進捗発表

3EP4-35 鶴瀬和輝

研究テーマ

YouTube動画の予測再生数と予測高評価数を
取得するシステムの実装

やったこと

・Word2vecについて学習

・Doc2vecについて学習

・チャンネルIDのデータ収集

・データの整形

やったこと

・Word2vecについて学習

・Doc2vecについて学習

・チャンネルIDのデータ収集

・データの整形

Word2vecってなに？

単語の意味や文法を捉えるために単語をベクトル化して学習する技術

単語の分散表現を獲得できる

分布仮説

単語の意味は周囲の単語によって形成される

カウントベースと推論ベース

カウントベース

　　周囲の単語の頻度によって単語を表現する手法

推論ベース

　　ニューラルネットワークを用いて少量の学習サンプルを見ながら
　　重みを繰り返し更新する手法

Word2vecで使用するニューラルネットワークのモデル

・CBOW(continuous bag-of-words)

・skip-gram

CBOWモデル

コンテキストからターゲットを推測することを

目的とするニューラルネットワーク

Corpus

0 1 2 3 4 5 6

You say goodbye and I say hello.

You say goodbye and I say hello.

You say goodbye and I say hello.

You say goodbye and I say hello.

You say goodbye and I say hello.

You say goodbye and I say hello.

Contexts	Target
You, goodbye	say
say, and	goodbye
goodbye, I	and
and, say	I
I, hello	say
say, .	hello

Contexts	Target
[0 2]	1
[1 3]	2
[2 4]	3
[3 1]	4
[4 5]	1
[1 6]	5

one-hotベクトル化

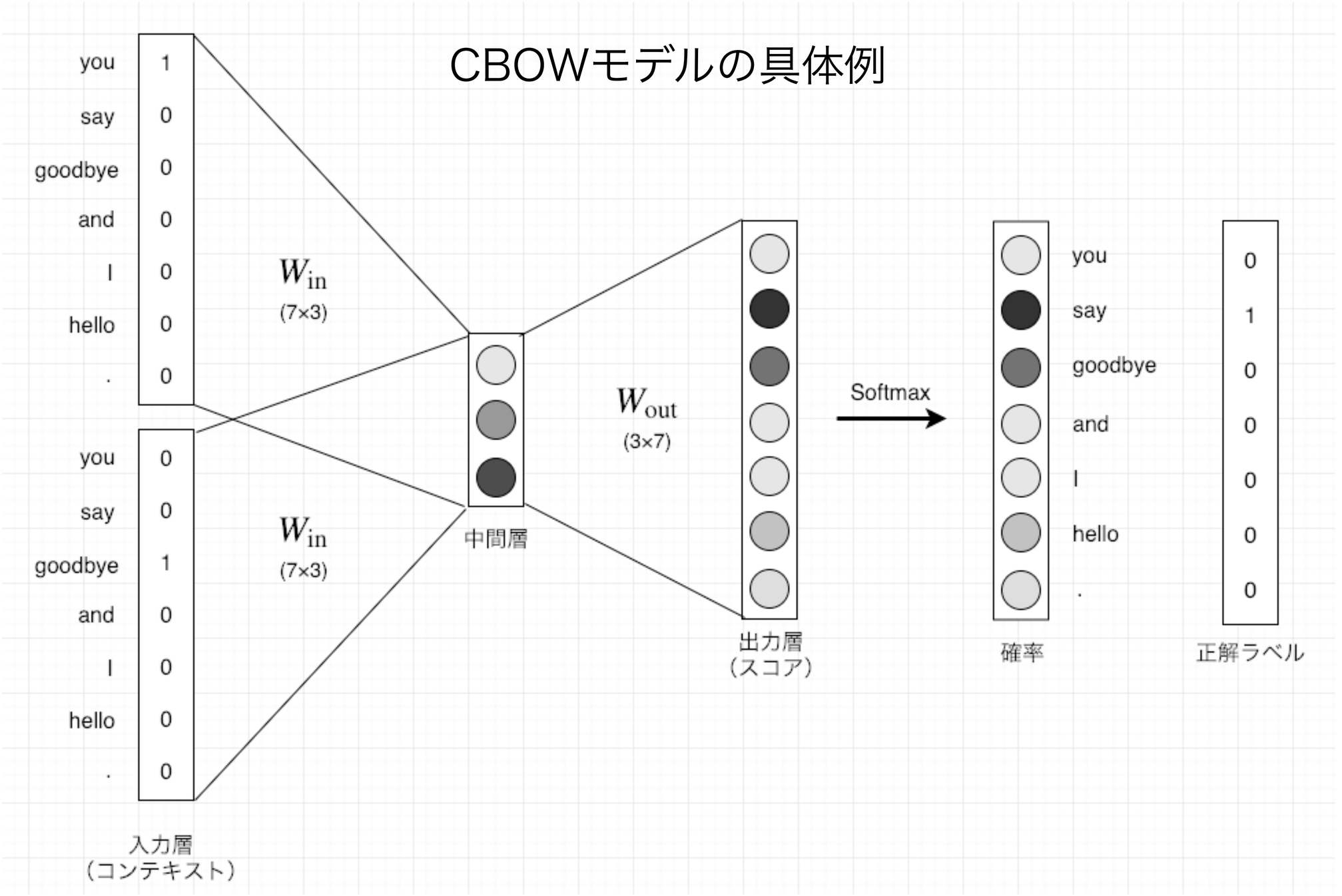
Contexts

```
[[[1 0 0 0 0 0 0]
  [0 0 1 0 0 0 0]]
 [[0 1 0 0 0 0 0]
  [0 0 0 1 0 0 0]]
 [[0 0 1 0 0 0 0]
  [0 0 0 0 1 0 0]]
 [[0 0 0 1 0 0 0]
  [0 1 0 0 0 0 0]]
 [[0 0 0 0 1 0 0]
  [0 0 0 0 0 1 0]]
 [[0 1 0 0 0 0 0]
  [0 0 0 0 0 0 1]]]
```

