1. Giới thiệu chung

* Word2Vec là một mô hình học sâu được phát triển bởi Google, dùng để chuyển các từ thành các vector số học trong không gian nhiều chiều.
* Đây là phương pháp trong Deep
* Phương pháp này có thể đưa ra các từ có ngữ nghĩa giống nhau dựa trên ngữ cảnh xung quanh

1. Các khái niệm cơ bản
   1. Vector từ ( Word vector)

* Mỗi từ tương ứng với một vector trọng số. Vector này được xây dựng từ các dữ liệu văn bản lớn sao.
* Các trọng số trong mỗi vector từ được tính toán trong quá trình huấn luyện mô hình học sâu để phản ánh mối quan hệ ngữ nghĩa giữa các từ xung quanh
  1. Vector-Oriented Reasoning
* Là việc sử dụng vector từ để tính toán và suy luận liên quan đến ngữ nghĩa các từ
* A diagram of a line graph

  Description automatically generated
* Ví dụ:
  + - **"Seattle"**: Là một thành phố ở Washington, Mỹ.
    - **"Portland Timbers"**: Là một đội bóng đá tại Portland, Oregon.
    - **"Portland"**: Là thành phố nơi đội bóng Portland Timbers có trụ sở.
    - Mô hình Word2Vec Mô hình Word2vec "biết" rằng các thuật ngữ Portland và Port land Timbers cách nhau gần bằng khoảng cách giữa Seattle và Seattle Sounders. Và các khoảng cách đó (sự khác biệt giữa các cặp vectơ) có cùng hướng. Vì vậy, mô hình Word2vec có thể được sử dụng để trả lời câu hỏi về phép loại suy đội thể thao của bạn. Sau khi cộng và trừ các vectơ từ, vectơ kết quả của bạn hầu như không bao giờ bằng chính xác một trong các vectơ trong vốn từ vựng vectơ từ của bạn. Các vectơ từ Word2vec thường có hàng trăm chiều, mỗi chiều có giá trị thực liên tục. Tuy nhiên, vec trong vốn từ vựng của bạn gần nhất với vectơ kết quả thường sẽ là câu trả lời cho câu hỏi NLP của bạn
    - A number and numbers in a line

      Description automatically generated with medium confidence

1. Hai kiến trúc được sử dụng trong Word2Vec
   1. Phương pháp skip-gram:

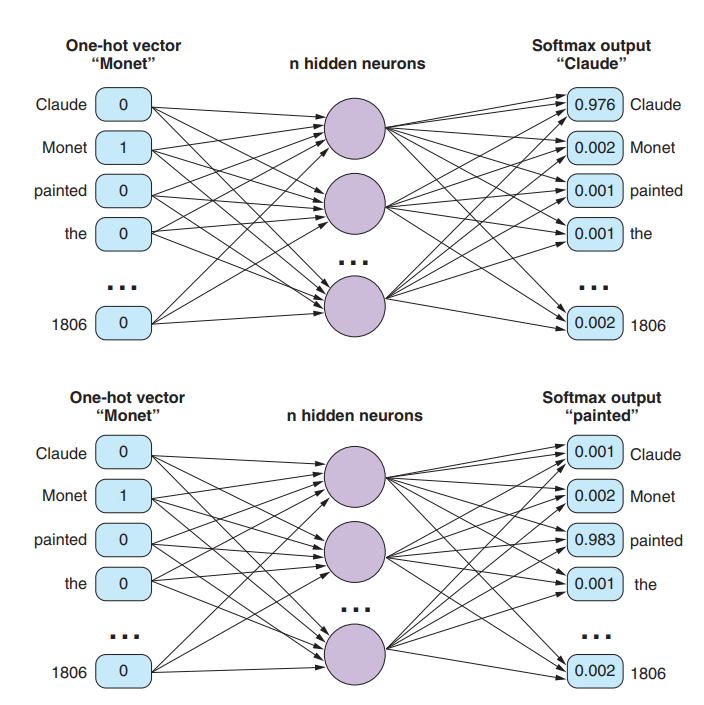
* Đây là phương pháp dự đoán các từ xung quanh dựa trên từ trung tâm
* Ví dụ: Claude Monet painted The Grand Canal of venice in 1908
* Trong câu về Monet trên, từ painted là đầu vào. Đầu ra sẽ đi dự đoán các từ như “Claude”, “Monet”, “The”, and “Grand”.
  + 1. Softmax
* **Softmax** thường được sử dụng trong mô hình **Skip-Gram** để dự đoán các từ ngữ cảnh (context words) từ một từ mục tiêu (target word)
* Đầu ra của các vector thường không nằm trong khoảng 0-1 và tổng không bằng 1 nên sử dụng softmax để chuẩn hóa về dạng xác suất
* Hàm softmax sẽ nén kết quả đầu ra giữa 0 và 1 và tổng của tất cả các đầu ra sẽ
* luôn cộng lại bằng 1. Theo cách đó, kết quả của lớp đầu ra có hàm softmax có thể được coi là xác suất.
* Đối với mỗi nút đầu ra trong số K, giá trị đầu ra softmax có thể được tính bằng hàm mũ chuẩn hóa:
* A mathematical equation with numbers and symbols

  Description automatically generated
* Ví dụ: nếu vector đầu vào của bạn như này A black and white image of a square with numbers

  Description automatically generated

Sau khi áp dụng hàm softmax thì kết quả sẽ thu được là: A math equation with numbers

