Distributed Representations of Words and Phrases

and their Compositionality

# introduction

distributed trình bày trong không gian vector giúp học thuật toán để đạt được hiệu xuất tốt hơn trong các tác vụ xữ lý ngôn ngữ tự nhiên bằng cách nhóm các từ tương tự với nhau. Một trong những trình bay ngôn từ vựng sớm nhát trờ về 1986 của Rumelhart, Hinton và Williams. Ý tưởng này đã được áp dụng cho mô hình hóa thống kê ngôn ngữ với thành công đáng kể. công việc tiếp theo bao gồm áp dụng tự động nhận dạng âm thanh và dịch máy và các nhiệm vụ NLP

gần đây, Mikolov giới thiệu Skip-gram model, một phương pháp hiệu quả chất lượng cho việc trình bày vector của word từ lượng lớn dữ liệu chữ không có cấu trúc. Không giống như hầu hết các mạng neuron trước đây để hoặc word vector, huấn luyện Skip-gram model không bao gồm các phép nhân ma trận dày đặc. điều này thật sự hiệu quả: một máy tính hiện thực có thể huấn luyện trên 100 tỉ word một ngày.

Trình bày word đã được tính toán sử dụng mạng neuron là rất thú vị vì các vector đã được học chính xác được mã hóa rõ ràng nhiều quy tắc ngôn ngữ và các mẫu. một vài cái thật sự bất ngờ, nhiều trong số chúng có thể trình bày dịch một cách tuyến tính. Ví dụ, kết quả của phép tính vector (“Madrid”)-(“Spain”)+(“France”) thì rất gần với vec(“Paris”) hơn tất cả các từ khác

Hình 1. Skip-gram model architecture. Huấn luyện đối tượng để học word vector trình bày là tốt để tiên đoán các từ gần kề

Trong văn bản này chúng ta sẽ trình bày một số mở rộng của mô hình skip-gram . chúng ta cho thấy mẫu cua 3nhu7ng4 từ thường xuyen suốt quá trình huấn luyện trong một …, và cải thiện độ chính xác của việc trình bày ít hơn những từ thường xuyên, thêm vào đó, chúng ta sẽ trình bày một biến thể của NCE cho quá trình huấn luyện mô hình Skip-gram kết quả sẽ là huấn luyện nhanh hơn và tốt hơn việc trình bày word vector, được so sánh tới lớp phức tập hơn …

…

…

…

# The Skip-gram model