Nome: **Rafael Claro Ito (R.A.: 118430)**

Principais destaques do artigo: **A neural probabilistic language model**

O principal problema em modelagem de linguagem estatística está na "maldição da dimensionalidade", expressão criada por Richard E. Bellman em seu trabalho sobre programação dinâmica (1957). Isso ocorre pois a função de probabilidade conjunta de um número grande de variáveis discretas resulta em um crescimento exponencial de parâmetros livres. A principal proposta deste artigo é resolver esse problema usando um modelo de rede neural que aprenda a representação distribuída das palavras, de modo que cada frase de treinamento representa um número exponencial de frases semanticamente vizinhas para o modelo. Desta forma o modelo aprende simultaneamente a representação distribuída das palavras e a função de probabilidade para sequência de palavras. Assim, o número de parâmetros livres cresce linearmente com o tamanho do vocabulário e também com o valor de *n*.

A rede neural proposta é resumidamente composta de:

- camada de entrada composta de (*n* - 1) palavras anteriores;
- camada que mapeia essas (*n* - 1) palavras em um vetor real de dimensão *m*;
- camada escondida com função de ativação sendo a tangente hiperbólica;
- camada de saída com função *softmax* que indica a probabilidade das palavras no vocabulário serem a próxima da sequência.

Neste esquema proposto o número de atributos (foram testados *m*=30, 60 e 100 neste artigo) é bem menor do que o tamanho do vocabulário (geralmente entre 16 e 18 mil palavras). Além da rede neural também foi explorado uma mistura de modelos, combinando a predição fornecida pela rede neural com a do modelo trigram, apresentando uma performance melhor.

Talvez a principal contrapartida deste modelo proposto seja o aumento de recurso computacional necessário quando comparado com os tradicionais modelos n-grams. Para contorná-lo, explorou-se paralelização de duas formas: paralelização com relação aos dados (utilizando processadores de memória compartilhada) e paralelização com relação aos parâmetros (utilizando clusters Linux), ambos detalhados na seção 3 do artigo.

Em seguida, comparou-se os resultados obtidos com este modelo de rede neural com outros modelos (*trigram* suavizado/interpolado, *back-off n-gram* e *class-based n-gram*). Para isto, utilizou-se dois *corpus*, *Brown* (conjunto de textos em inglês) e *AP News* (notícias entre 1995 e 1996), usando como métrica para avaliação de performance o cálculo da perplexidade. Os resultados obtidos foram significativos quando comparado aos modelos estados da arte que apresentavam as melhores métricas. Obteve-se uma melhora de 24% na perplexidade utilizando o *corpus Brown* e 8% utilizando o *AP News*. Indicativos sugerem que este modelo é capaz de se beneficiar de contextos maiores do que os tradicionais modelos n-gram.

Também é mostrado como uma variação da rede neural em questão pode ser interpretada como um modelo de minimização de energia proposto por Hinton (*products of experts*). A habilidade desta nova arquitetura em lidar com palavras fora do vocabulário e ainda atribuí-las um valor de probabilidade é apontada como uma vantagem em relação a arquitetura anterior.

Por fim, são apontadas diretrizes para trabalhos futuros: interpretação da representação das palavras, inserção de informação a priori, decomposição da rede em sub-redes, representação da probabilidade condicional em uma estrutura de árvore, técnicas para melhora do tempo de treinamento, entre outras.