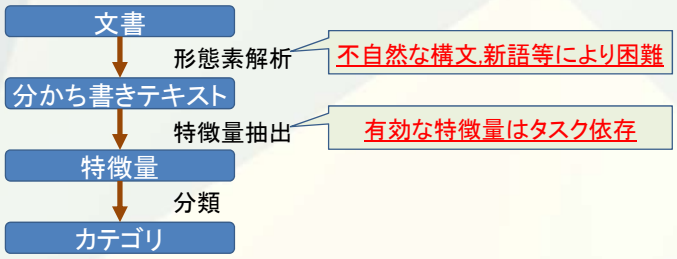
# 文字画像による Character-level Embeddingと文書分類

小谷 龍ノ介，島田 大樹，彌冨 仁
法政大学 理工学研究科
{ryunosuke.kotani.9p, daiki.shimada.9g} @stu.hosei.ac.jp

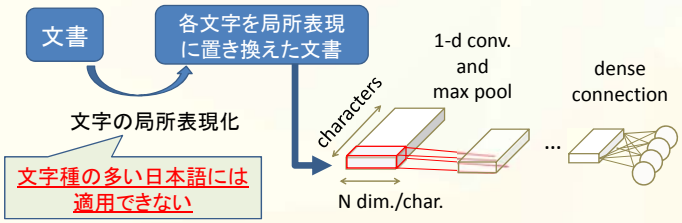
## 1. Background

### 従来の日本語文書分類



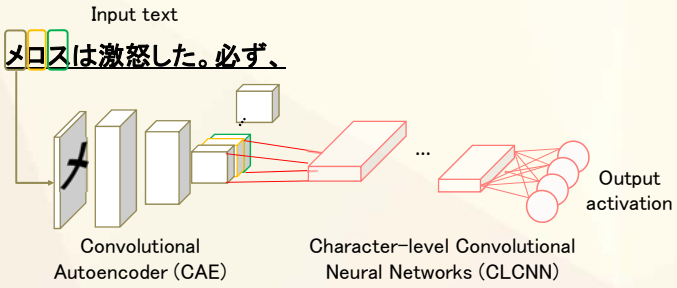

### Character-Level Convolutional Neural Networks (CLCNN)

➢ 英語文書において，特徴量抽出・分類を自動化




X. Zhang, J. Zhao, and Y. LeCun, Character-level convolutional networks for text classification, In Advances in Neural Information Processing Systems, pp. 649 657, 2015.

## 2. Method

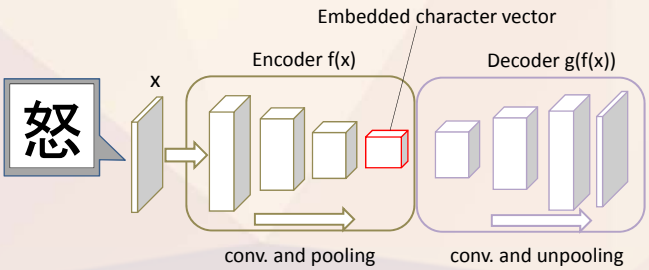



➢ 文字画像を入力としたCAEにより，文字の密なベクトル表現を獲得

➢ CLCNNを適用することにより，形態素・単語分割が不要

➢ ***Wildcard Training*** によるData Augmentationによって汎化性能を向上

## 2.1. Convolutional Autoencoder (CAE)




➢ 対象となる文字画像を入力
- ひらがな・カタカナ・漢字(JIS 第一・二水準)・英字・記号を含む計6,625字
- 32x32ピクセルの文字画像

➢ 文字を任意の次元の密ベクトルにすることが可能

➢ 形の似た文字は意味的にも近い文字となることを仮定し，意味的情報を考慮したベクトルを獲得することを期待

## 2.2. CLCNN

文書から任意の文字数分を切り出し，1次元方向に畳み込むCNN

➢ 文字表現の畳み込みにより，各文字の共起について考慮される

**本実験で用いたCLCNNのアーキテクチャと学習**
- 入力文字数は128字
- CAEで対象となった文字以外は零ベクトルとして入力
- 全4層構造
- 事前学習によるパラメータの選定は行わない
- 目的関数はクロスエントロピー誤差関数
- 誤差逆伝播法により重みを更新

## 2.3. Wildcard Training

CLCNNの汎化性能を高めるためには**大量のデータセット**が必要

日本語のData Augmentationでは類似語置換など**単語分割等が前提**

➢ 学習時の入力の一部を零ベクトル (wildcard)とする

対象となった文字の情報がCLCNNを伝播しないため，任意の文字 (wildcard)としてCLCNN内で扱われることを期待

➢ 形態素解析等を前提としないData Augmentationが可能

## 3. Evaluation

● 64次元文字ベクトルからの再構成画像


● 評価用データセット＆タスク

**◆日本語文書の著者推定**
- 青空文庫で公開されている著者10名 計104作品
- 学習：81作品，評価：23作品

**◆Web新聞記事の新聞社推定**
- 朝日，毎日，産経，読売の各新聞社の政治・経済・国際カテゴリの記事
- 各社5,610件 計22,440件で構成
- 学習：17,952件，評価：4,488件

● 日本語文書の著者推定の結果

| method | accuracy[%] |
|---|---|
| CLCNN (+wildcard training) | **69.57** |
| CLCNN | 52.17 |
| 3-gram (n=50,000) | 56.52 |
| 形態素解析 (n=50,000) | 47.83 |

※3-gram, 形態素解析は最頻出語上位n語によるTF-IDFベクトルを用いたロジスティック回帰の結果

● Web新聞記事の新聞社推定の結果

| method | accuracy[%] |
|---|---|
| CLCNN (+wildcard training) | **86.72** |
| CLCNN | 80.95 |
| 3-gram (n=50,000) | 65.78 |
| 3-gram (n=100,000) | 81.71 |
| 3-gram (n=300,000) | 84.27 |
| 形態素解析 (n=49,684**) | 67.22 |

** 学習データに出現する全形態素を用いた

## 4. Conclusion

✓ CLCNNの日本語文書分類問題への適用手法と適用可能性を示した

✓ CLCNNの学習法としてWildcard Trainingを提案し，汎化性能が向上することを確認した

✓ 提案手法が，従来の分類手法を上回る性能が得られることを示した