Text Understanding from Scratch

12056 外山洋太

Text Understanding from Scratch

- 著者
  - Xiang Zhang, Yann LeCun
- Advances in Neural Information Processing Systems (NIPS) 2015

Text Understanding from Scratch

- 文書理解
  - 文字レベルの入力から文書の意味を理解
- 目的
  - 文字から文書全体まで幅広く意味を捉えたい
- 方法
  - 深層学習（畳み込みニューラルネットワーク）
- タスク
  - オントロジー分類
  - 評判分類
  - カテゴリ分類

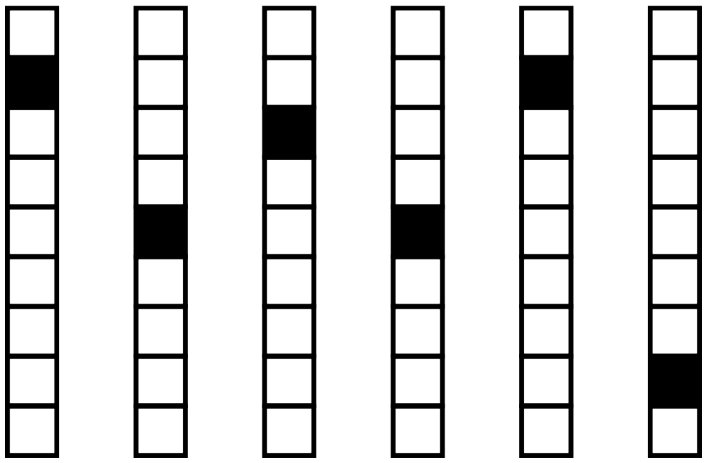
提案手法

- 入力：文書
- 出力：クラス
  (e.g. レビューのレーティング、文書のカテゴリなど)
- 処理の流れ
  1.文書内の文字を量子化
  2.畳み込みNNと全結合NNで文書を分類

文字の量子化

- 文字の量子化
  - 各文書を文字ベクトルの並びで置き換え
  - 文字ベクトルはone-hotベクトル
  - 実験では69文字（＝文字ベクトルの次元数）
    - a-z, 0-9, 「-,;.!?:'\”/\@#$%^&*~`+-=<>()[]{}\n」

toyota



文字の量子化

- 文書の長さがまちまちな問題
  - 文書からn文字取ってくる
  - 長さが満たない
    ⇒ 空白文字（ゼロベクトル）で埋める
  - 各文書は（考える文字の数）×（文書の長さ（固定））の行列
  - 実験では文書の長さに下限（100文字等）を設定
    → 短すぎる文書は無視

