BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NIÊN LUẬN CƠ SỞ
NGÀNH CÔNG NGHỆ THÔNG TIN

Đề tài

PHÂN LOẠI VĂN BẢN SỬ DỤNG CNN
Áp dụng cho Tiếng Việt

Sinh viên: Nguyễn Phước Thành
Mã số: B1610669
Khóa: 42

Cần Thơ, 02/2020

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG
BỘ MÔN CÔNG NGHỆ THÔNG TIN

NIÊN LUẬN CƠ SỞ
NGÀNH CÔNG NGHỆ THÔNG TIN

Đề tài

PHÂN LOẠI VĂN BẢN SỬ DỤNG CNN
Áp dụng cho Tiếng Việt

Người hướng dẫn
TS Lâm Nhựt Khang

Sinh viên thực hiện
Nguyễn Phước Thành
Mã số: B1610669
Khóa: 42

Cần Thơ, 02/2020

Lời cảm ơn

Cảm ơn TS. Lâm Nhựt Khang, Bộ môn Công nghệ Thông tin, Khoa Công nghệ Thông tin và Truyền thông, trường Đại học Cần Thơ đã tích cực hướng dẫn, giúp đỡ nghiên cứu đề tài này.

Mục lục

Danh mục đồ thị, biểu bảng và hình ảnh

Tóm lượt

Hiện nay, các tiến bộ khoa học kỹ thuật dần thay thế các hoạt động thủ công truyền thống của con người, cùng với đó là sự tiến bộ không ngừng của công nghệ thông tin trong thời đại ngày nay đòi hỏi chiếc máy tính của chúng ta phải làm được nhiều hơn nữa, cuộc cách mạng công nghệ 4.0 đã thúc đẩy các nhà khoa học trong việc biến ước mơ của con người về trí tuệ nhân tạo thành hiện thực, những ứng dụng của trí tuệ nhân tạo ngày càng đa dạng, những thành tựu mà trí tuệ nhân tạo đem lại khiến chúng một phần không thể thiếu trong một xã hội phát triển nhanh chóng như ngày hôm nay. Hoạt động của trí tuệ nhân tạo được dựa trên tri thức và hành vi của con người, bên cạnh đó là bề dày về số lượng thuật toán cũng như giải thuật được thiết kế một cách tỉ mỉ. Trí tuệ nhân tạo bao gồm nhiều thành phần: xử lý ảnh, xử lý ngôn ngữ tự nhiên, học máy, học sâu, ...

Phần giới thiệu

Hiện nay phần lớn các doanh nghiệp đang đối mặt với "cơn lũ" dữ liệu về mọi mặt: feedback của khách hàng, thông tin đối thủ cạnh tranh, emails của khách hàng, các văn bản về sản phẩm và kỹ thuật. Việc khai thác các dữ liệu này là điểm mấu chốt để doanh nghiệp có thể nhanh chóng đưa ra các kế hoạch hoặc hành động kịp thời so với các đối thủ cạnh tranh.

Vấn đề ở đây là có quá nhiều thông tin cần xử lý cùng lúc, và kích thước dữ liệu ngày càng tăng. Điều này sẽ là bất khả thi nếu chỉ dựa vào sức người trong một giới hạn về số lượng và thời gian nhất định.

Tiếng nói và chữ viết là hai phạm trù cơ bản của ngôn ngữ. Việc nhận biết cũng như phân biệt ý nghĩa của một câu nói xem ra khá đơn giản đối với con người, nhưng làm sao để máy tính có thể phân tích và đưa ra những nhận định gần giống con người đối với dữ liệu đầu vào là văn bản khi máy tính chỉ có thể hiểu ở những con số 0, 1 liên tiếp nhau.

Mục tiêu của đề tài này là tìm hiểu về việc phân loại văn bản (Text Classification) bằng việc sử dụng CNN (Convolutional Neural Network) để áp dụng cho Tiếng Việt.

