

ニューラルネットワークで自然言語処理をかじってみた

芝浦工業大学 数理科学研究会
市毛 竣

令和元年 11月 1日

1 研究背景と内容

何かと話題になっている人工知能. その中でも自然言語処理は数多くの研究がなされている. 特に深層学習の発達により, 以前よりも人間らしい会話ができるようになった. そのため深層学習の根幹であるニューラルネットワークと自然言語処理はどのように組み合わせられているのかを知りたいと思い, 後の研究につなげるため発表に至る. 内容としては, ニューラルネットワークのアルゴリズムと, それを用いる自然言語処理の理論の一つである word2vec を簡単な例でまとめた.

2 ニューラルネットワークのアルゴリズム

ニューラルネットワークのアルゴリズムでは, 与えられた入力 x に対し, 重みパラメータと呼ばれる ω と b を用いて計算し, 得た最終的な出力結果 y と実際のデータを比較し, そのデータの値に近づくように重みパラメータの値を更新するというを行う. この行為を千回, 一万回と行い, 重みパラメータを少しずつ実際のデータに合うように更新していく. このときのパラメータの更新を学習と呼ぶ.

3 word2vec

自然言語処理にはいろいろな方法があるが, ニューラルネットワークを使う場合, 推論ベースというものがよく用いられている. 推論ベースではどの場所に, 何の単語が入るかを推論する, というを行う. この推論問題を解くための手法が word2vec である. word2vec では次の二通りのモデルが使われている.

3.1 CBOW モデル

CBOW モデルでは, ある単語のコンテキスト (前後の単語) を入力データ x として扱い, コンテキストの中央の単語を推論する. 次に示す文では, 「you」と「goodbye」が入力データとして扱われ, ???を推論するというを行う.

???に入る単語は?

you ??? goodbye and I say hello .

3.2 skip-gram モデル

skip-gram モデルでは, ある単語を入力データ x として扱い, その単語のコンテキストを推論する. 次に示す文では, 「say」が入力データとして扱われ, その前後の二つの???を推論するというを行う.

???に入る単語は?

??? say ??? and I say hello .

4 損失関数

学習するには, 出力結果と実際のデータの差の指標として, 損失関数 L を定めなければならない. t 番目の単語を ω_t , 単語の総数を T とし, 事後確率を使う.

CBOW モデルでは

$$L = -\frac{1}{T} \sum_{t=1}^T \log P(\omega_t | \omega_{t-1}, \omega_{t+1}) \quad (1)$$

skip-gram モデルでは

$$L = -\frac{1}{T} \sum_{t=1}^T (\log P(\omega_{t-1} | \omega_t) + \log P(\omega_{t+1} | \omega_t)) \quad (2)$$

のように損失関数を定める.

今後について

実際にデータを集めて自然言語処理をしてみたい. 特に, データを集める際, 特定の人物やキャラクターの言葉のみをデータにすることで, どのように結果が変わるかなども研究していきたい.

参考文献

- [1] ゼロから作る Deep Learning -Python で学ぶディープラーニングの理論と実装-, 斎藤 康毅, O'Reilly 社, 2016.
- [2] ゼロから作る Deep Learning2 -自然言語処理編-, 斎藤 康毅, O'Reilly 社, 2018.