枝刈り・量子化・符号化を用いたカプセルネットワークの軽量化

尾崎ゼミ 4年
島田研太

サマリー

Capsule Networkの圧縮

ベクトルに着目した圧縮の適応

Dynamic routing between capsules.
Sabour, S., Frosst, N., and Hinton, G. E.
NIPS 2017

目次

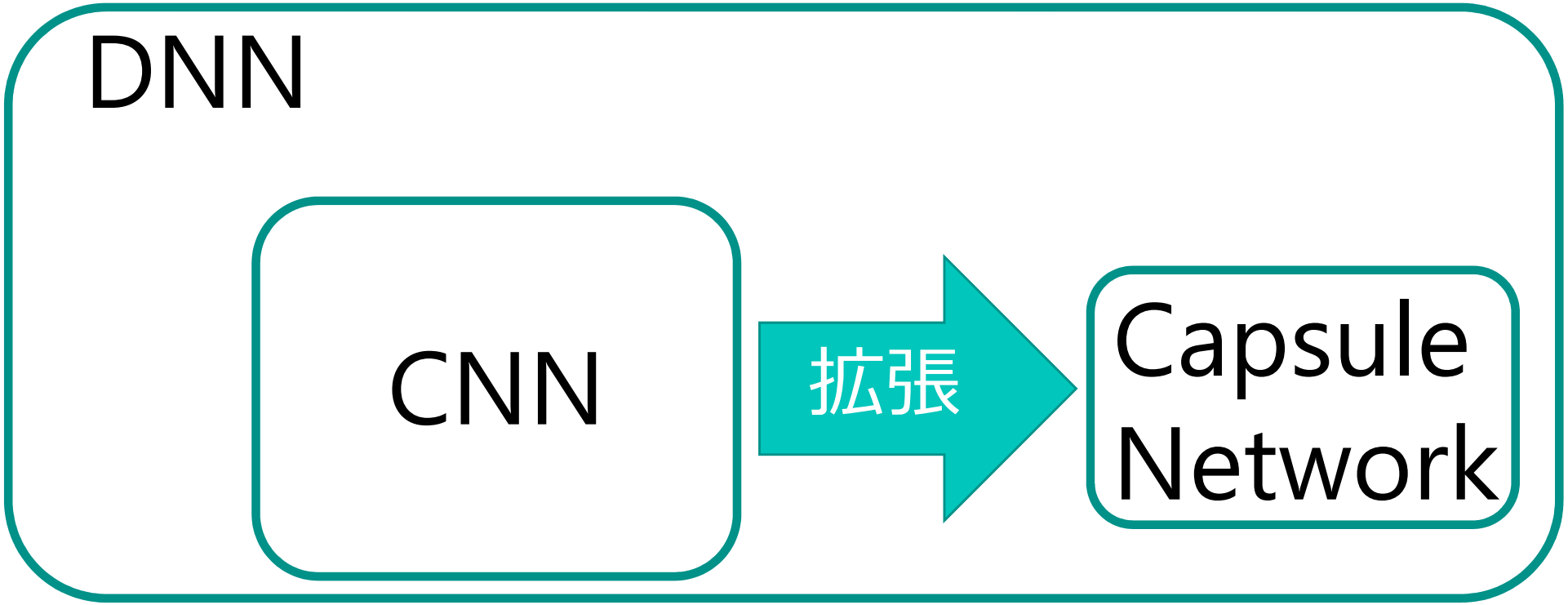
1. DNN,CNN,Capsule Network
2. モデル圧縮手法
3. Capsule Networkへの適応にあたって
4. 評価実験および考察