Bố cục của bản báo cáo gồm 3 phần: phần giới thiệu, phần nội dung và phần kết luận. Trong đó phần nội dung gồm có 4 chương:

Chương 1 - Đặc tả yêu cầu: Mô tả phương pháp phân loại văn bản.

Chương 2 - Thiết kế giải pháp: Trình bày các lý thuyết có liên quan được sử dụng trong đề tài.

Chương 3 - Cài đặt giải pháp: Mô tả hoạt động của phương pháp phân loại văn bản sử dụng CNN (Convolutional Neural Network) thông qua ví dụ minh họa

Chương 4 - Đánh giá kiểm thử: Đánh giá và hướng phát triển

Phần nội dung

Chương 1 - Đặc tả yêu cầu

Xử lý ngôn ngữ là một phạm trù trong xử lý thông tin với dữ liệu đầu vào là ngôn ngữ, có thể được xem như văn bản hay tiếng nói. Đây là các dạng văn bản chính được lưu trữ dưới dạng tài liệu, đặc điểm chung của chúng là không có cấu trúc (non-structured) hoặc nửa cấu trúc (semi-structured) và không thể lưu lại dưới dạng bảng biểu. Vì vậy chúng ta cần phải xử lý chúng để máy tính có thể hiểu được.

Một trong những ứng dụng rộng rãi của xử lý ngôn ngữ tự nhiên (NLP) và máy học có giám sát đó là "phân loại văn bản". Mục tiêu là nhằm phân loại một cách tự động một văn bản (câu / từ) thuộc vào một hoặc nhiều nhóm đã được định nghĩa từ trước.

Phương pháp hiện tại là sử dụng các kỹ thuật có sẵn cũng như các thuật toán điển hình để xử lý theo từng giai đoạn:

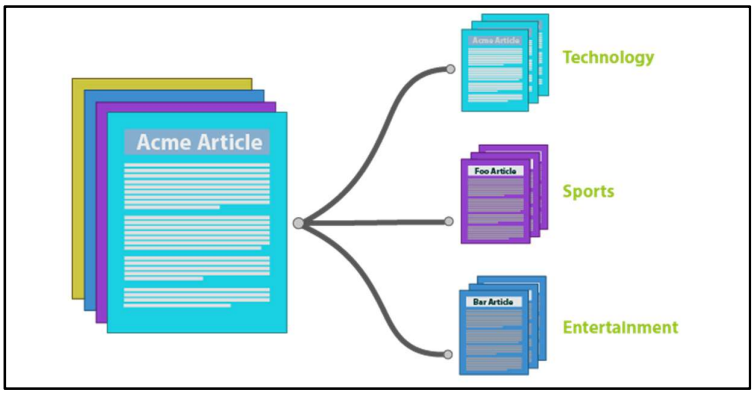
Giai đoạn 1: Tiếp cận và tiền xử lý dữ liệu trước khi thực hiện chuyển các văn bản thành các vector đưa vào ma trận.

Giai đoạn 2: Sử dụng mô hình CNN cho việc huấn luyện tập dữ liệu và đưa ra mô hình phân loại văn bản.

Chương 2 - Thiết kế giải pháp

Việc tiếp cận một bài toán liên quan đến xử lý ngôn ngữ tự nhiên, thông thường sẽ trải qua các mức phân tích:

- Phân tích hình thái: cách từ được xây dựng, các tiền tố và hậu tố của từ.
- Phân tích cú pháp: mối liên hệ về cấu trúc và ngữ pháp giữa các từ và ngữ.
- Phân tích ngữ nghĩa: nghĩa của từ, cụm từ và cách diễn đạt.
- Tích hợp văn bản: quan hệ giữa các ý hoặc các câu.
- Phân tích thực nghĩa: mục đích phát ngôn, cách sử dụng ngôn ngữ trong giao tiếp.

