食事画像に対する
Few/Zero-shot Segmentation

電気通信大学大学院　情報学専攻
本部 勇真, 柳井 啓司

研究背景・目的

従来の問題点

- セグメンテーションモデルの学習には、大量の学習データを必要
- 無数のカテゴリが存在する食事データには、データ不足の問題が生じる

数枚もしくは0枚の学習データで学習できる高精度なモデルの作成

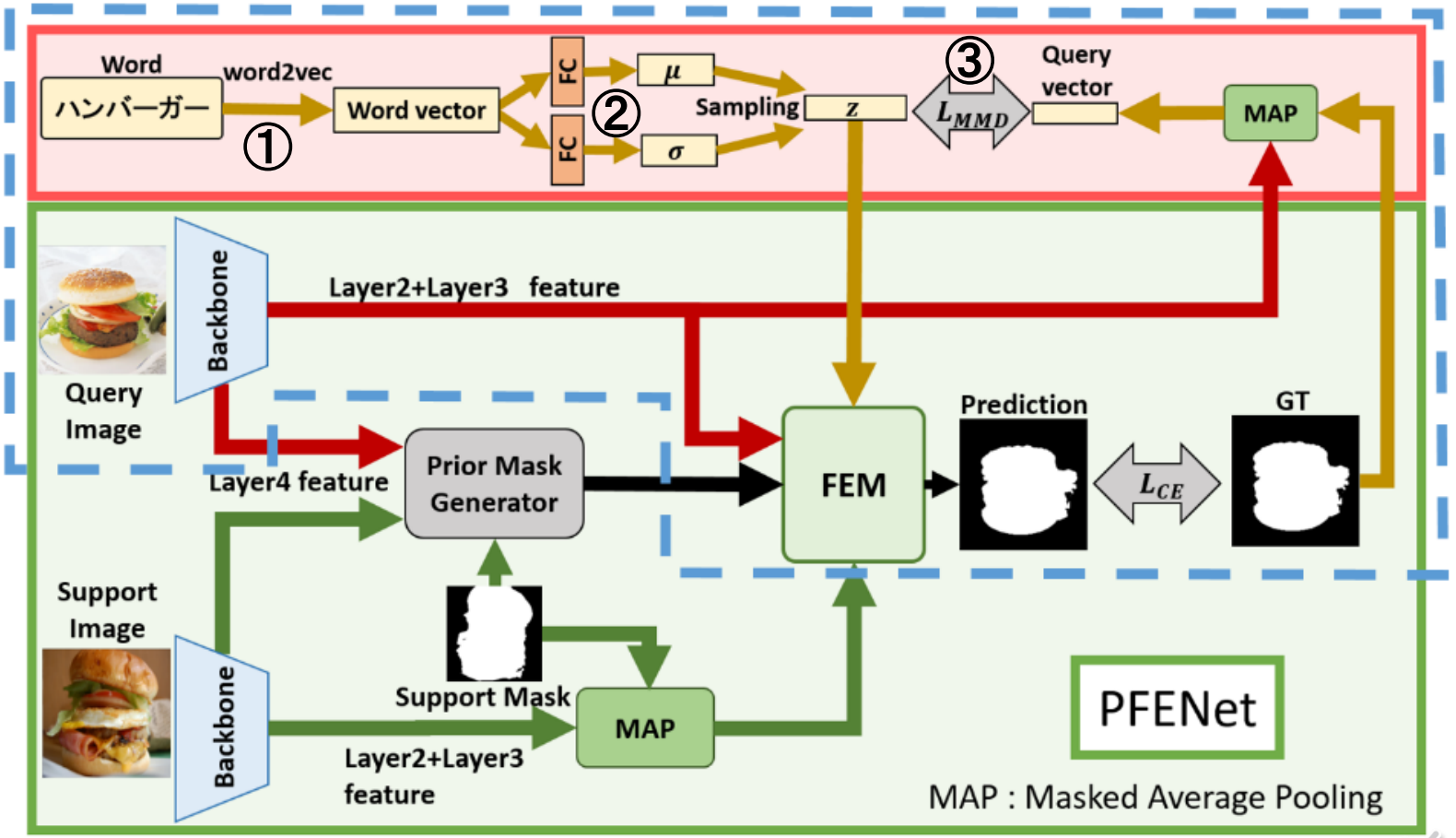
食事データセットの量の不十分さを解消

提案手法