目次

1. DNN,CNN,Capsule Network
2. モデル圧縮手法
3. Capsule Networkへの適応にあたって
4. 評価実験および考察

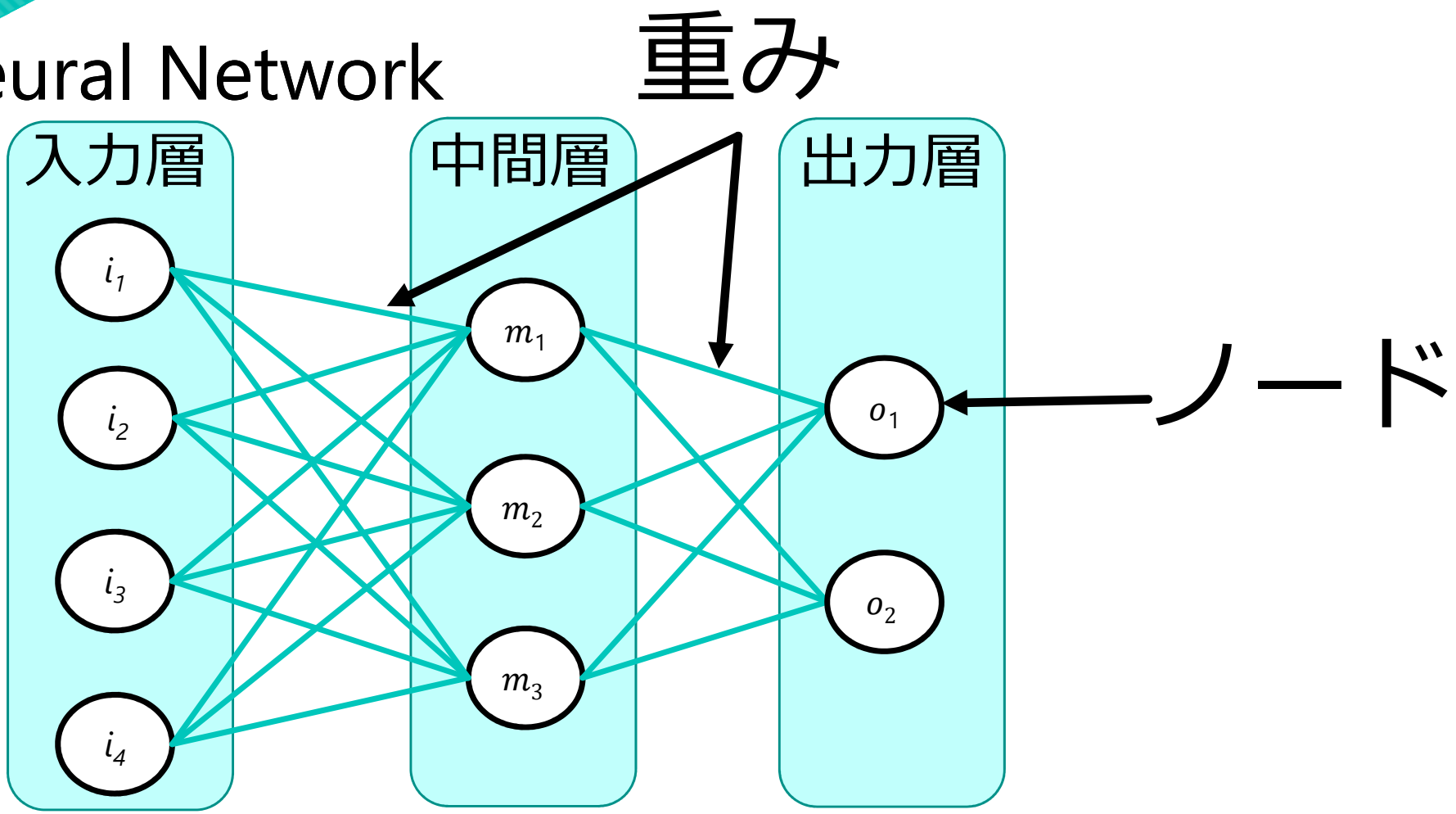
深層学習について

DNN,CNN,Capsule Networkの関連性

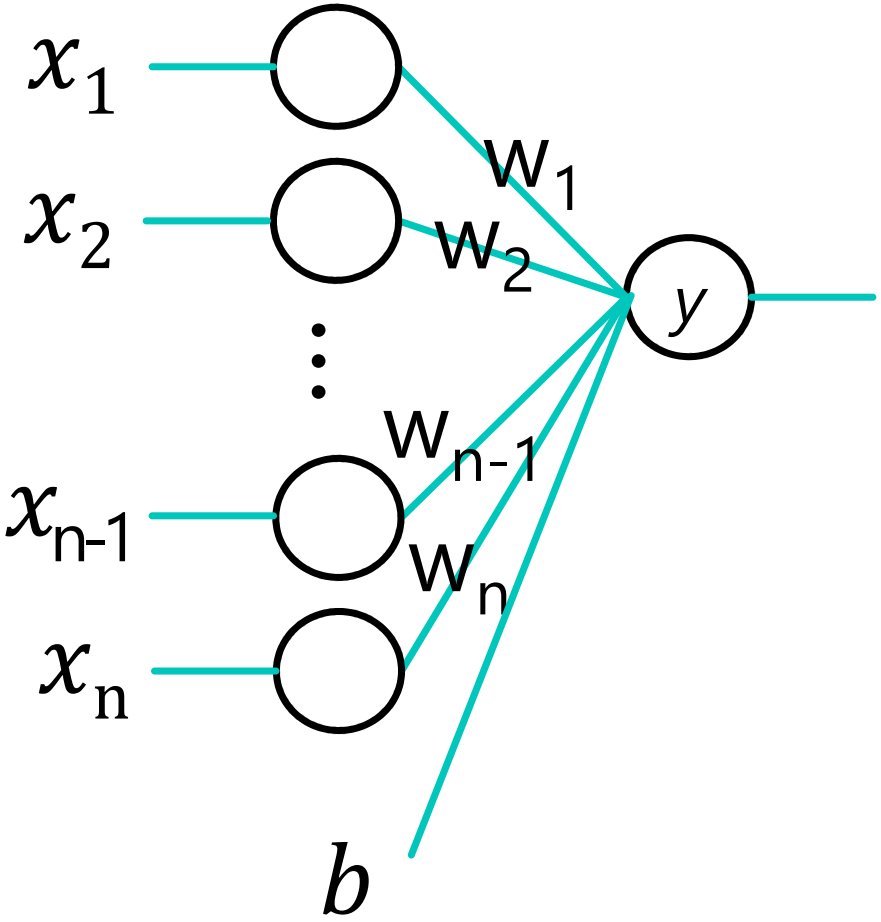


DNN⊇CNN⊇Capsule Network

Deep Neural Network



DNN⊇CNN⊇Capsule Network



yの出力（スカラー）

$$=f(\Sigma w_n \chi_n + b)$$

χ_n :入力値（スカラー）