Description automatically generated

* Các bước mà Word2Vec sử dụng skip-gram để dưa ra 1 từ đầu ra là:
* Biểu thị các từ bằng 1 vector one-hot
* Sử dụng mô hình huấn luyện neutral network
* Đưa ra trọng số đại diện cho mỗi từ
* Ví dụ: Trong ví dụ dưới đây thì đầu vào là “Monet” và đầu ra mong đợi là “painted” hoặc “Claude”. Tùy thuộc vào hoàn cảnh và mô hình huấn luyện sẽ đưa ra đầu ra phù hợp
* 
* Các hoạt động của skip-gram:
* Bước 1: chọn cửa số ( window-size )
* Bước 2: Tạo one-hot-endcoding cho từng từ
* Bước 3: Huấn luyện dữ liệu
* Bước 4: Đưa ra từ
* Xét ví dụ: Claude Monet painted The Grand Canal of venice Đầu vào là Monet

B1: chọn window-size = 1

Bước 2:

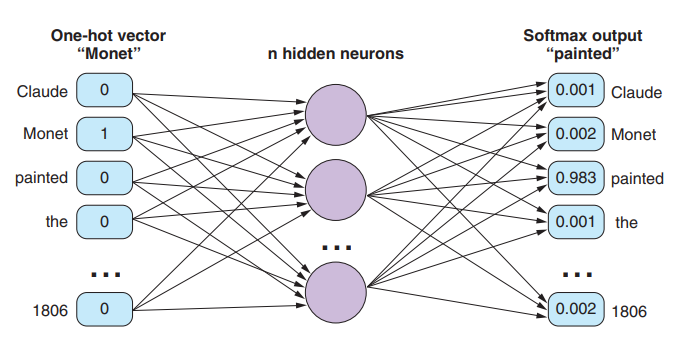
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Claude | 1 |  |  |  |  |  |  |  |
| Monet |  | 1 |  |  |  |  |  |  |
| painted |  |  | 1 |  |  |  |  |  |
| The |  |  |  | 1 |  |  |  |  |
| Grand |  |  |  |  | 1 |  |  |  |
| Canal |  |  |  |  |  | 1 |  |  |
| of |  |  |  |  |  |  | 1 |  |
| venice |  |  |  |  |  |  |  | 1 |

Bước 3: Do với window-size =1 nên khi có input word thì ra chỉ lấy trước và lấy sau 1 từ là từ target

|  |  |
| --- | --- |
| Input word | Target |
| Monet | Claude |
|  | painted |
| painted | Monet |
|  | The |
| The | painted |
|  | Grand |

Bước 4: Huấn luyện và đưa ra kết quả

Ví dụ trên chúng ta lấy từ Monet và mong muốn đầu ra là painted:



* Trọng số của một lớp ẩn trong mạng nơ-ron thường được biểu diễn dưới dạng ma trận: một cột cho mỗi nơ-ron đầu vào, một hàng cho mỗi nơ-ron đầu ra. Vì vậy, nếu bạn nhân (tích vô hướng) một vectơ hàng one-hot-coding đại diện cho 1 từ với ma trận trọng số đã được đào tạo, bạn sẽ có được một vectơ biểu diễn trọng số cho từ mà bạn quan tâm.
  1. Phương pháp CONTINUOUS BAG-OF-WORDS APPROACH
* Trái với skip-gram thì phương pháp này là dự đoán từ trung tâm dựa trên các từ xung quanh
* Ví dụ: input: Con mèo …. Trên ghế sofa
* Output: … có thể là “nằm”, “ngồi”, “ăn”, “ngủ”,…
* Cách hoạt động
* Tương tự như skip-gram
* B1:Chọn window-sized
* B2: one-hot-endcoding
* B3: Chuẩn bị dữ liệu
* B4: huyến luyện và đưa ra kết quả
* Ví dụ: Claude Monet painted The Grand Canal of venice in 1908

B1: chọn window – size = 2

B2:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Claude | 1 |  |  |  |  |  |  |  |
| Monet |  | 1 |  |  |  |  |  |  |
| painted |  |  | 1 |  |  |  |  |  |
| The |  |  |  | 1 |  |  |  |  |
| Grand |  |  |  |  | 1 |  |  |  |
| Canal |  |  |  |  |  | 1 |  |  |
| of |  |  |  |  |  |  | 1 |  |
| venice |  |  |  |  |  |  |  | 1 |

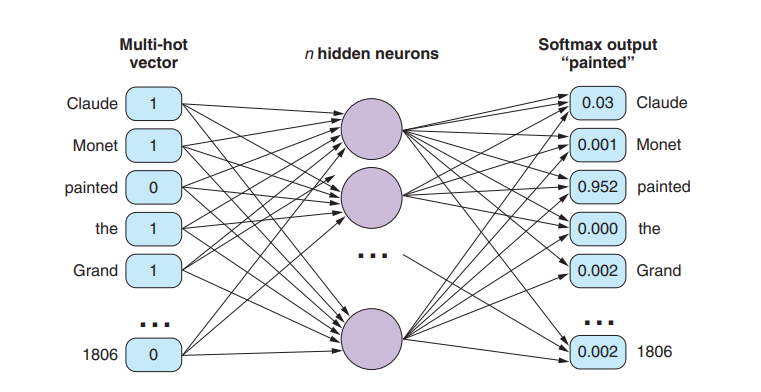
B3:

|  |  |
| --- | --- |
| Context word | Target |
| Claude Monet, The Grand | painted |
| Monet painted, Grand Canal | The |
| painted The, Canal of | Grand |

B4:

Ví dụ: input: Claude Monet, The Grand

và output là : painted



1. Các sử dụng Word2Vec vào dự án