Resumo – Scaling Laws for Neural Language Models

O artigo investiga possíveis fatores que determinem o desempenho de modelos de linguagem, com foco na arquitetura Transformers, por meio da função de perda entropia cruzada. São investigados fatores como a arquitetura do modelo, o poder computacional empregado e a quantidade de dado disponível para treino.

Cita estudos anteriores que relacionavam desempenho com o tamanho do modelo e da base de treino já em problemas de estimativa de densidade e modelos *random forest*, bem como achados em alocação ótima de poder computacional. Também estudos que acharam essas relações em vários tipos de corpus distintos, que relacionaram profundidade e tamanho das camadas neurais, e trabalhos similares no campo de processamento de imagens.

Para o estudo, foi empregado principalmente o corpus WebText2, com tamanho de 10^{10} palavras, vocabulário de cerca de 50 mil tokens, sentenças de 1024 tokens, 512 sentenças por lote. Treinamento realizado com *decoder Transformers*, sem bias, otimizador Adam, 250 mil passos, variadas configurações de taxas de aprendizado. Verificou-se então o desempenho conforme o tamanho do modelo, tamanho do *dataset*, formato dos modelos, menores tamanhos de sentenças, tamanhos de lotes, diferentes corpora e modelos. Em modelos maiores utilizou-se o Adafactor por conta de restrições de memória.

Seguem alguns resultados encontrados. O desempenho do modelo (L) depende intensamente do número de parâmetros do modelo (N), excetuando os de *embedding*; depende ainda do número de tokens do *dataset* (D), e do poder computacional empregado (C), e pouco do formato da rede. Há uma relação de lei de potência entre L e N, D, C quando um desses valores não é restrito pelos outros dois, em mais de seis ordens de magnitude. N e D devem crescer juntos para melhor desempenho, em uma relação aproximada de 8 para 5. Curvas de treino seguem leis de potência previsíveis, ou seja, extrapolando a partir dos primeiros passos conseguimos predizer aproximadamente a perda alcançada ao longo do treinamento. Ao avaliar modelos em texto com distribuição diferente do treino, há uma penalidade constante sobre o desempenho no treinamento, mas a melhoria no treino é acompanhada por uma melhoria correspondente no teste. Modelos maiores alcançam melhor performance com menos passos de otimização e menos dados de treino.

Ao ampliar o tamanho do modelo, dados e computação, é esperado que futuros modelos tenham melhor desempenho e sejam mais eficientes em relação a dados para treino do que os modelos atuais. Maiores modelos deverão ser mais importantes que mais dados. Os autores esperam que as relações obtidas deixem de ser meras observações para produzir um pacote preditivo, de forma similar ao que ocorre com a lei dos gases ideais, de onde obtemos relações entre propriedades macroscópicas de um gás de forma universal e independente da maior parte dos detalhes em seus constituintes microscópicos.

Ao sugerir trabalhos futuros, admitem que ainda não sabem quais de seus resultados dependem da estrutura dos dados de linguagem natural, que é universal. E advogam pela importância da investigação se a redução continuada da perda nos modelos se traduzirá em maior efetividade das tarefas relevantes de linguagem natural.