Prova 02 -Processamento de Linguagem Natural

Aluno: Leonam Rezende Soares de Miranda
matrícula: 2020681492

1)

a) Overfitting não é desejável, pois apesar do modelo se ajustar muito próximo ou exatamente ao conjunto de treinamento, não consegue ajustar a dados adicionais ou prever observações futuras de forma confiável. Para corrigir essa situação podem ser utilizados mais dados para treinamento ou utilizar a regularização L2.

b) A ideia do skip-gram é maximizar a similaridade entre os vetores de palavras que aparecem juntas (no contexto umas das outras) no texto e minimizar a similaridade de palavras que não aparecem. No entanto, computar isso pode ser muito lento, porque existem muitos contextos. A amostragem negativa é uma das maneiras de resolver esse problema. A ideia é se uma palavra aparece no contexto de outra, então seus vetores são mais próximos do que de várias outras palavras escolhidas aleatoriamente, ao invés de avaliar todas as outras palavras do corpus.

c) Utilizar taxas de aprendizagem menores, ou uma técnica de decaimento da taxa de aprendizagem durante o treinamento, pois ao utilizar taxas de aprendizagem altas durante o treinamento pode ocorrer “overshoot” da função de custo divergindo, ou ficar oscilando sobre um mínimo local sem convergir.
Usar uma inicialização de pesos adequada: como a inicialização de Xavier. Por si só, essa opção não garante que não ocorrerá a explosão do gradiente, divergindo a função de custo, mas torna a rede mais robusta para que não ocorra esse problema.

d) i - Modelos que usam vetores densos de palavras generalizam melhor para palavras novas do que aqueles que usam vetores esparsos.

e) Não irá, pois inicializações diferentes levam a resultados diferentes. Em geral, inicializar todos os pesos em zero resulta na falha da rede em quebrar a simetria, e a rede não é mais poderosa do que um classificador linear. A inicialização aleatória é usada para quebrar a simetria e garantir que diferentes neurônios possam aprender coisas diferentes. A inicialização ruim dos pesos, como valores grandes, pode levar ao problema desaparecimento / explosão de gradientes.

f) iii) Faz uso de estatísticas globais de co-ocorrência.

2) Nos modelo sequence-to-sequence o desempenho cai bastante ao trabalhar com sentenças longas. Transformers utilizam camadas de self-attention, permitindo que o modelo se concentre nas partes relevantes da entrada.

3)
Sendo o conjunto de treinamento formado por sequências e um vetor de suas respectivas entidades nomeadas:

- Atribuir para cada classe um número. Ex.: pessoas =1, localidades =2, etc.
- Utilizar um vetor de palavras pré-treinado.
- A partir do vetor de palavras, atribuir a cada palavra um número, a partir do dicionário word2index. Ex.: carro = 1093, Japão = 19744, etc. Assim cada sequência é transformada num vetor de números, onde cada número corresponde ao índice da palavra;
- Realizar padding, já que será utilizado **LSTM** que requer que todas as sequências de entrada possuam o mesmo tamanho;

A primeira camada do modelo obtém os vetores das palavras a partir de seus índices, a segunda camada é composta por LSTMs. As saídas dos LSTMs alimentam uma camada densa e a última camada possui função de ativação softmax para obter as probabilidades da sequência possuir cada uma das categorias (classes).