w_n :重み（スカラー）

b :バイアス（スカラー）

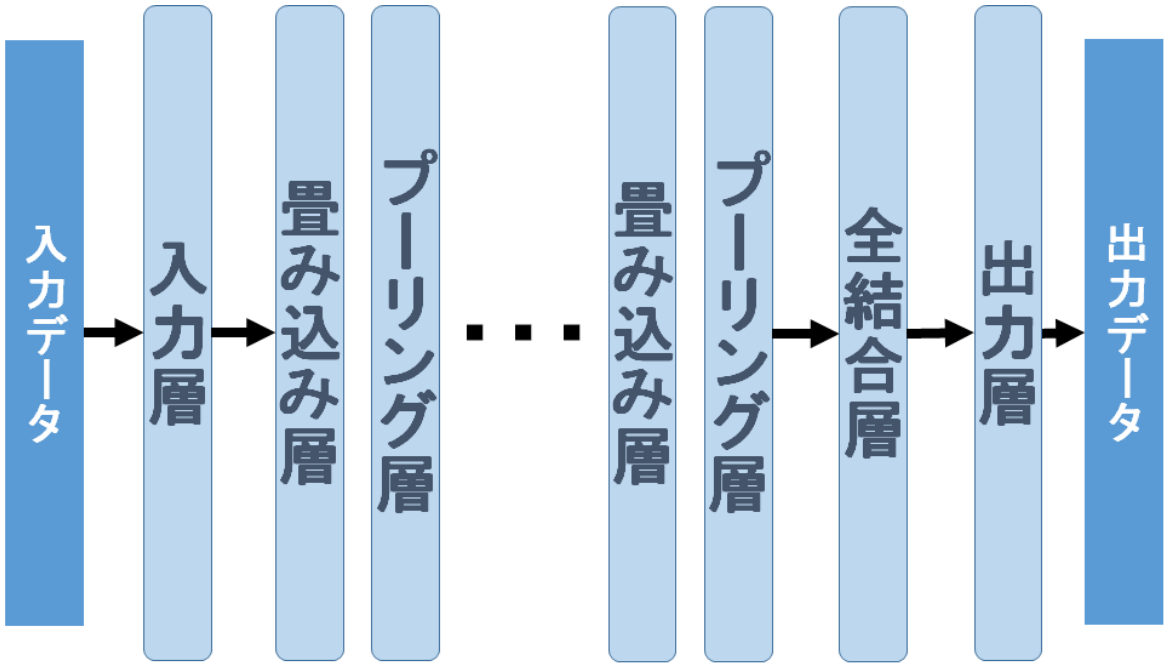
$f()$:活性化関数

DNN⊇CNN⊇Capsule Network

Convolutional Neural Network

DNNに畳み込み層とプーリング層を追加

高精度な画像認識



https://thinkit.co.jp/sites/default/files/639906.png

DNN⊇CNN⊇Capsule Network

畳み込み層

入力データ

0	2	0	1	2	0
1	0	3	0	0	2
3	1	0	2	2	1
2	2	2	1	0	1
1	0	1	3	2	2
0	3	1	0	2	0

フィルタ
(重み)