① レシピデータで学習した word2vecの使用

② 単語再構成にVAE

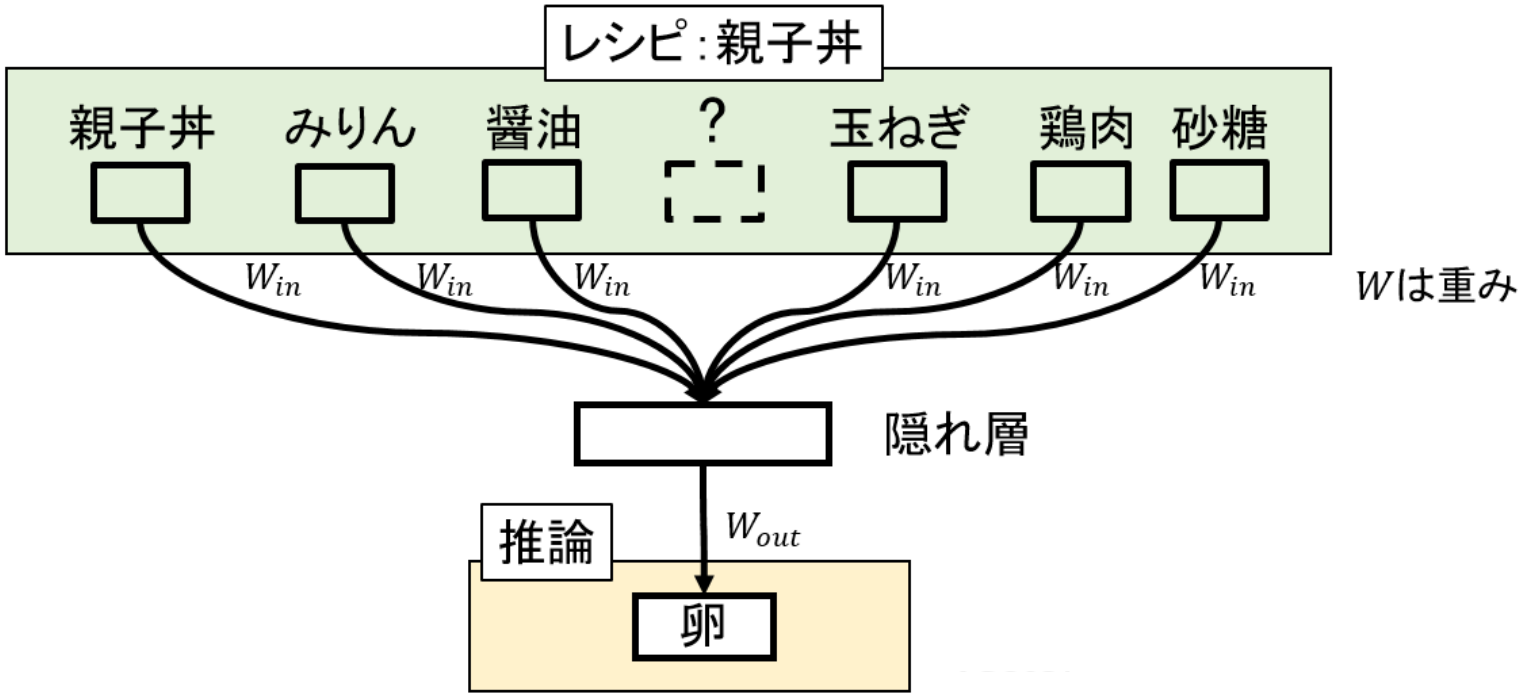
③ MMDを用いたロス



提案モデル (wPFE, zPFE: ----)

提案手法：word2vec

- レシピの素材単語の共起関係に基づく単語埋め込みの追加
  - クックパッドデータセットに含まれる16万のレシピテキストを利用して料理単語のword2vecを学習
  - 埋め込み単語は、VAEによって潜在変数として変換