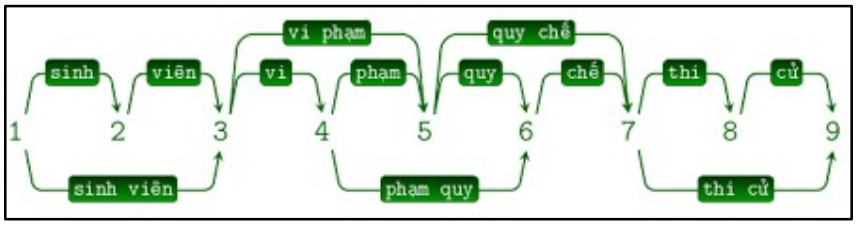


Mô hình được mong đợi sẽ có khả năng phân tích được loại bài viết dựa theo nội dung được dự đoán bởi mô hình phân loại.

2.1 Phương pháp tách từ Tiếng Việt - Tokenization and Word Segmentation

Việc tách từ là một trong cách phương pháp đầu tiên trong việc xử lý dữ liệu đối với các bài toán xử lý ngôn ngữ tự nhiên, mục đích của phương pháp này nhằm:

- Các khoảng trắng giữa các từ. (Điều này chỉ được phép đối với các ngôn ngữ như tiếng Việt, trong đó khoảng trắng được sử dụng để đánh dấu các ranh giới âm tiết thay vì ranh giới từ.)
- Loại bỏ các từ viết tắt, gây thừa thãi trong câu.
- Lọc các từ khóa chính làm nổi bật nội dung của cả câu.



2.2 Các phương pháp biểu diễn ngôn ngữ

Định nghĩa ***word embeddings*** (tập tự nhúng):

"Tập nhúng từ *là tên chung cho một tập hợp các mô hình ngôn ngữ và các phương pháp học đặc trưng trong xử lý ngôn ngữ tự nhiên (NLP), nơi các từ hoặc cụm từ từ vựng được ánh xạ tới vectơ số thực. Về mặt khái niệm, nó liên quan đến việc nhúng toán học từ một không gian với một chiều cho mỗi từ vào một không gian vectơ liên tục với kích thước thấp hơn nhiều."*

Tóm lại **word embedding** là phương pháp ánh xạ mỗi từ vào một không gian số thực nhiều chiều có kích thước nhỏ hơn nhiều so với kích thước từ điển.

Trước khi word embedding ra đời đã có rất nhiều phương pháp được áp dụng và nghiên cứu, từ mã hóa **one-hot encoding** cho đến **biểu diễn theo ma trận đồng nhất**,... nhưng những giải pháp này vẫn gặp phải vấn đề về chi phí tính toán. Cho đến năm 2013, nhóm Mikolov giúp giới NLP thở phào với giải pháp mới mang tên word2Vec. Từ thời điểm này hàng loạt bài toán NLP được giải quyết với độ chính xác cao hơn nhiều so với trước.

2.3 Tổng quát phương pháp

Ý tưởng chính của **word2vec** là:

- Thay vì lưu thông tin xuất hiện của các từ bằng cách đếm trực tiếp như ma trận đồng xuất hiện, word2vec học để đoán từ lân cận của tất cả các từ.
- Các giải pháp sau đó như Glove cũng tương tự word2vec được đề xuất bởi nhóm Pennington năm 2014.