1	0	0
0	1	0
0	0	1

0*1 + 2*0 + 0*0 +
1*0 + 0*1 + 3*0 +
3*0 + 1*0 + 0*1 = 0

特徴マップ

0			

https://axa.biopapyrus.jp/deep-learning/cnn.html

DNN⊇CNN⊇Capsule Network

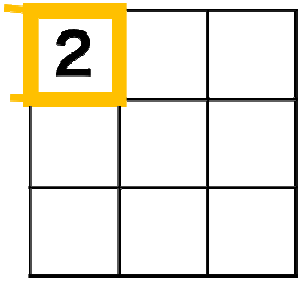
プーリング層

入力データ

0	2	0	1	2	0
1	0	3	0	0	2
3	1	0	2	2	1
2	2	2	1	0	1
1	0	1	3	2	2
0	3	1	0	2	0

https://axa.biopapyrus.jp/deep-learning/cnn.html

最大値を出力



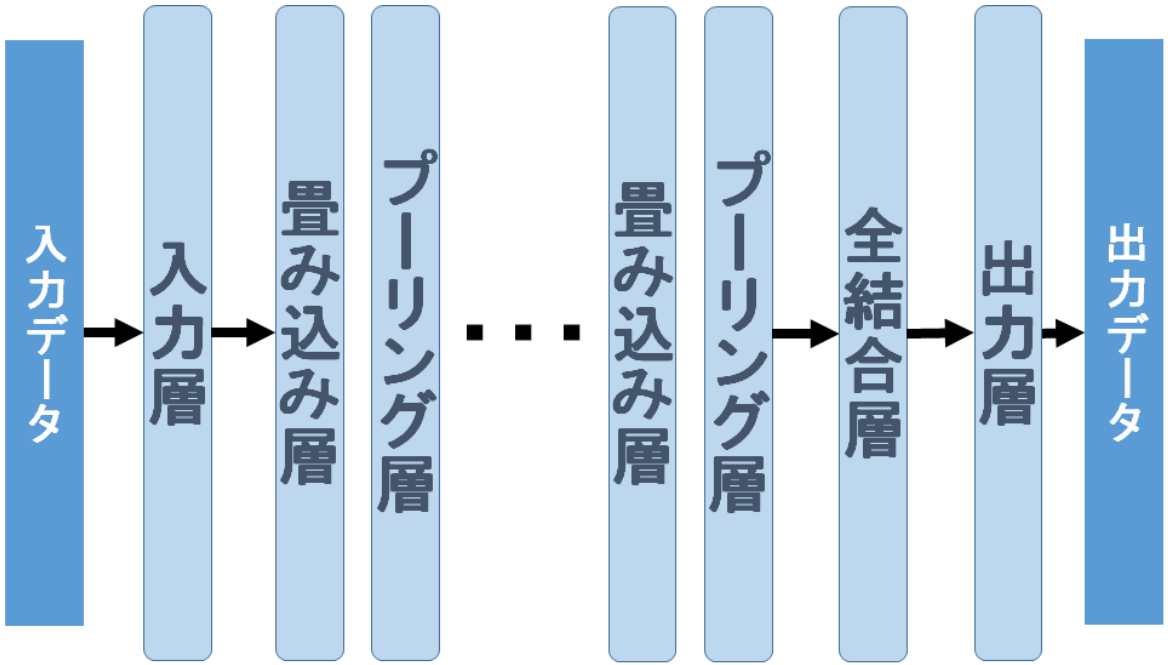
微小な位置の変化や平行移動
に対する頑健性を獲得

DNN⊇CNN⊇Capsule Network

Convolutional Neural Network

DNNに畳み込み層とプーリング層を追加

画像認識に特化

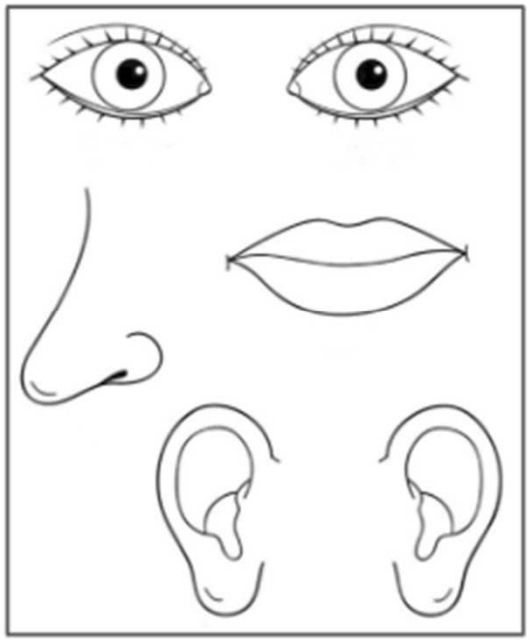


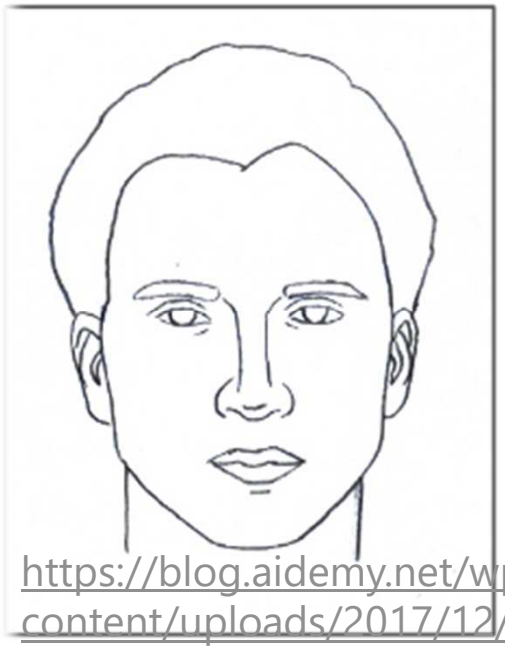
https://thinkit.co.jp/sites/default/files/639906.png