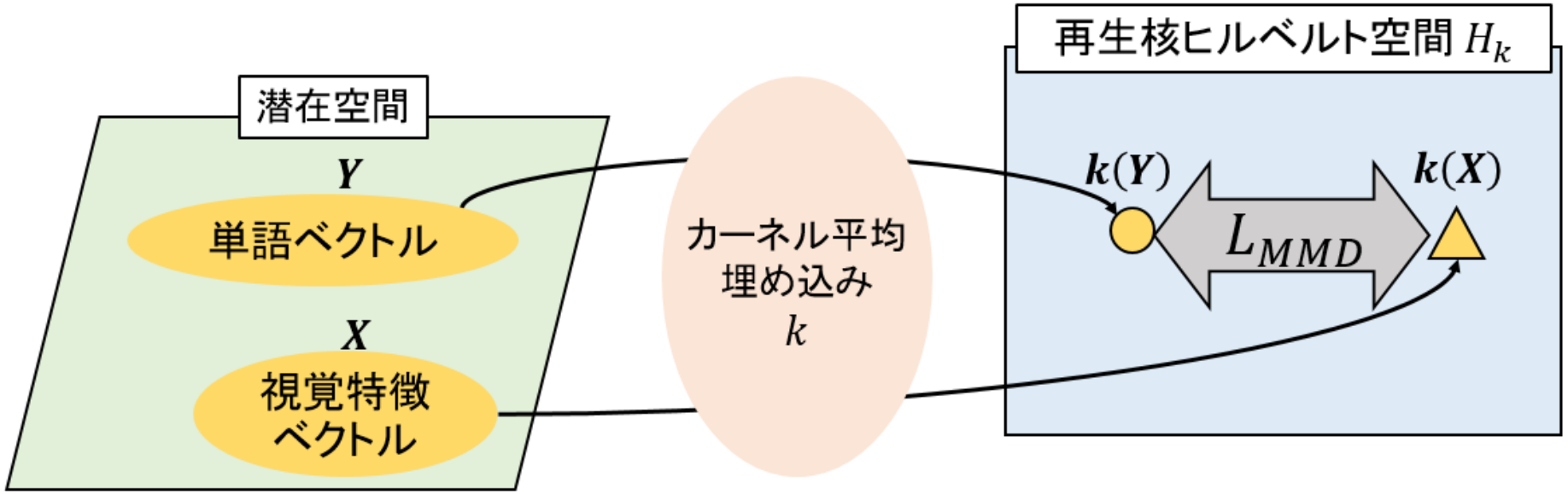


提案手法：MMDによる損失関数

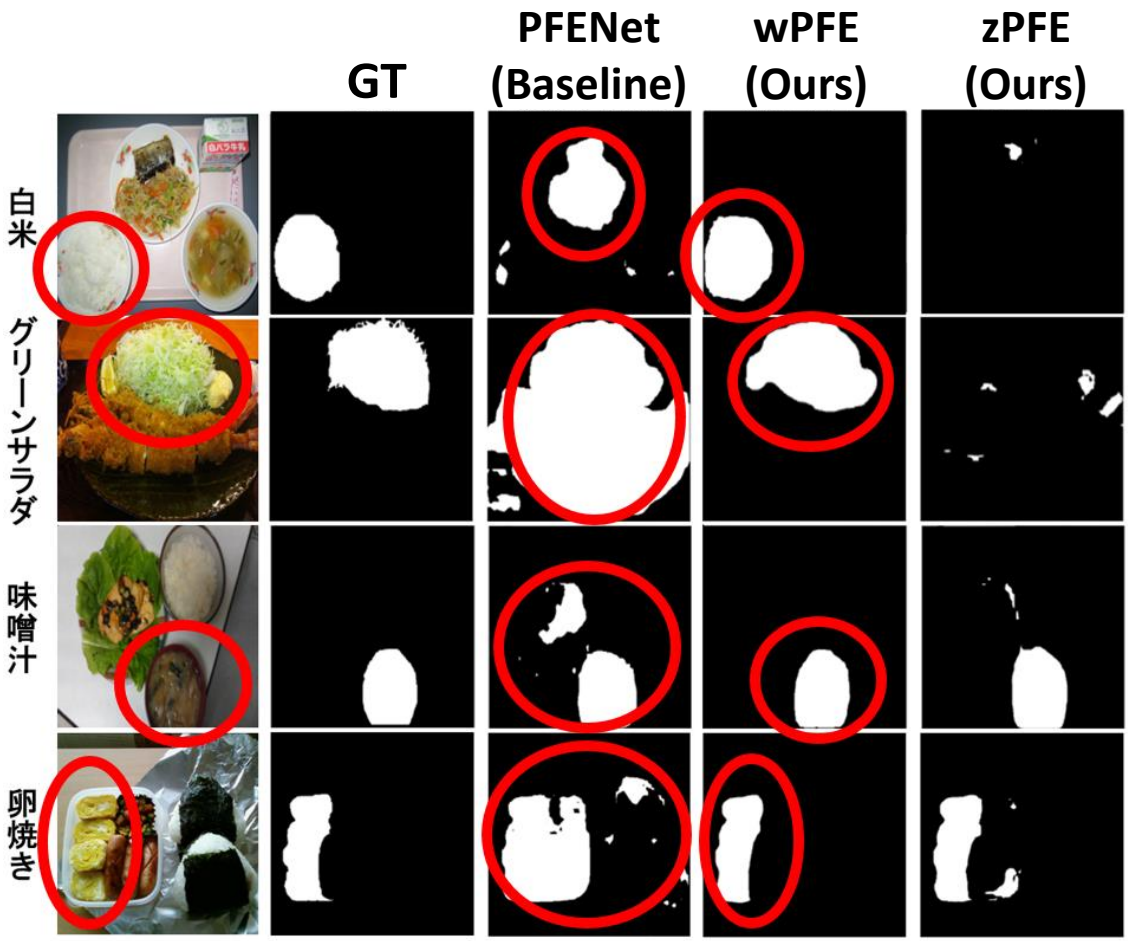
- Maximum Mean Discrepancy に基づく損失関数 L_{MMD} の追加
  - 2つの分布をGaussianカーネルで再生核ヒルベルト空間に写像した分布間の平均2乗誤差を使用することで分布の違いを定量