Target

```
[[0 1 0 0 0 0 0]
 [0 0 1 0 0 0 0]
 [0 0 0 1 0 0 0]
 [0 0 0 0 1 0 0]
 [0 1 0 0 0 0 0]
 [0 0 0 0 0 1 0]]
```

CBOWモデルの具体例



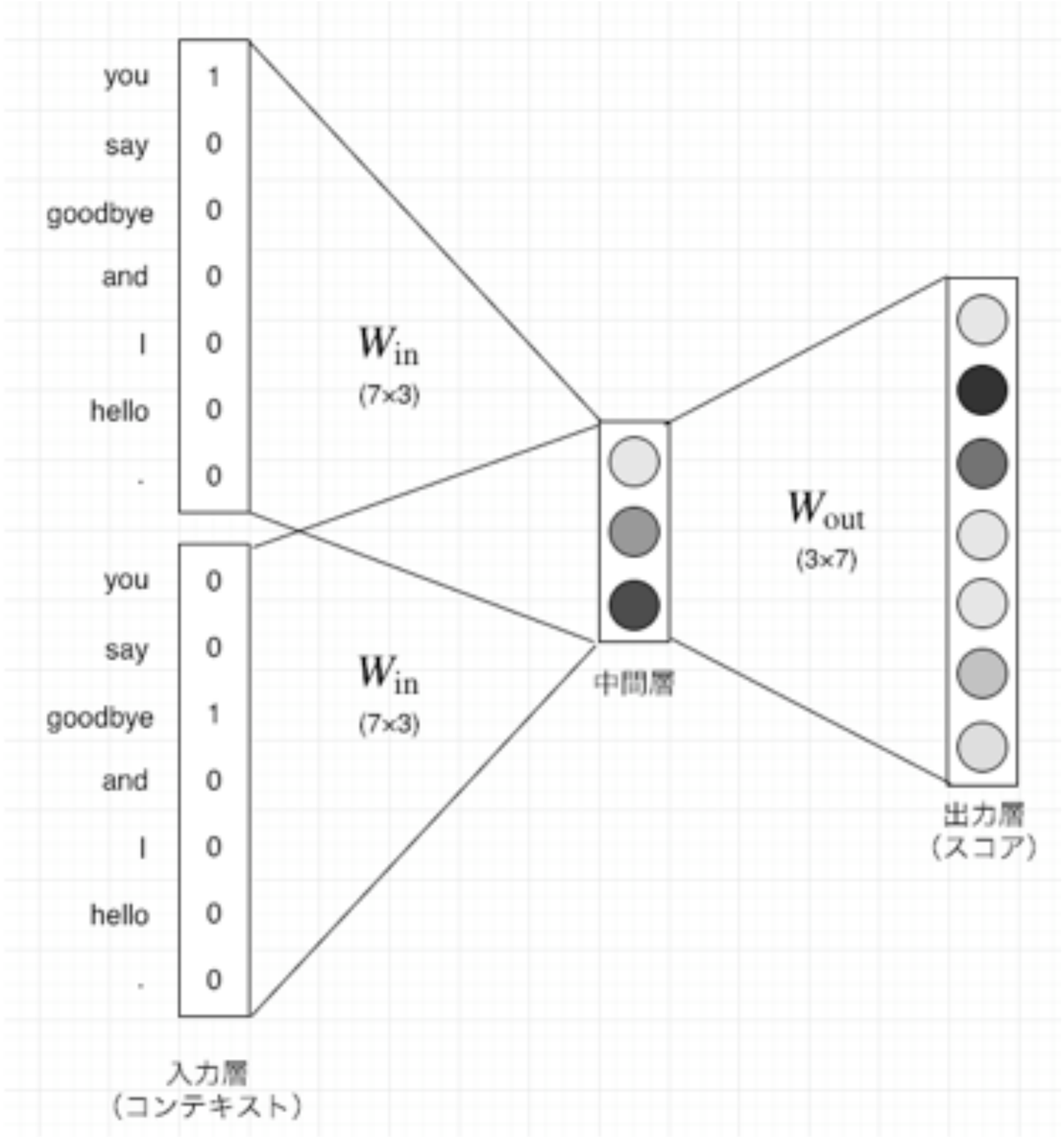
CBOWモデルのネットワーク

Ｗ ： 重み

入力層から中間層への変換を全結合層によって行う

↓

中間層から出力層のニューロンの変換は別の全結合層で行う

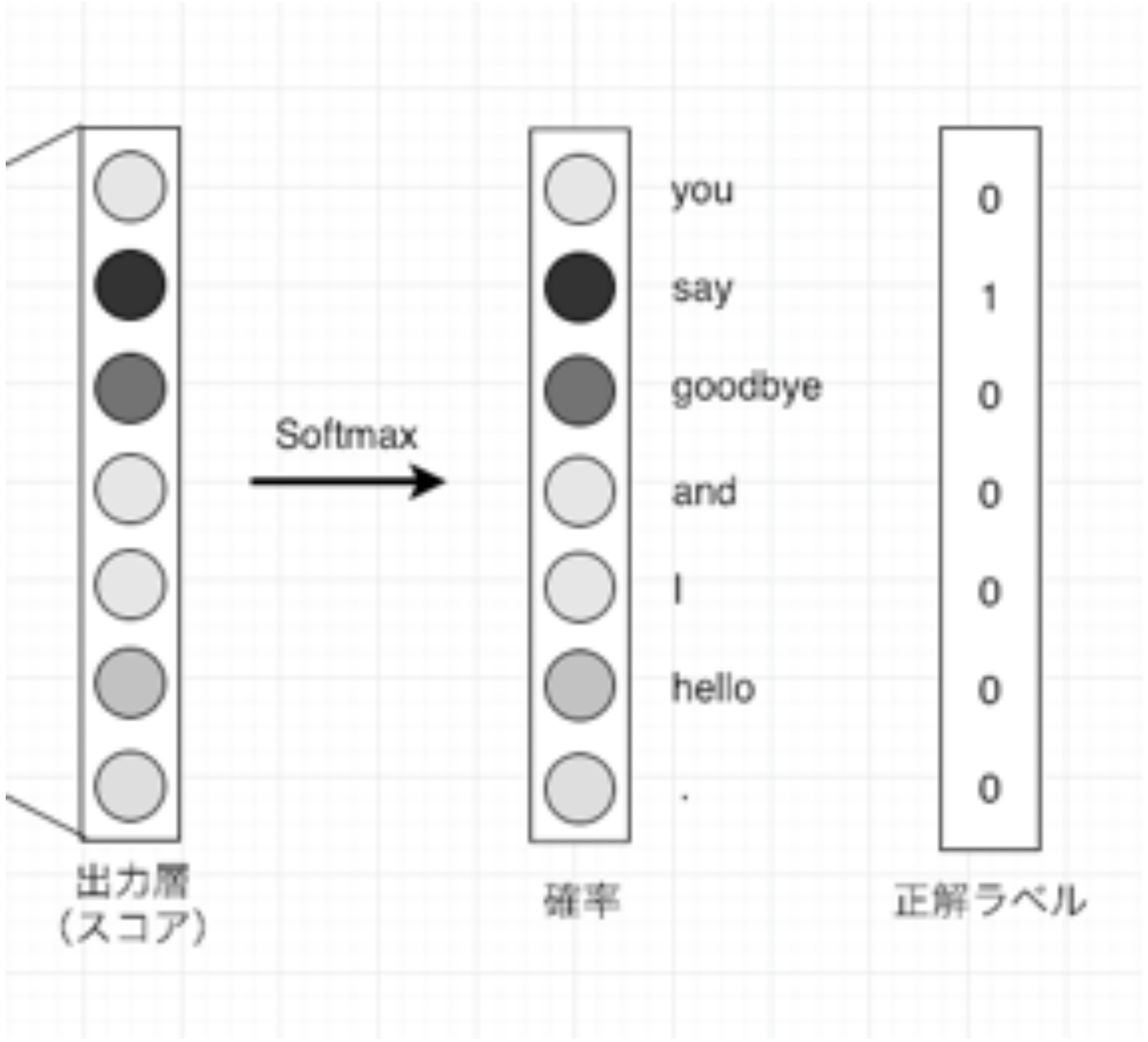


出力層

出力層のニューロンは
各単語の「スコア」

softmax関数

$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_i}}$$



今後について

Kazuki Tsuruse 17:09

活動記録

- 研究に使うデータを取得してgithubにアップロード
- doc2vecの学習

疑問点
自分の研究は動画タイトル・動画カテゴリ・チャンネル登録者数の入力からYouTube動画の予測再生数と予測高評価数を取得するというテーマだが、
ユーザーが入力する動画タイトルとすでに存在している(学習に使う)動画タイトルの類似度を特徴量として扱いたいと思っている。
そのための方法として以下がある。

方法1
ユーザーが新しく入力するタイトルの中にある重要度が高い単語をdoc2vecで学習させたモデルの
similarメソッドのパラメータに渡し、文書間の類似度を求める。
方法2
コーパス(動画タイトルが入っている)の中に、ユーザーが新しく入力するタイトルを入れ、コーパスの中にあるタイトルをベクトル化させ、doc2vecで学習させる。
類似度の確認方法は、「新しく入力するタイトル」の文書IDをmost_similar メソッドのパラメータとして扱えば、類似度が取得できるはず

方法1では一度作成したモデルから類似度を求めることが出来るが、方法2ではタイトルを入力するたびにモデルを作る必要があるため処理に時間がかかる。
しかし方法1では調べたところ、パラメータに文書ではなく単語しか渡す事しかできないっぽいので、ユーザーが新しく入力したタイトルを渡すことが出来ないので方法としてはイマイチだと思っています。

自分としてはユーザーが新しく入力したタイトル一つだけをモデルに組み込み、類似度を算出したいと思っているが、イマイチ方法がわからないです。