DNN⊇CNN⊇Capsule Network

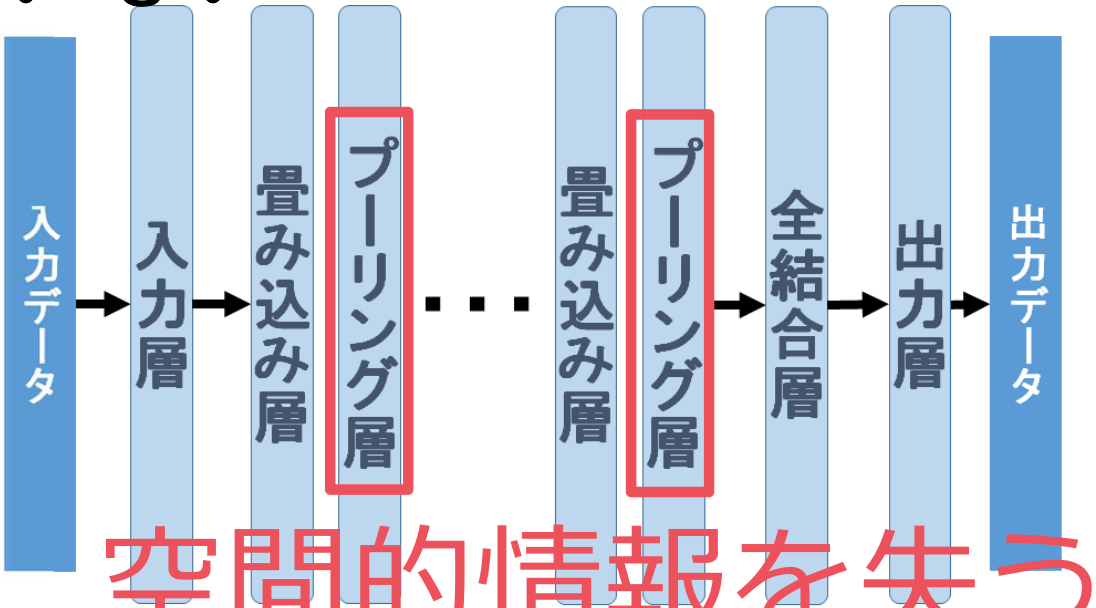
CNNの欠点

空間的情報を考慮していない





https://blog.aidemy.net/wp-content/uploads/2017/12/20171202201831.png

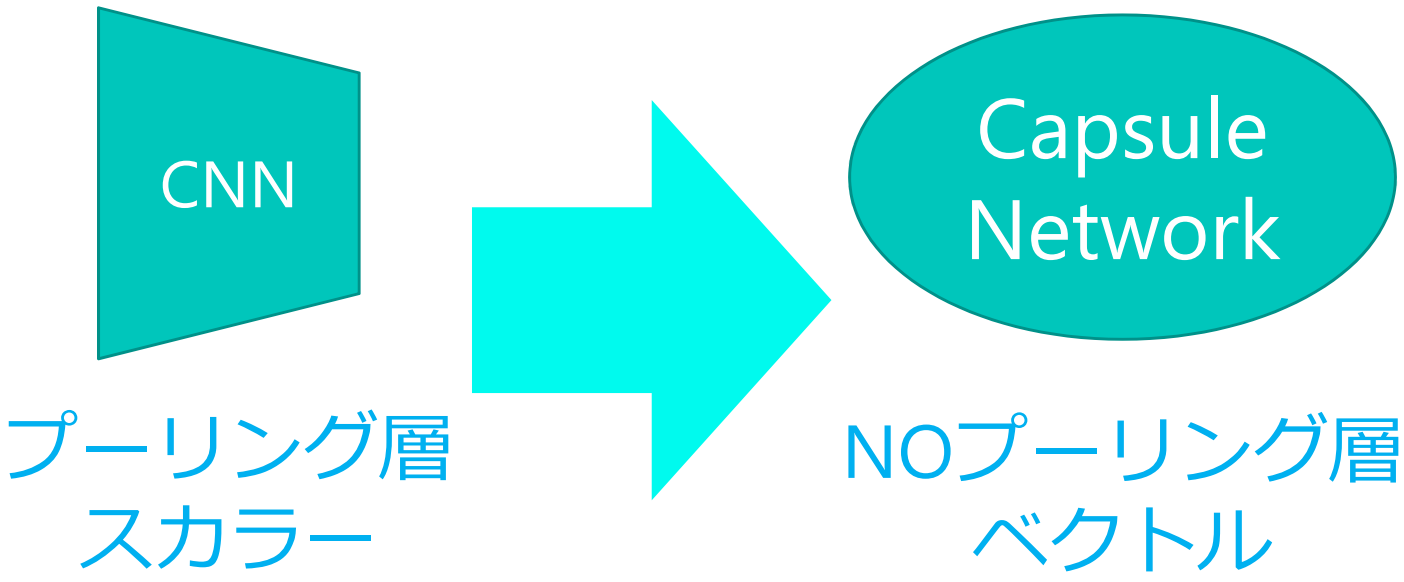


空間的情報を失う

https://thinkit.co.jp/sitesdefault/files/639906.png

DNN⊇CNN⊇Capsule Network

特徴間の存在確立を計算するためにベクトルを使用



特徴間の関連性を保持

DNN⊇CNN⊇Capsule Network

CNN

yの出力（スカラー）

$$=f(\Sigma w_n \chi_n + b)$$

χ_n :入力値（スカラー）

w_n :重み（スカラー）

b :バイアス（スカラー）

$f()$:活性化関数

Capsule Network

yの出力(ベクトル)

$$=f(\Sigma w_n \chi_n + b)$$

χ_n :入力値(ベクトル)