$$L^2_{MMD}(X_n, Y_m) = \|k(X) - k(Y)\|^2_{H_k}$$

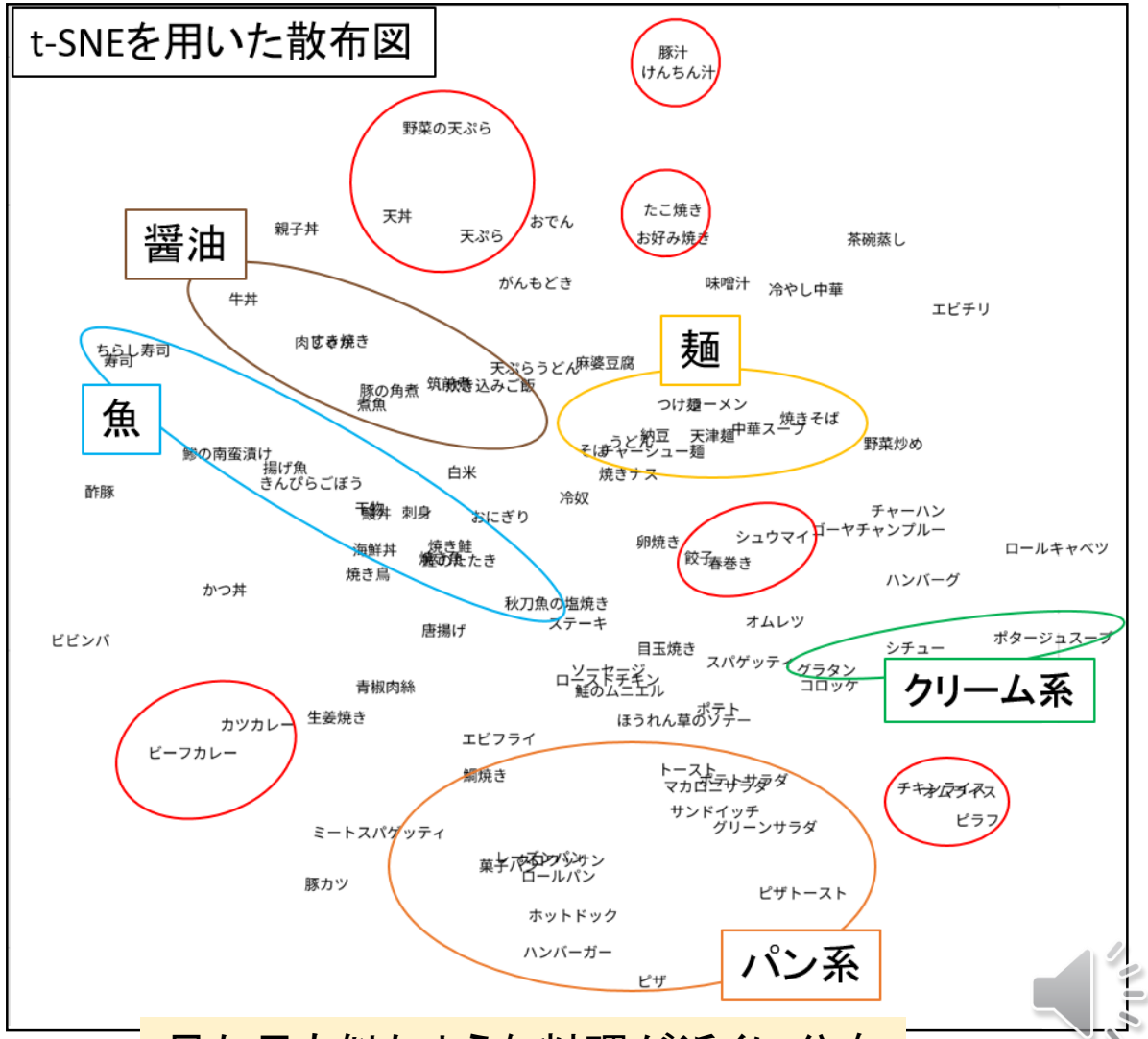
$$= \frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i, X_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(Y_i, Y_j) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j)$$



