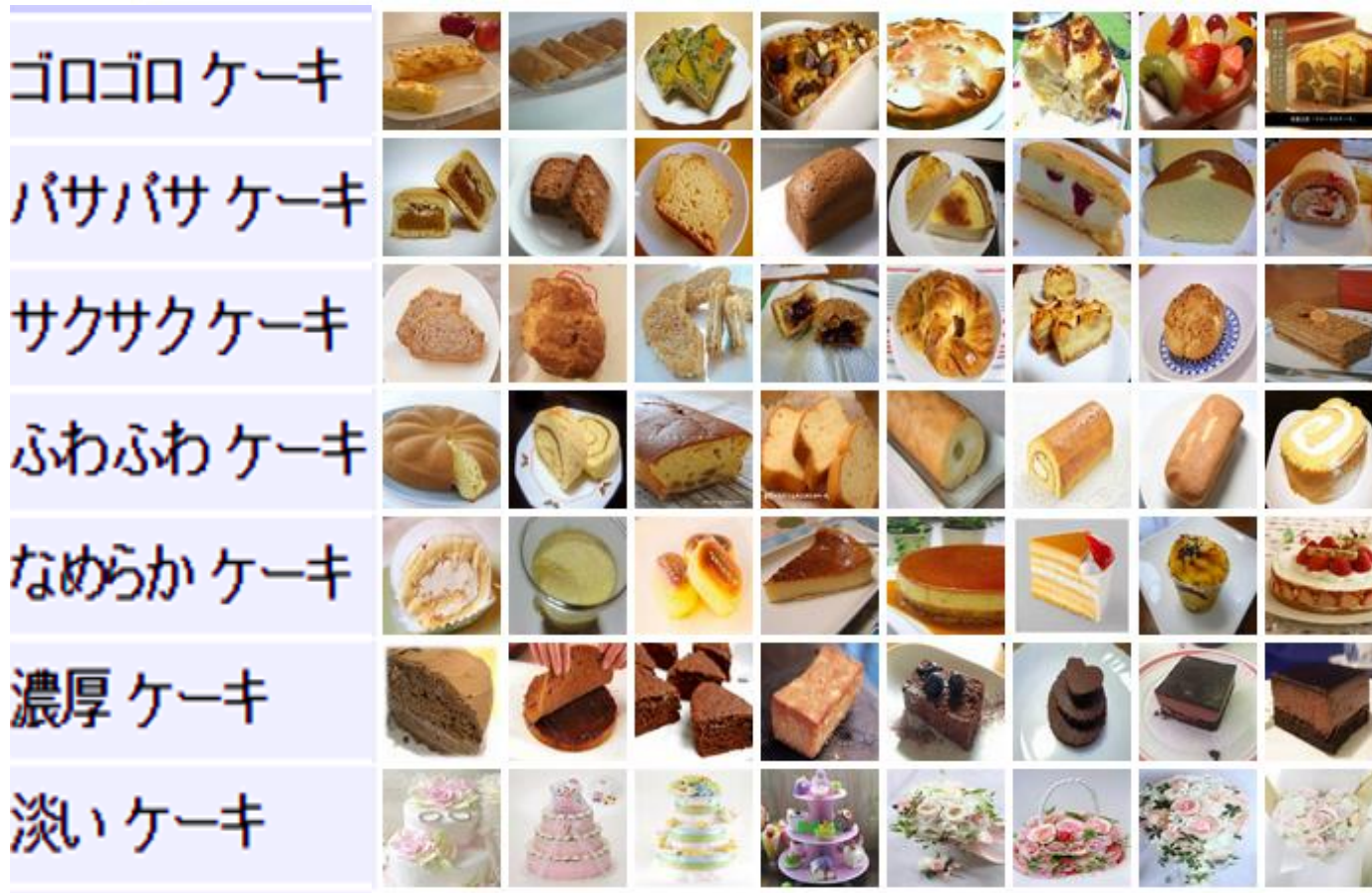
同一名詞内オノマトペマルチクラス分類

- DCNN特徴 5 fold cross varidation で実験
- 名詞と組み合わせたオノマトペは、Webテキスト検索の共起頻度に基づき自動選択
 - クラス数が少なくなるので形容詞も使用

名詞	犬 5クラス	靴 6クラス	ケーキ 7クラス	花 7クラス
分類率(%)	52.5	85.7	72.3	84.6

名詞＋オノマトペ画像の認識

・ ケーキ＋オノマトペ画像



ケーキ内オノマトペマルチクラス分類

		confusion matrix							
ゴロゴロ ケーキ		31	3	10	2	3	1	0	62.0
パサパサ ケーキ		3	26	3	7	5	6	0	52.0
サクサク ケーキ		8	4	24	8	2	3	1	48.0
ふわふわ ケーキ		1	2	2	42	3	0	0	84.0
なめらか ケーキ		3	2	3	3	37	2	0	74.0
濃厚 ケーキ		1	0	2	1	0	46	0	92.0
淡い ケーキ		0	0	0	0	3	0	47	94.0
		66.0	70.3	54.5	66.7	69.8	79.3	97.9	72.3%

名詞＋オノマトペ画像の認識

- 花＋オノマトペ画像



名詞内オノマトペマルチクラス分類

		confusion matrix							
ポンポン 花		35	3	2	2	4	0	4	70.0
ふわふわ 花		7	36	2	3	2	0	0	72.0
フレッシュ 花		2	1	38	7	1	0	1	76.0
メイン 花		0	2	3	44	1	0	0	88.0
ブルー 花		1	0	0	0	49	0	0	98.0
黄色い 花		0	0	0	0	0	50	0	100.0
赤い 花		3	0	0	0	3	0	44	88.0
		72.9	85.7	84.4	78.6	81.7	100.0	89.8	84.6%

まとめ

- オノマトペ画像のWeb画像を自動収集し、認識可能性を評価した
 - 認識可能性は単語による差が大きかった
 - 「ふわふわ」などは認識可能性が低く、「ざらざら」、「じゅわじゅわ」などは認識可能性が高い
- さらに、名詞＋オノマトペに対応する画像の分類可能性も評価した
 - 同一名詞内での分類であってもオノマトペを分類することができた

今後の課題

- 高精度な特徴量の導入
 - DCNN特徴のFine-tuningを行う
- 質感語に対応する領域の推定
- 単一オノマトペ＋様々な名詞の組み合わせに対応した画像の収集と分析
 - 名詞が異なってもオノマトペが認識できるかどうか（例 ふわふわケーキ VS ふわふわ犬）



ご清聴ありがとうございました