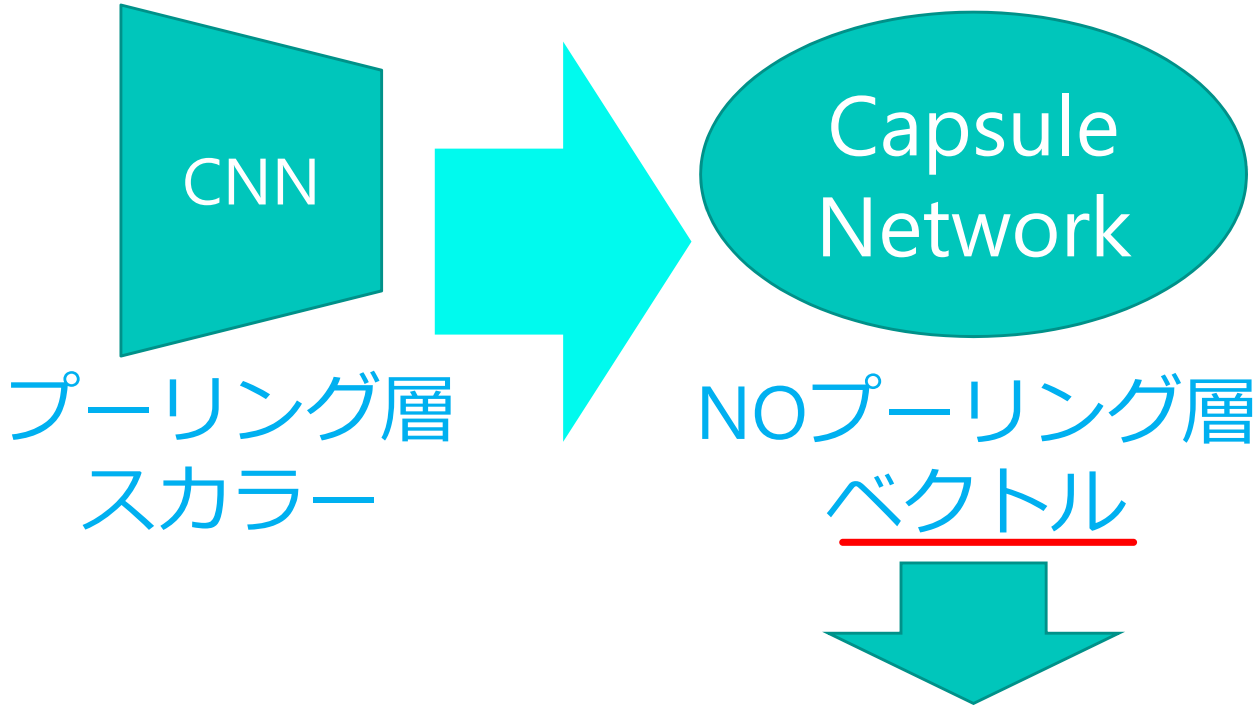
w_n :重み行列(実数行列)

b :バイアス(ベクトル)

$f()$:活性化関数

DNN⊇CNN⊇Capsule Network

軽量化の必要性



CNN	パラメータ数
LeNet	7M
AlexNet	80M
VGG-16	140M
ResNet	58M

目次

1. DNN,CNN,Capsule Network
2. モデル圧縮手法
3. Capsule Networkへの適応にあたって
4. 評価実験および考察

軽量化の関連研究

Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding

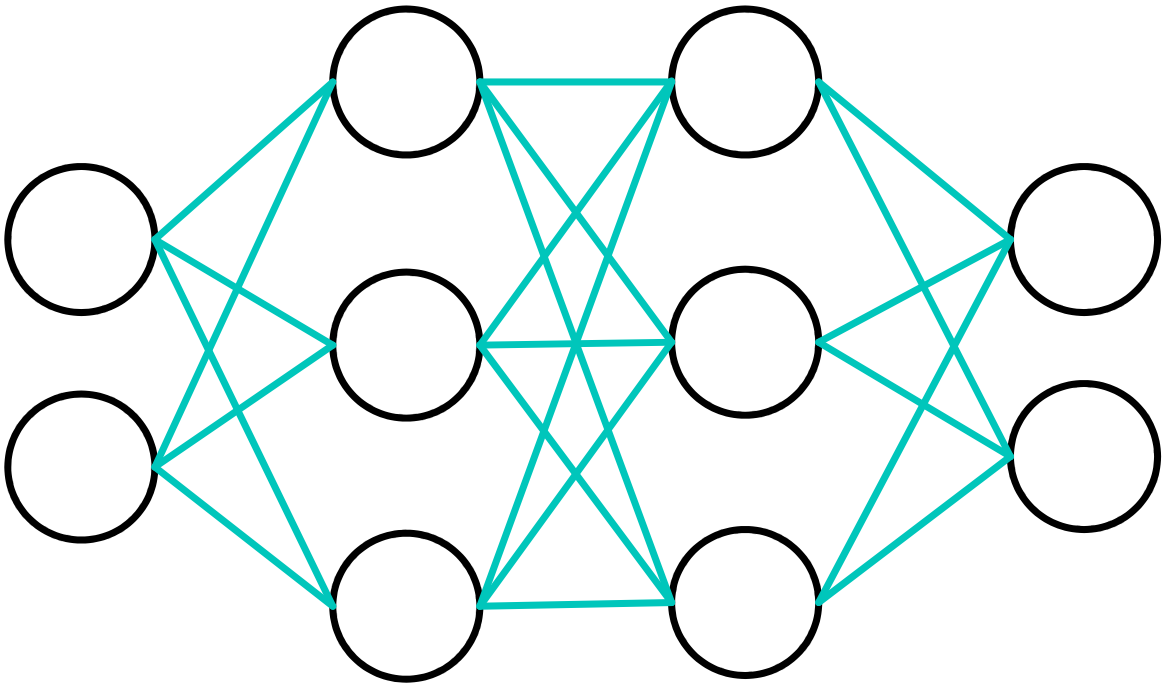
S. Han, H. Mao, and W. J. Dally,
International Conference on Learning Representations 2016