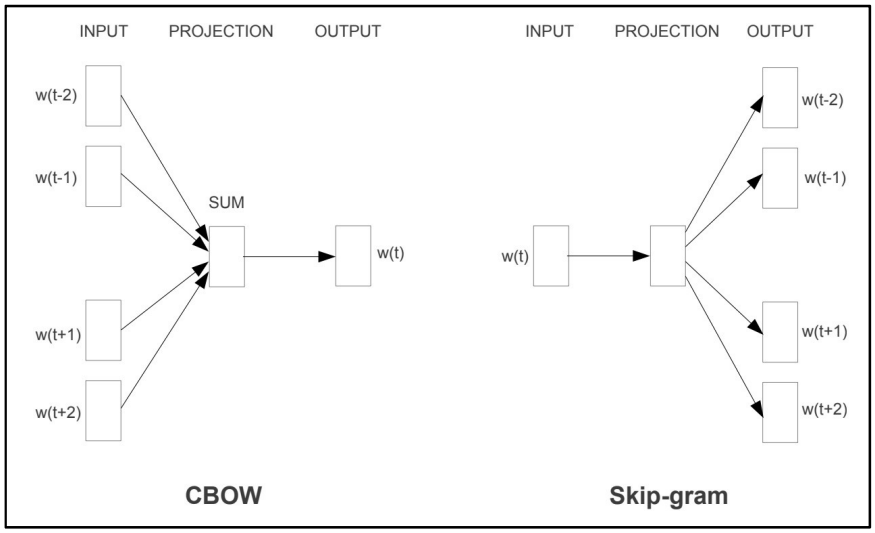
- Tính toán nhanh hơn và dễ dàng thêm dữ liệu mới vào trong mô hình

Phương pháp:
Đoán các từ lân cận trong cửa sổ m của mỗi từ 1.1. Với mỗi từ t = 1 . . . T1.2. Đoán các từ trong cửa sổ bán kính m của tất cả các từ.

Hàm mục tiêu(object function): tối ưu hợp lý hóa cực đại của bất kì từ ngữ cảnh (context word) đối với một từ đang xét hiện tại (center word).

$$J(\theta) = -\frac{1}{T}\prod_{t=1}^{T}\prod_{-m\leq j\leq m, j\neq 0} p(w_{t+j}|w_t;\theta)$$

Các kiến trúc khác nhau: (1) cho ngữ cảnh đoán từ hiện tại (CBoW), và (2) cho từ hiện tại đoán ngữ cảnh (Skip-gram).



CBoW:

- Cho các từ ngữ cảnh
- Đoán xác suất của một từ đích

Skip-gram:

- Cho từ đích
- Đoán xác suất của các từ ngữ cảnh

Tốc độ học mô hình:

Model	*Vector Dimensionality*	*Training Words*	*Training Time*	*Accuracy [%]*
Collobert NNLM	50	660M	2 months	11
Turian NNLM	200	37M	few weeks	2
Mnih NNLM	100	37M	7 days	9
Mikolov RNNLM	640	320M	weeks	25
Huang NNLM	50	990M	weeks	13
Skip-gram (hier.s.)	1000	6B	hours	66
CBOW (negative)	300	1.5B	**minutes**	**72**

Hình trên[1] so sánh hiệu năng của các mô hình khác nhau. Mô hình CBOW cho kết quả ấn tượng với thời gian học bằng phút thay vì bằng giờ hay tháng như các mô hình trước đây. Tất nhiên đây không phải là những so sánh tuyệt đối công bằng do các mô hình được đánh giá trên các tập dữ liệu khác nhau trên những máy tính có tốc độ xử lý khác nhau. Nhưng phần nào cũng thể hiện được ưu điểm mà word2vec mang lại.

Chương 3 - Cài đặt giải pháp

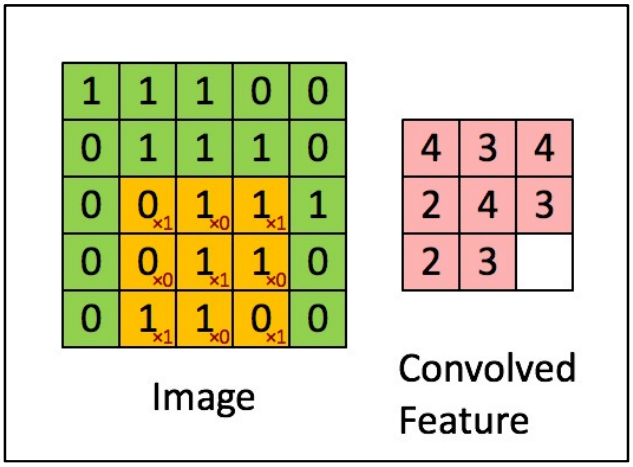
3.1. Khái quát CNN (Convolutional Neural Network)