実験結果：食事データセットUECFoodPix-25^i

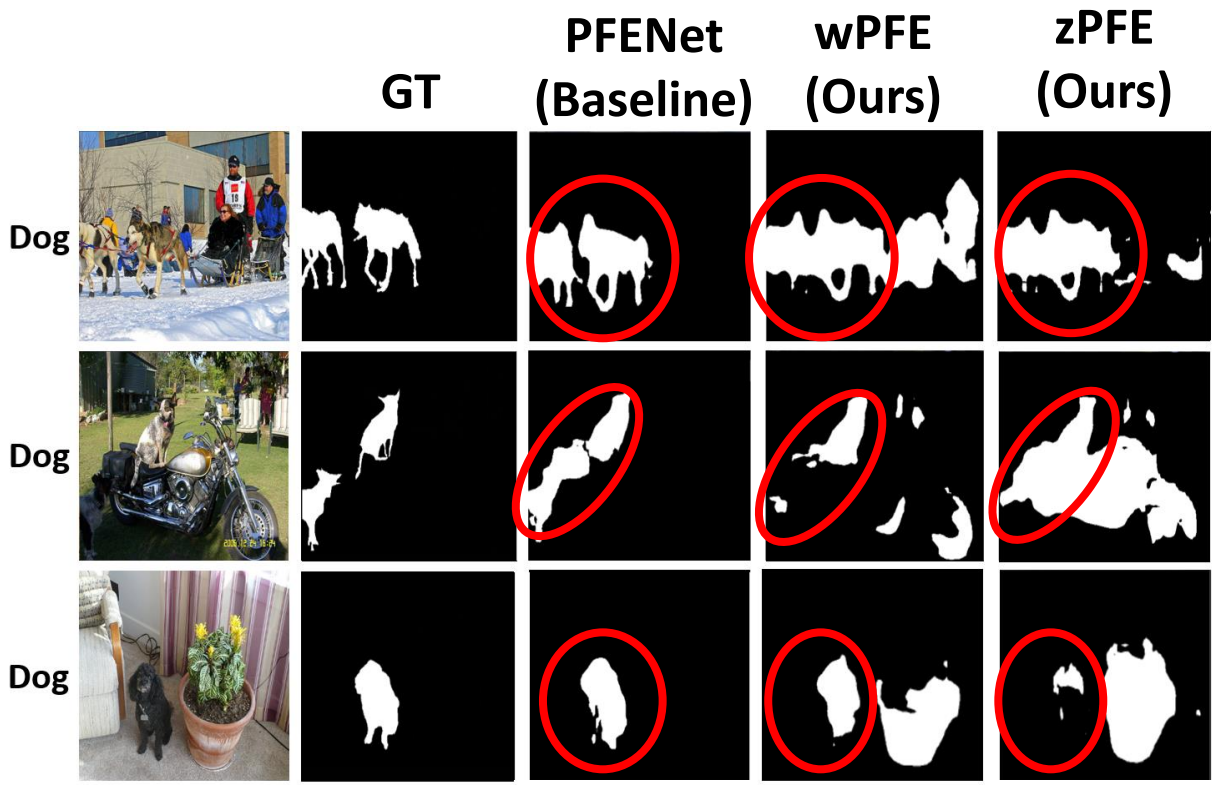


	UECFoodPix-25^i						
	Five-shot		One-shot		Zero-shot		
	wPFE (Ours)	PFE	wPFE (Ours)	PFE	zPFE (Ours)	PFE	Kato
Mean	**0.845**	0.838	**0.843**	0.832	**0.816**	0.786	0.736