モデル圧縮に関する論文

枝刈り、量子化、符号化を適応

Deep Compression　枝刈り

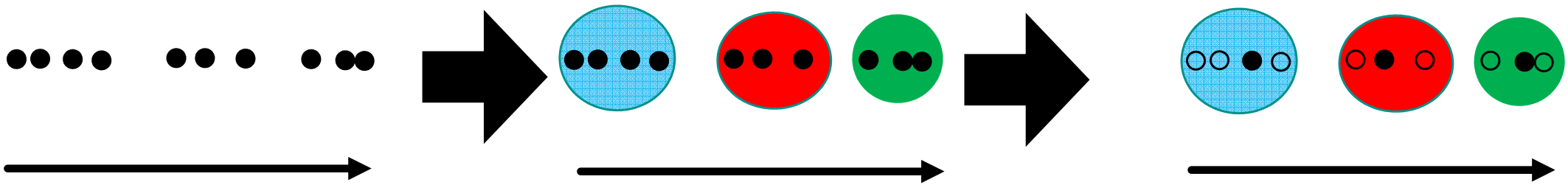
小さい値の重みを０に
することでモデル全体
のデータ量を減らす
枝刈り後に再訓練をする
ことで精度を保つ



Deep Compression　量子化と重み共有

重みをクラスタリング

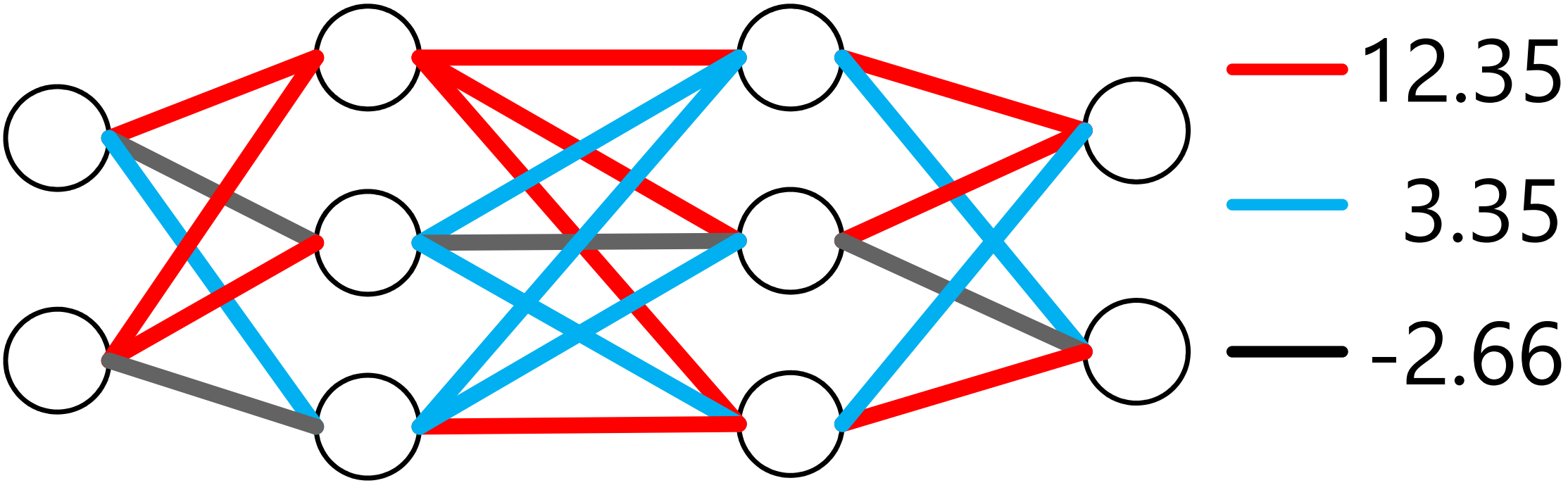
各クラスタの重みをセントロイドに置き換える



グループ化　　グループ内の重みを1つに

Deep Compression　ハフマン符号化

登場頻度の高い重みから順番に短い符号を割り当てることで情報圧縮を行う



Deep Compression　ハフマン符号化

登場頻度の高い重みから順番に短い符号を割り当てることで情報圧縮を行う

12.35	10回	0
3.35	7回	10
-2.66	4回	100

12.35　3.35　12.35
-2.66　12.35　3.35
3.35　12.35　12.35

0 10　0
100　0 10
10　0　0

目次

1. DNN,CNN,Capsule Network
2. モデル圧縮手法
3. Capsule Networkへの適応にあたって
4. 評価実験および考察

Capsule Networkへの適応

スカラー単位をベクトル単位へ

高い圧縮率が期待できる（と予想できる）

DNN CNN Capsule Network

CNN

yの出力（スカラー）

$$=f(\Sigma w_n \chi_n + b)$$

χ_n :入力値（スカラー）

w_n :重み（スカラー）

b :バイアス（スカラー）

$f()$:活性化関数

Capsule Network

yの出力(ベクトル)

$$=f(\Sigma w_n \chi_n + b)$$

χ_n :入力値(ベクトル)

