Efficient Estimation of Word Representations in Vector Space

O artigo visa propor dois novos tipos de arquitetura para computar representações em vetores contínuos para palavras em grandes quantidades de dados, capaz de utilizar e treinar em uma quantidade de até 1.6 bilhões de palavras, enquanto muitos trabalhos ainda utilizavam palavras como unidades totalmente atômicas, perdendo totalmente a noção de similaridade entre as palavras, para aproveitar as vantagens, como simplicidade, robustez e modelos mais simples através deste conjunto. Porém tais técnicas também apresentam suas limitações, como a utilização de uma grande quantidade de dados que podem ser aproveitados, visto que tais técnicas não passavam de alguns milhões de palavras.

Representação de palavras por vetores contínuos não era algo novo, porém utilizavam arquiteturas mais complexas, este trabalho tenta simplificar estas arquiteturas, alcançando uma representação mais usual para grandes conjuntos de dados, mantendo a eficiência.

O modelo apresentado foca em uma representação de palavras distribuída é apreendida por uma rede neural, tentando aprimorar a acurácia final enquanto minimiza a complexidade total do modelo, que em geral, para os propostos gira em torno de E x T x Q, sendo E o número de épocas de treinamento, T o número de palavras e Q que dependente da arquitetura, se utilizada uma rede neural Feedforward ou recorrente.

O modelo Feedforward consiste em uma entrada, projeção, camadas escondidas e camada de saída. Como entrada tem-se N palavras com codificação utilizando-se 1-para-V, sendo V o tamanho do vocabulário. Esta camada de entrada é projetada em uma camada P com dimensão N x D, utilizando-se uma matriz de projeção. A parte mais complexa deste caminho é entre a projeção e as camadas escondidas densas que é utilizada posteriormente para computar a probabilidade da distribuição das palavras no vocabulário. Alguns mecanismos para melhorar a arquitetura também foram utilizados, como por exemplo o softmax hierárquico.

Outro modelo apresentado é dado pela utilização de uma rede neural recorrente, que pode solucionar alguns problemas limitações do modelo anterior, como exemplo a necessidade de especificar o tamanho do contexto (N), além de poder representar modelos mais complexos. Esta arquitetura também não contém a camada de projeção. Aproveita-se também da memória que estes tipos de redes formam para lembrar informações passadas.

Em todos os modelos apresentados e testados implementou-se modelos de paralelismo, utilizando-se do framework DisBelief, capaz de executar múltiplas instâncias do mesmo modelo em paralelo e sincronizar a atualização dos gradientes. Também vale ressaltar a utilização de um learning rate adaptativo, Adagrad.

As duas arquiteturas propostas tentam então minimizar a complexidade computacional, explorando modelos mais simples como princípio. A primeira, chamada de Continuous Bag-of-Words Model, remove-se a camada não linear e compartilha a camada de projeção entre todas as palavras, em relação com o modelo Feedforward apresentado anteriormente. O segundo modelo, chama Continuous Skip-gram Model similar ao primeiro, porém ao invés de predizer as palavras considerando o contexto, maximizando a classificação em relação às outras palavras na mesma sentença.

Como resultado, inicialmente verificou-se a existência de outros tipos de similaridade, geralmente não utilizados em outros artigos. Ótimos resultados foram obtidos em referência a descrição semântica encontrada pelo algoritmo e suas acurácias, atingindo valores do estado da arte no conjunto de dados do desafio da Microsoft, como exemplo. Tudo isto utilizando uma grande quantidade de dados realizando o processamento em paralelo.