畳み込みニューラルネットワーク

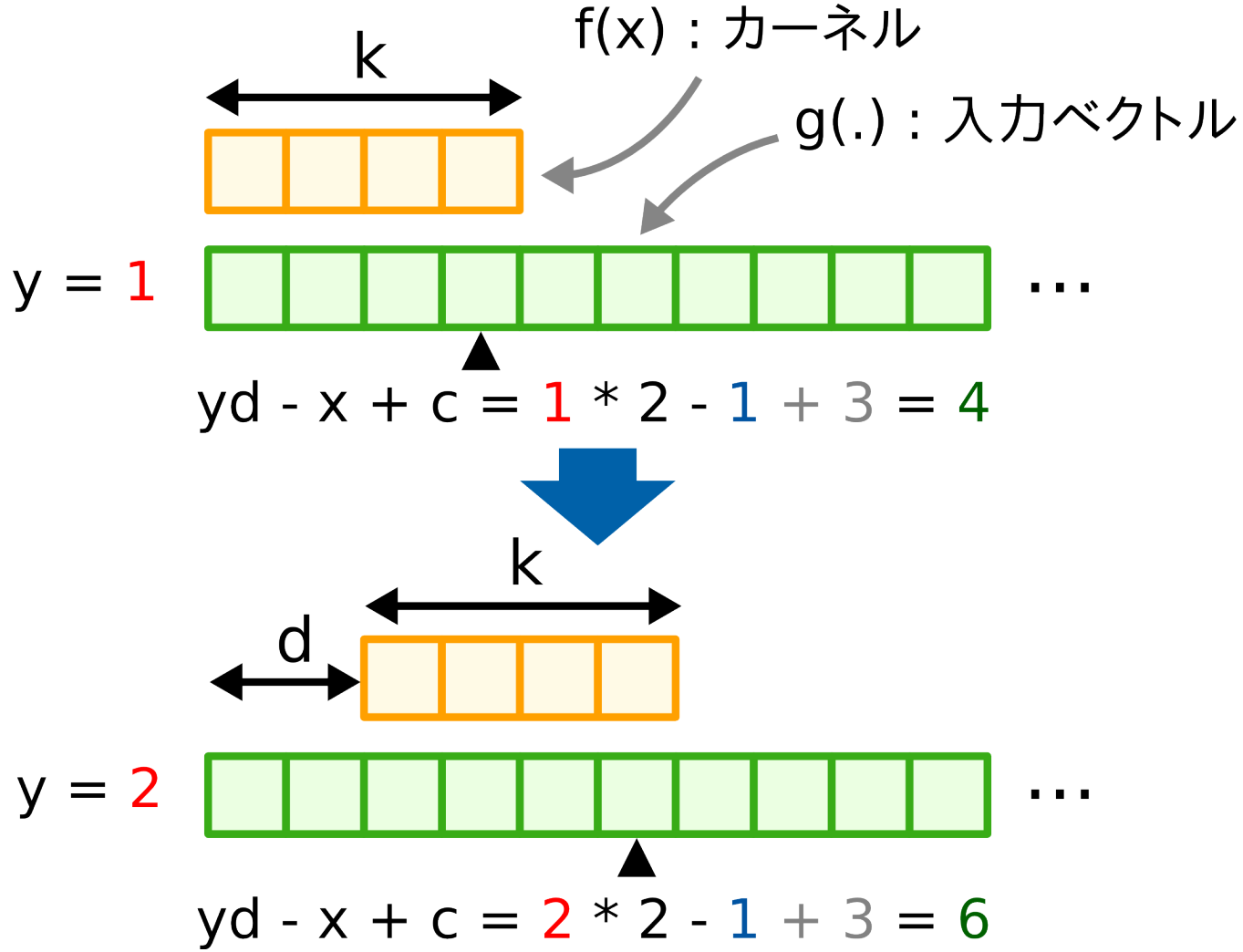
- 畳み込み層

$$h(y) = \sum_{x=1}^{k} f(x) \cdot g(y \cdot d - x + c),$$

  - g(a) : a番目の文字のベクトル（入力）
  - h(y) : y番目の出力ベクトルの要素（出力）
  - f(x) : カーネル中の位置xにおける重み
  - k : カーネルサイズ
  - x : カーネル中の位置
  - d : カーネルが一度に進む量
  - c = k – d + 1