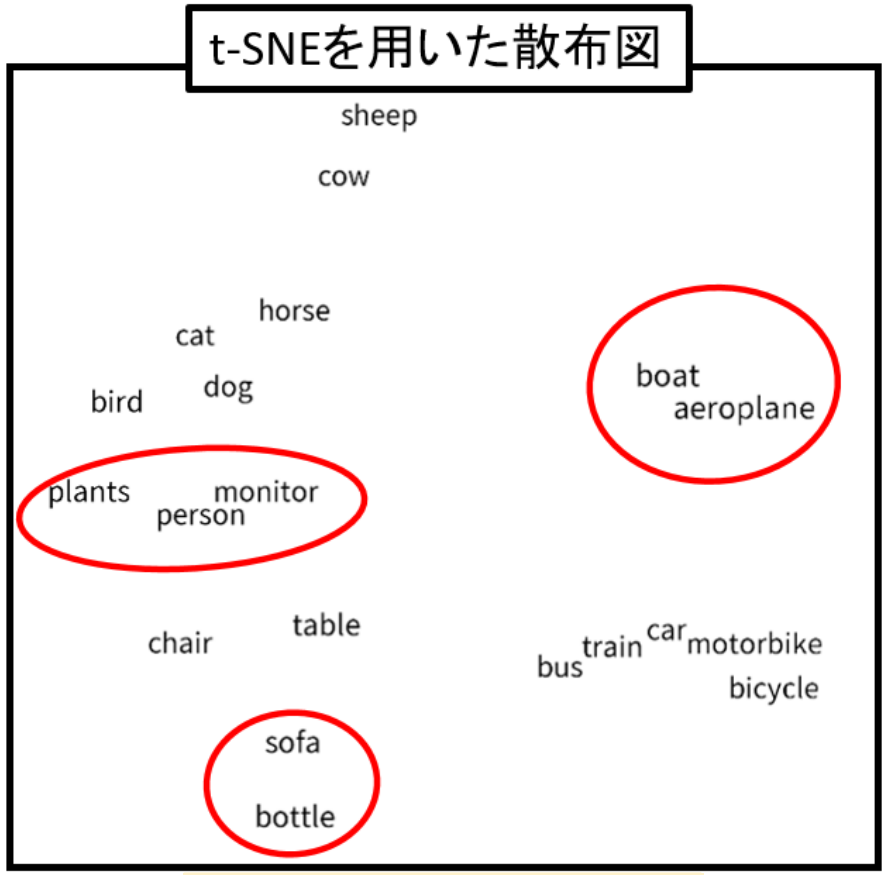


見た目も似たような料理が近くに分布

実験結果：一般物体データセットPascal-5^i



	Pascal-5^i						
	Five-shot		One-shot		Zero-shot		
	wPFE (Ours)	PFE	wPFE (Ours)	PFE	zPFE (Ours)	PFE	Kato
Mean	0.591	**0.621**	0.586	**0.607**	0.517	**0.551**	0.454

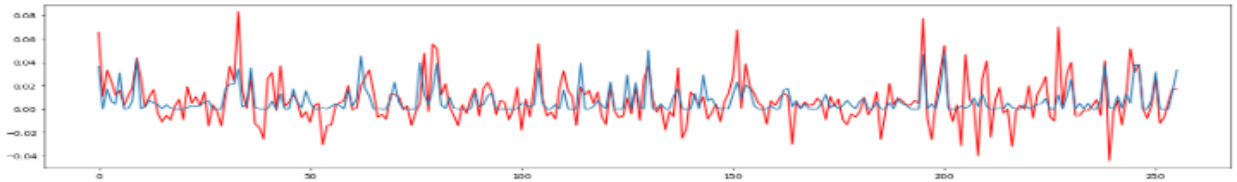


視覚的には近くない

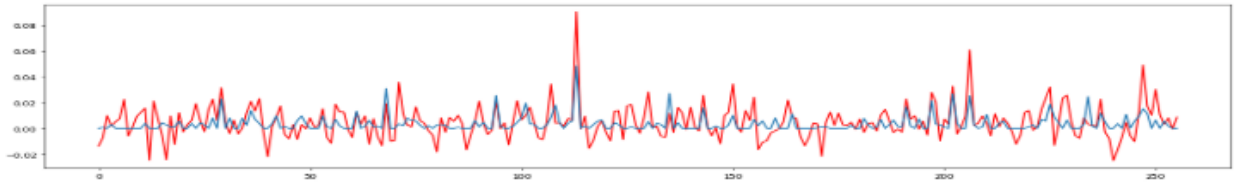
実験結果：埋め込みの有効性

- 学習データにwikiの文章データを用いる場合と、レシピデータの違いの実験
- wikiの文章データよりもレシピデータを使用するほうが料理の視覚情報を反映