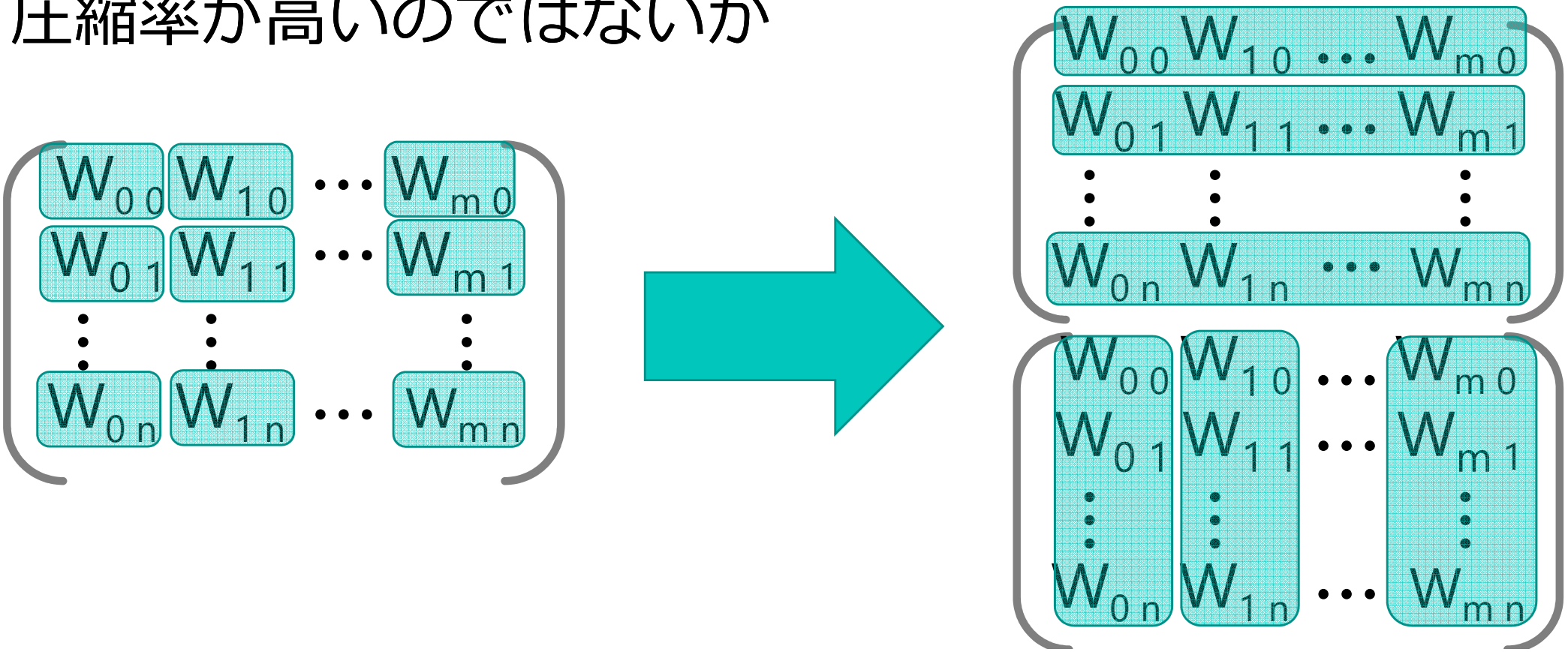
w_n :重み行列(実数行列)

b :バイアス(ベクトル)

$f()$:活性化関数

Capsule Networkへの適応

スカラー単位の圧縮よりベクトル単位での圧縮の方が圧縮率が高いのではないか



目次

1. DNN,CNN,Capsule Network
2. モデル圧縮手法
3. Capsule Networkへの適応にあたって
4. 評価実験および考察

評価実験

データセット : MNIST ,CIFAR-10

モデル形式 : MNIST　Dynamic Routing Between Capsulesのモデル形式

CIFAR-10　VGG16の全結合層をCapsule Networkの層へ変更

データセット

MNIST

白黒の手書き数字

訓練データ6万枚+テストデータ1万枚

＝7万枚

CIFAR-10

10種類色付き画像

訓練データ5万枚+テストデータ1万枚

＝6万枚

結果 分類精度

MNIST
軽量化前99.62%

average		クラスタ数		
		32	64	128
次元数	1	99.53	99.54	99.53
	8	89.45	97.37	98.78
	16	78.81	90.31	96.13

CIFAR-10
軽量化前64.36%

average		クラスタ数		
		32	64	128
次元数	1	64.33	64.41	64.38
	8	55.57	53.34	56.61
	16	44.71	49.27	53.88

結果　モデル容量の変化

各圧縮後のモデル容量

MNIST	容量(MB)
圧縮前	50.78
枝刈り	44.63
量子化	40.81
符号化	15.47

CIFAE-10	容量(MB)
圧縮前	99.89
枝刈り	99.92
量子化	99.92
符号化	88.53

まとめと今後の課題

まとめ
- ベクトルに着目したCapsuleNetworkの圧縮
  - データセット：MNIST、CIFAR-10
- スカラー単位よりベクトル単位のほうが高精度
- せいどてきには上手くいかず

今後の課題
クラスタリングの異なる手法や他の圧縮手法の適応