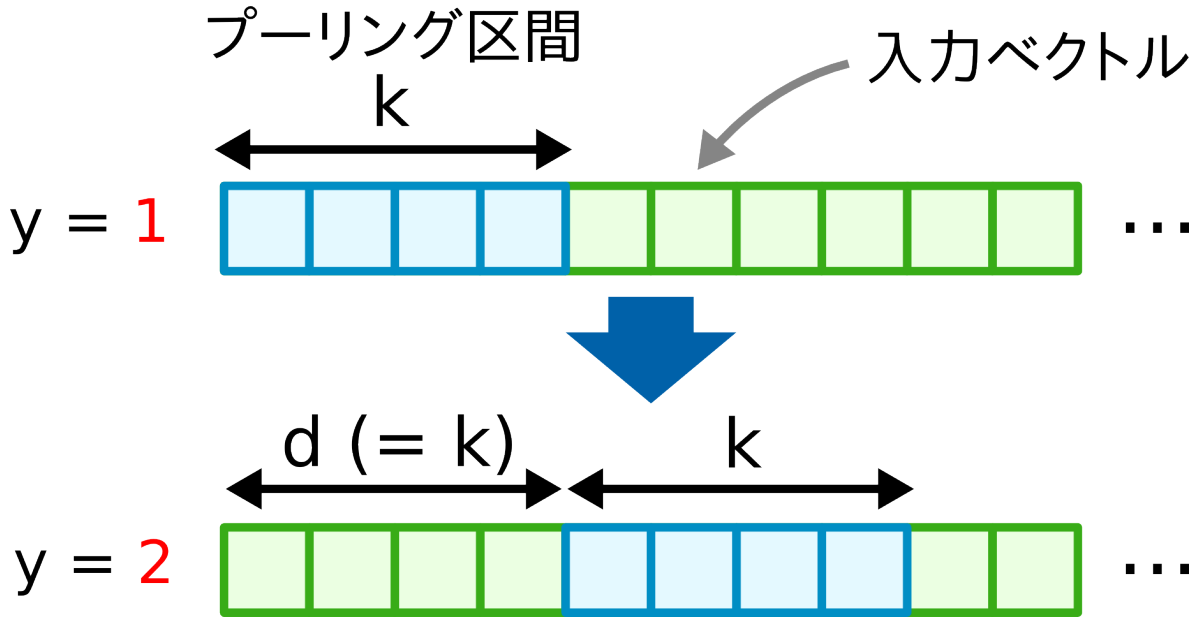
畳み込みニューラルネットワーク

- 畳み込み層



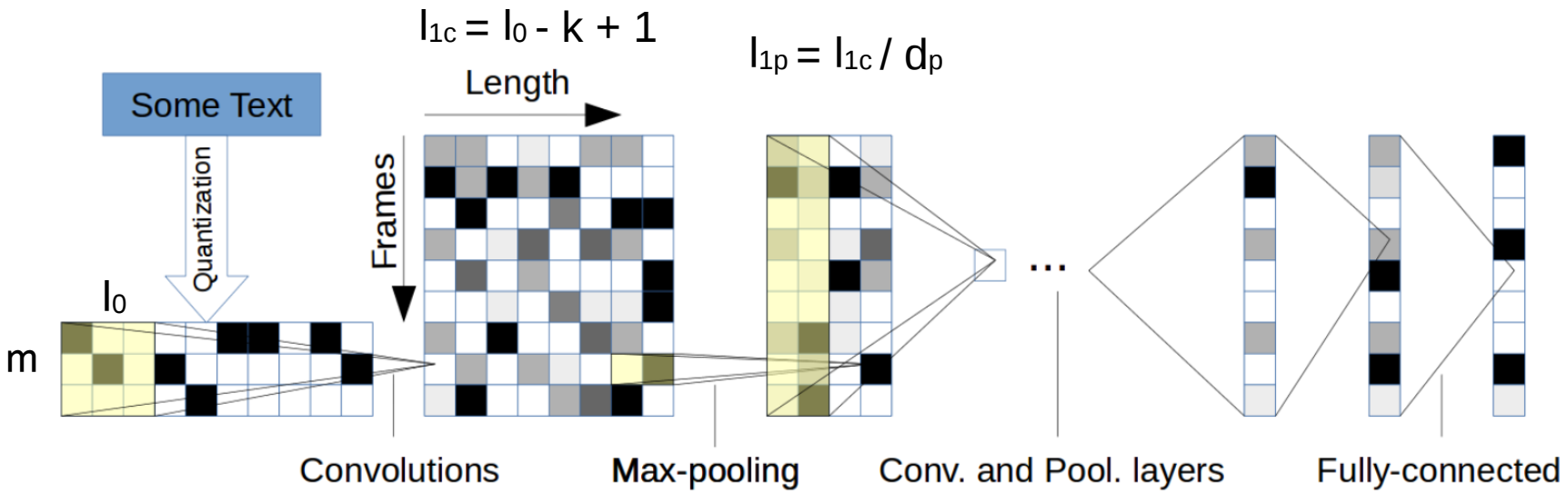
畳み込みニューラルネットワーク

- プーリング層
  - プーリング区間の中で平均／最大値を取る
  - 一般にプーリング区間はオーバーラップしない



モデル

- 畳み込みNN + 全結合NN



  - l_0 : 入力文書の長さ（文字数）
  - m (= 69) : 文字ベクトルの次元数

シソーラスによるデータ補完

- 少ないデータ→過学習
  ⇒ 文書内の単語を同義語で置き換えて
  データ数を補完
- 置き換えの方法
  - r個の単語を置き換え
    P[r] ~ p^r（幾何分布）
  - シソーラスからs番目に意味が近い同義語を選択
    P[s] ~ q^s
  - 実験では、p = q = 0.5

実験

- DPpediaのクラス分類
  - DPpedia : Wikipediaベースの構造化された情報
  - サンプルをオントロジークラスに分類
  - 各サンプルで「タイトル」と「概要」を持つものを使用

Table 3. DBpedia ontology classes. The numbers contain only samples with both a title and a short abstract.

Class	Total	Train	Test
Company	63,058	40,000	5,000
Educational Institution	50,450	40,000	5,000
Artist	95,505	40,000	5,000
Athlete	268,104	40,000	5,000
Office Holder	47,417	40,000	5,000
Mean Of Transportation	47,473	40,000	5,000
Building	67,788	40,000	5,000
Natural Place	60,091	40,000	5,000
Village	159,977	40,000	5,000
Animal	187,587	40,000	5,000
Plant	50,585	40,000	5,000
Album	117,683	40,000	5,000
Film	86,486	40,000	5,000
Written Work	55,174	40,000	5,000