- VAEを使用したほうがより複雑な表現を獲得できる

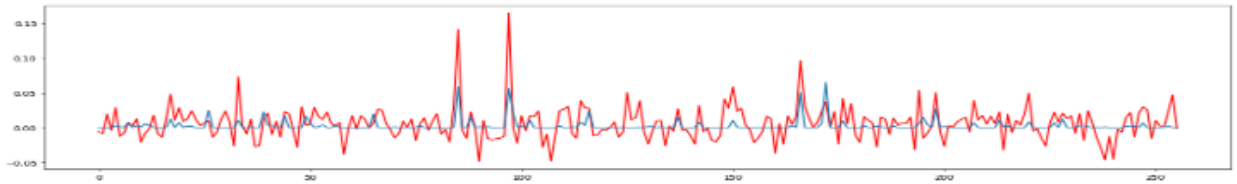
酢豚



けんちん汁



冷やし中華



赤色：再構成した特徴量、青色：クエリのMAPベクトル

学習データの違いによる比較			
	CookPad (10K)	CookPad (160K)	Wiki (40M)
Mean	0.796	**0.816**	0.802

単語の視覚特徴への再構手法の比較			
	+GMMN +MMD	+VAE +MMD	None
Mean	0.834	**0.843**	0.832

まとめ

- PFENetをベースに単語埋め込みと新たな損失を加えた新しいモデルを提案

- UECFoodPix-25^i ： 精度改善
  食事画像同士の類似度が高く、食材単語の共起関係が効果的

- Pascal-5^i ： 精度低下
  wikiの文章によるword2vecの学習は、視覚情報を反映していない