Khi nghe đến CNN chúng ta thường nghĩ đến Computer Vision. CNN hầu hết đảm nhận các công việc liên quan đến phân loại ảnh (Image Classification) và xuất hiện trong đa số các hệ thống thị giác máy tính hiện nay từ tự động gán nhãn trên Facebook đến xe tự lái.

3.1.1. Tích chập (Convolution) là gì ?

Một trong những cách hiểu đơn giản về tích chập đó là cửa sổ trượt (ma trận) trên một ma trận.

[1] Theo Mikolov tại #mostly.ai/summit



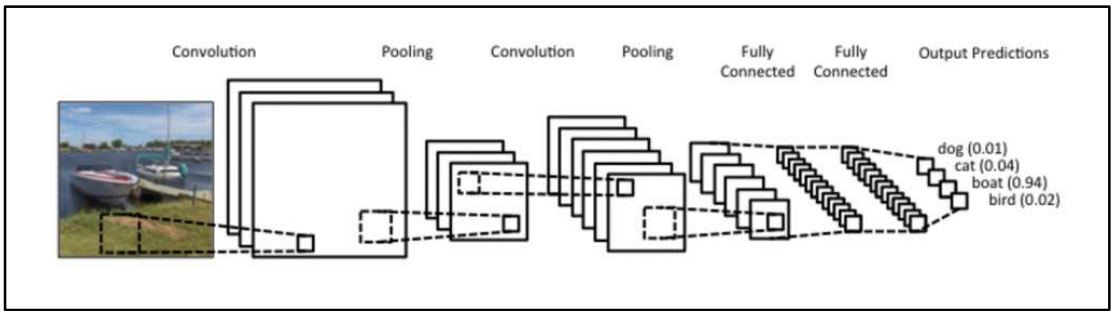
Bộ lọc sẽ biến đổi ma trận f và g (thường được gọi là ma trận đầu vào) để tạo thành một ma trận mới. Đối với ảnh đầu vào thì bộ lọc sẽ được gia giảm các tham số trên ma trận lọc để có thể dễ dàng lấy được các đặc trưng của ảnh.

3.1.2. Mạng nơ-ron tích chập (Convolutional Neural Network) là gì ?

Mạng nơ-ron tích chập (CNN) đơn thuần chỉ gồm một vài lớp (layer) tích chập (convolutional) kết hợp với hàm truyền phi tuyến (non-linear activation function) như Relu hay tanh để tạo ra thông tin trừu tượng hơn cho các lớp (layer) tiếp theo.

Các lớp liên kết với nhau thông qua cơ chế tích chập. Lớp tiếp theo là kết quả của lớp trước, nhờ vậy ta có các kết nối cục bộ. Nghĩa là mọi nơ-ron ở lớp tiếp theo sinh ra từ ma trận lọc (filter) áp đặt lên một vùng cục bộ nơron lớp trước đó.