Table 4. DBpedia results. The numbers are accuracy.

Model	Thesaurus	Train	Test
Large ConvNet	No	**99.96%**	98.27%
Large ConvNet	Yes	99.89%	**98.40%**
Small ConvNet	No	99.37%	98.02%
Small ConvNet	Yes	99.62%	98.15%
Bag of Words	No	96.29%	96.19%
word2vec	No	89.32%	89.09%

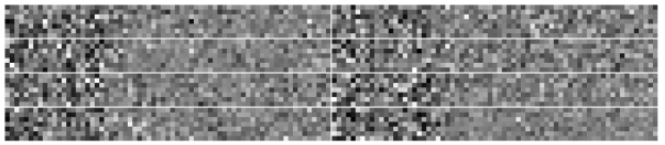


Figure 3. Visualization of first layer weights

実験

- 中国語のニュース分類
  - 中国語のニュースを各カテゴリに分類
  - 中国語の文書はピンインで英語に変換

Table 12. Sogou News dataset

Category	Total	Train	Test
Sports	645,931	90,000	12,000
Finance	315,551	90,000	12,000
Entertainment	160,409	90,000	12,000
Automobile	167,647	90,000	12,000
Technology	188,111	90,000	12,000

Table 13. Result on Sogou News corpus. The numbers are accuracy

Model	Thesaurus	Train	Test
Large ConvNet	No	**99.14%**	**95.12%**
Small ConvNet	No	93.05%	91.35%
Bag of Words	No	92.97%	92.78%

考察

- 日本語や中国語だと文字数が多い
  ⇒ 文字ベクトルの次元が大きい
  - ピンインやローマ字化では文字の意味が失われる
- 比較手法が余り強そうでない
  - word2vecを用いた手法では単語ベクトルをクラスタリングして、クラスタの重心のbagを分類
  - 最近の深層学習を用いた手法との比較が分からない
- 文書を記号の列として扱う
  - 自然言語処理以外にも使えそう

まとめ

- 文字から畳み込みNNを用いて文書理解を行うモデルを提案
- 文字からでも文書分類、評判分類が可能
  - ほぼ事前知識なし
  - 辞書や形態素解析も要らない