Resumo – A Neural Probabilistic Language Model

O artigo propõe um novo método de modelagem de linguagem que enfrenta a maldição da dimensionalidade encontrada em métodos tradicionais, por meio do aprendizado de uma representação distribuída para palavras.

A maldição da dimensionalidade ocorre em modelos com muitas variáveis discretas, como palavras em uma sentença. A generalização desses modelos é difícil pois qualquer mudança nessas variáveis causam um impacto grande no valor da função alvo, e quando o número de variáveis discretas é muito grande, ocorrências mais comuns estão distantes quase ao máximo umas das outras por métrica hamming. Em modelos com variáveis contínuas, por outro lado, a generalização é mais fácil porque a função a ser aprendida possui suavidades locais.

Um modelo estatístico de linguagem pode ser representado pela probabilidade condicional da próxima palavra dadas todas as palavras anteriores, o que corresponde ao produto das probabilidades condicionais de cada palavra anterior, dada a antecessora. Este modelo tem aplicações em linguagem natural como reconhecimento de fala, tradução e recuperação de informações. Melhorias no modelo podem impactar significativamente estas aplicações. Pode-se aproveitar a informação da ordem das palavras, e o fato de que palavras mais próximas são estatisticamente mais dependentes. Assim, modelos n-gram tomam probabilidades condicionais em combinações de n–1 palavras anteriores, o contexto. Quando uma sequência é desconhecida, pode-se usar um contexto menor.

A proposta do trabalho é associar cada palavra do vocabulário com um vetor de características distribuído, ou seja, um vetor de valores reais em um espaço m dimensional; expressar a função de probabilidade conjunta das sequências de palavras como vetores destas palavras em sequência; e aprender de forma simultânea os vetores de características de palavras e os parâmetros da função de probabilidades.

Esses parâmetros podem ser aprendidos pela otimização de uma função alvo de forma a maximizar a log-likelihood do dado de treino. Palavras similares semanticamente ou sintaticamente tendem a possuir vetores similares. A função de probabilidades é uma função suave, que é modificada pouco a pouco conforme pequenas mudanças nas características.

O modelo proposto é composto por uma função de cada palavra do vocabulário para o vetor de características, e uma função de probabilidade formada por 2 camadas de combinação linear (a segunda sem bias) interpoladas por uma função de ativação tangente hiperbólica, e uma softmax. Opcionalmente, usa-se uma conexão direta dos vetores acumulando com a saída da segunda camada, antes da softmax. A função de perda é a média aritmética do log do softmax para cada palavra da sequência, opcionalmente adicionando um termo de regularização. Otimiza-se a função por gradiente estocástico ascendente, com uma taxa fixa de aprendizagem de 10^{-3}. Experimentou-se a mistura deste modelo com modelos estatísticos tradicionais, e computação paralela.

Foram realizados experimentos comparativos sobre o corpus Brown com um fluxo de 1.181.041 palavras, das quais 800 mil usadas para treino, 200 mil para validação e as restantes para teste. O vocabulário contendo 47.528 palavras distintas, retirando as raras chegou-se a 16.383. Também empregados os textos da Associated Press News de 1995 e 1996, com 14 milhões de palavras, 148 mil distintas, reduzidas para 18 mil. Os resultados foram comparados com diversos modelos de n-grams. A métrica empregada foi a perplexidade, formada pela média geométrica das probabilidades condicionais obtidas.

Foram obtidas reduções na perplexidade de 24% na base Brown e 8% na AP News, em relação ao melhor modelo n-gram. Interpolação com modelos n-gram traz melhores resultados, sugerindo que os modelos erram em locais diferentes. Parece que se o uso de conexões diretas piora o desempenho, mas de forma não conclusiva.