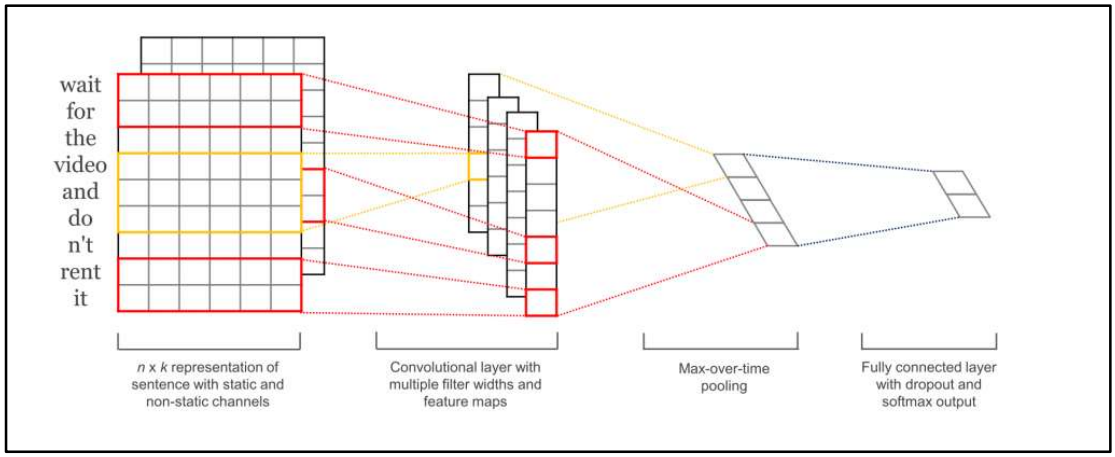
Trong bài toán phân loại ảnh, CNN sẽ học để phát hiện các đặc trưng của ảnh thông qua các cạnh trong lớp (layer) đầu tiên, sau đó sử dụng các cạnh này để phát hiện các đặc trưng phức tạp hơn trong các lớp tiếp theo. Lớp cuối cùng là lớp phân loại.



3.2. Áp dụng CNN vào NLP

Thay vì các pixel của ảnh, dữ liệu đầu vào của các bài toán NLP sẽ là các câu, các đoạn văn bản được trình bày dưới dạng ma trận. Mỗi hàng trong ma trận sẽ đại diện cho một từ. Thông thường các từ sẽ được vectơ hóa thông qua các kĩ thuật như word2vec hay Glove.

Trong CNN bộ lọc (ma trận) sẽ có chiều rộng bằng với chiều rộng của ma trận đầu vào:



Ta có thể thấy các lớp đầu tiên được sử dụng để giảm chiều của các vectơ đầu vào. Lớp tiếp theo thực hiện các phép tích chập sử dụng nhiều bộ lọc có kích thước khác nhau. Ví dụ: trượt hơn 3, 4 hoặc 5 từ cùng một lúc. Tiếp theo, gộp tối đa kết quả của lớp tích chập thành một vectơ đặc trưng dài, thêm các lớp pooling nhằm giữ lại các đặc trưng kem theo bỏ các phần thừa và phân loại kết quả bằng cách sử dụng lớp softmax.

Chương 4 - Đánh giá kiểm thử

Bài toán vận dụng sẽ được sử dụng trong bài toán phân loại bài báo tiếng Việt thuộc 1 trong 10 thể loại: chính trị xã hội, đời sống, khoa học, kinh doanh, pháp luật, sức khỏe, thế giới, thể thao, văn hóa, vi tính.

Ví dụ bài báo mẫu: "*Thủ tướng Đức nhận lời tham dự lễ kỷ niệm D-Day Thủ tướng Gerhard Schroeder sẽ trở thành nguyên thủ Đức đầu tiên tham dự lễ kỷ niệm ngày quân đồng minh đổ bộ lên bãi biển Normandy trong Thế chiến II (mang mật danh D-Day) vào tháng 6 tới. Ông đã chấp nhận lời mời tham gia lễ kỷ niệm 60 năm ngày D-Day của Tổng thống Pháp Jacque Chirac. Phát ngôn viên của Berlin cho biết: "Tổng thống Chirac đã mời Thủ tướng Schroeder... ".*

Phía trên là bài báo thuộc thể loại **thế giới.**

4.1 Chuẩn bị dữ liệu

Dữ liệu phải được bỏ những ký tự đặc biệt trong văn bản như dấu chấm, phẩy, dấu đóng ngoặc, mở ngoặc... bằng cách sử dụng thư viện **gensim**. Sau đó sẽ sử dụng thư viện **PyPi** để tách từ tiếng Việt.

