Nome: Rafael Claro Ito (R.A.: 118430)
Resumo do artigo: Effective Approaches to Attention-based Neural Machine Translation

Apesar de mecanismos de atenção estarem sendo usados para tarefas de tradução de texto usando redes neurais, denominada NMT (*neural machine translation*), quando o artigo foi publicado em 2015, pouco foi explorado sobre diferentes arquiteturas. Foi nesse contexto que a pesquisa descrita no artigo foi feita. Aqui, são apresentadas duas formas de mecanismos de atenção, uma global e outra local.

Inicialmente é descrita de forma sucinta o funcionamento de uma NMT, que nada mais é do que uma rede neural que modela a probabilidade condicional *p(y|x)*, onde *x* é a sequência de palavras de entrada e *y* é sua tradução, ou seja, sequência de saída. Uma forma básica de NMT consiste de dois componentes: um *encoder* que computa a representação $\boldsymbol{s}$ de cada sentença de entrada, e um *decoder* que gera uma palavra de saída por vez. Também são citados usos de redes recorrentes (RNN) e LSTM em outros trabalhos da literatura, sendo que para os modelos propostos foi usada uma arquitetura LSTM empilhada.

Em seguida, o artigo descreve as arquiteturas da abordagem global e local. Para ambos modelos o procedimento é o mesmo: para o passo $\boldsymbol{t}$, usar os estados atuais de saída $\boldsymbol{h_t}$ e de entrada $\bar{\boldsymbol{h}}_{\boldsymbol{s}}$ para calcular o vetor de contexto $\boldsymbol{c_t}$, que é usado para a predição da palavra de saída $\boldsymbol{y_t}$. Entretanto, algumas diferenças são levadas em conta para cada modelo na hora de computar $\boldsymbol{c_t}$.

No modelo global todas palavras da sequência de entrada são usadas para se calcular a atenção. Para o cálculo do vetor de contexto $\boldsymbol{c_t}$ é utilizado o vetor de alinhamento $\boldsymbol{a_t}$, calculado a partir de funções *scores*, que podem ser três tipos diferentes de operações com os vetores $\boldsymbol{h_t}$ e $\bar{\boldsymbol{h}}_{\boldsymbol{s}}$ (*dot*, *general* e *concat*).

O modelo local tenta suprir a desvantagem que o modelo global apresenta ao traduzir sequências longas. Para este modelo, apenas uma janela de palavras da sequência de entrada é levada em conta para se calcular a atenção. Aqui um novo vetor de posicionamento $\boldsymbol{p_t}$ é levado em conta para o cálculo de $\boldsymbol{c_t}$. São utilizadas duas variantes do modelo local para o cálculo de $\boldsymbol{p_t}$, uma de alinhamento monotômico (***local-m***) e uma de alinhamento preditivo (***local-p***).

Interessantemente, o artigo também propõe uma abordagem *input-feeding*, que é uma espécie de *coverage set* para rastrear quais palavras foram traduzidas. Nessa abordagem, vetores atencionais $\tilde{\boldsymbol{h}}_{\boldsymbol{t}}$ são concatenados com entradas para o próximo passo. Os objetivos desta implementação são prover ao modelo informações de escolhas de alinhamento passadas e também criar camadas profundas horizontais e verticais.

Diversas comparações com modelos de outros autores e artigos são feitas. Ambos modelos de atenção propostos podem ser interpretados a partir de modelos já encontrados na literatura: o modelo de atenção global pode ser visto como um modelo parecido com o de (Bahdanau et al., 2015), porém com uma arquitetura mais simples, enquanto que o modelo de atenção local pode ser visto como uma mistura entre os modelos de atenção soft e hard propostos por (Xu et al., 2015), com a vantagem de o modelo proposto no artigo ser diferenciável quase que em todos os pontos.

Em seguida, são apresentadas informações do treinamento e também os resultados obtidos. O modelo foi treinado no dataset WMT'14 e usou o dataset newstest2013 para validação e seleção de hiperparâmetros. Para isso, limitou-se o vocabulário para as 50 mil palavras mais frequentes e o restante foi convertido para o token <unk>. Na tabela 1 do artigo é possível ver a colaboração de cada técnica (*reverse*, *dropout*, atenção local/global, *feed input*, *unk replace*) para as métricas de perplexidade e BLEU (*bilingual evaluation understudy*). Aqui, duas técnicas me chamaram a atenção: a *reverse*, que consiste em treinar o modelo com a ordem das palavras de entrada invertida, e a *unkown replace*, que treina o modelo para emitir a posição na sequência de entrada para cada palavra OOV (*out-of-vocabulary*) encontrada na sequência de saída, permitindo em uma fase de pós-processamento traduzir cada OOV usando um dicionário.

Os testes foram feitos no dataset WMT, a partir de tarefas de tradução de inglês para alemão e vice-versa. Para a tarefa de tradução de inglês para alemão, o modelo de atenção local demonstrou um ganho de 5.0 BLEU com relação a sistemas sem mecanismos de atenção, mas com técnicas como dropout. Já o *ensemble* de 8 modelos diferentes atingiu o status de SOTA (*state-of-the-art*) no dataset WMT'15, atingindo 25.9 BLEU, que corresponde a um aumento de 1.0 BLEU acima do melhor sistema até então. Um outro resultado interessante é a habilidade dos modelos atencionais em lidar com sentenças longas.