Kết quả thu được: "*thủ_tướng đức nhận_lời tham_dự lễ kỷ_niệm day thủ_tướng gerhard schroeder sẽ trở_thành nguyên_thủ đức đầu_tiên tham_dự lễ kỷ_niệm ngày quân đồng_minh đổ_bộ lên bãi biển normandy trong thế_chiến ii mang mật_danh day vào tháng tới ông đã chấp_nhận lời mời tham_gia lễ kỷ_niệm năm ngày day của tổng_thống pháp jacque chirac phát_ngôn_viên của berlin cho biết tổng_thống chirac đã mời thủ_tướng schroeder...*"

4.2 Feature Engineering

Ở giai đoạn này dữ liệu dạng văn bản sẽ được chuyển thành dạng vector thuộc tính có dạng số học, có rất nhiều phương pháp có thể được sử dụng: Count vectors as features, TF-IDF Vectors, Word Embeddings, ... Trong ví dụ chúng ta sẽ sử dụng phương pháp **Word Embeddings** phương pháp này sẽ chuyển một từ trong từ điển về một vector **n** chiều, bằng cách sử dụng thuật toán bag-of-words.

Trong ví dụ phần triển khai, một từ sẽ được biểu diễn bằng một **vector 300** chiều. Từ đó chúng ta có thể sử dụng chúng cho các mô hình *Deep Learning* như *Convolutional Neural Network* để phân loại văn bản.

4.3 Xây dựng mô hình và đánh giá mô hình

Sau khi tiến hành huấn luyện và kiểm thử trên mô hình CNN, kết quả cho ra tương đối khả quan sau 20 lần chạy huấn luyện.

```
Epoch 18/20
30383/30383 [==============================] - 8s 248us/step - loss: 0.1788 - a
Epoch 19/20
30383/30383 [==============================] - 7s 246us/step - loss: 0.1811 - a
Epoch 20/20
30383/30383 [==============================] - 7s 245us/step - loss: 0.1688 - a
Validation accuracy:  0.9016587677725119
Test accuracy:  0.9044924860540369
```

Bên cạnh đó mô hình cũng được chạy so sánh với các mô hình deep learning khác.

Mô hình	Độ chính xác
Naive Bayes	~0.86
Deep Neural Network	~0.90

Phần kết luận

Việc sử dụng mạng nơ-ron tích chập (CNN) hay các mô hình deep learning trong bài toán xử lý ngôn ngữ tự nhiên đem lại kết quả tương đối khả quan, nhưng đối với từng loại văn bản khác nhau đòi hỏi các kỹ thuật xử lý khác nhau, thêm vào đó việc lựa chọn các tham số thích hợp cho mô hình cũng đòi hỏi việc mất thời gian.

Vì là mô hình cơ bản được áp dụng nên việc cải tiến về mô hình là việc cần thiết nhằm đem lại kết quả tốt hơn.

Tài liệu tham khảo

[1] Deshpande, A. (2018). *The 9 Deep Learning Papers You Need To Know About (Understanding CNNs Part 3).*

[2] Yoon Kim (2015). *Convolutional Neural Networks for Sentence Classification.*

[3] Ronan Collobert (NEC Labs America, Princeton NJ), Jason Weston (Google, New York, NY), L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa (NEC Labs America, Princeton NJ) . *Natural Language Processing (almost) from Scratch.*

[4] N. T. Duyen, N. X. Bach, and T. M. Phuong, 2014."An empirical study on sentiment analysis for Vietnamese"International Conference on Advanced Technologies for Communications (ATC 2014), pp. 309–314.

[5] Hồ Tú Bảo[a, b], Lương Chi Mai[a] ([a]Viện Công nghệ Thông tin, [b]Viện Khoa học và Công nghệ Tiên tiến Nhật Bản), Lương Chi Mai (2015). *Về xử lý tiếng Việt trong công nghệ thông tin* .