Deep Learning for Natural Language Processing

Leonardo Santos Miranda - Mestrado

1) Respostas curtas:

a) Suponha que você tenha uma rede neural que está operando em overfitting em relação aos dados de treinamento. Por que essa não é uma situação desejável? Descreva duas maneiras de corrigir esta situação.
R: Ter um modelo operando em overfitting significa que o modelo produz bons resultados em cima da base de dados em que foi treinado, mas é somente nesses dados que o resultado é de fato, bom. Em outras palavras, um modelo em overfitting não consegue se generalizar quando novos dados aparecem e o resultado não será bom, por isso esta não é uma situação desejável. Uma maneira de reduzir o overfitting (e a mais óbvia) é aumentar a quantidade de dados no treinamento. Uma outra maneira de reduzir o overfitting é aplicar técnicas de regularização sobre a rede neural.

b) Por que o skip-gram com amostragem negativa (negative sampling) é mais rápido de treinar do que o modelo skip-gram original (sem negative sampling)?
R: A amostragem negativa nos permite que o modelo seja treinado modificando apenas uma pequena porcentagem dos pesos, ao invés de todos os pesos para cada amostra de treino.

c) Suponha que a função de perda de uma arquitetura neuronal esteja sendo avaliada de maneira que a perda aumenta com o número de épocas. Sugira duas opções para contornar essa situação. Justifique as sugestões.
R: Uma das opções para contornar a esta situação é diminuir a taxa de aprendizado: pode ser que o modelo esteja se movendo muito na direção oposta ao gradiente, podendo se afastar do ponto mínimo da função de perda.

d) Vimos que vetores densos de palavra tem muitas vantagens sobre o uso de vetores esparsos, como one-hot-encoding. Qual das alternativas a seguir NÃO é uma vantagem que os vetores densos têm sobre os vetores esparsos?
i) Modelos que usam vetores densos de palavras generalizam melhor para palavras novas do que aqueles que usam vetores esparsos.
ii) Modelos que usam vetores densos de palavras generalizam melhor para palavras raras do que aqueles que usam vetores esparsos. - X
iii) Vetores densos de palavras codificam semelhanças entre as palavras, enquanto os vetores esparsos não.

R: A alternativa ii) não é uma vantagem.

e) Um modelo de rede neural multi-camadas treinado usando-se descida de gradiente no mesmo conjunto de dados, porém com inicializações diferentes para seus parâmetros, irá garantidamente chegar aos mesmos parâmetros ao fim do treinamento? Justifique sua resposta.
R: Não. A descida de gradiente é um algoritmo de otimização usado para minimizar funções movendo de forma interativa na direção da descida mais íngreme, conforme definido pelo negativo do gradiente. Em um modelo de rede neural MLP, se os parâmetros deste algoritmo forem alterados, como por exemplo, a taxa de aprendizagem, o gradiente irá se comportar de maneira diferente, podendo demorar ainda mais o tempo de convergência ou fazer com que o gradiente seja ultrapassado, resultando com parâmetros diferentes.

f) Qual das afirmações a seguir sobre Skip-gram está correta?
i) Ele prevê a palavra central a partir das palavras do contexto circundante.
ii) O vetor final para uma palavra é a média ou soma do vetor de entrada v e vetor de saída u correspondendo a essa palavra.
iii) Faz uso de estatísticas globais de co-ocorrência.
iv) Nenhuma dessas. - X
R: Afirmação correta é a iv) Nenhuma dessas.

2) Discuta os problemas/limitações inerentes a arquiteturas sequence-to-sequence e como a arquitetura Transformers lida com essas limitações.
R: Modelos sequence-to-sequence sofrem e perdem desempenho ao trabalhar com problemas que envolvem sentenças mais longas – é difícil para a rede neural memorizar contextos grandes. A arquitetura transformers, diferente da arquitetura sequence-to-sequence padrão, não utiliza LSTM e faz o uso somente de camadas de Self-Attention. No transformers, os vetores de embedding recebem uma sentença inteira como entrada, ao invés de somente uma palavra e desta maneira o desempenho não decai por conta de sentenças longas, como acontecia no modelo sequence-to-sequence. Além disso, por utilizar camadas de Self-Attention, por mais que sentenças longas sejam lidas, o modelo dará um foco maior em pequenas partes da sentenças, calculando qual a palavra mais importante para cada parte.

3) A tarefa de “Reconhecimento de Entidades Nomeadas” consiste em classificar nomes em categorias pré-estabelecidas, como Pessoas, Localidades etc. Proponha uma solução completa para o problema, incluindo o que espera-se de dados de treinamento. A solução deve usar uma arquitetura sequence-to-sequence.

R: Utilizando duas LSTM:
Uma LSTM com Attention para a codificação das sentenças de entrada, visto que o uso de Attention não tem perda de desempenho mesmo se a sentença for longa e caso seja, que foque as palavras mais importantes, facilitando o reconhecimento de entidades. A outra LSTM para a decodificação, filtrando as sentenças de forma que caso o seu vetor não tenha nenhuma entidade, a LSTM possa ignorá-la, diminuindo a perda de desempenho. O método de treinamento para a decodificação seria baseado no algoritmo de descida gradiente estocástica para minimizar o negativo do logaritmo likelihood. É esperado que os dados de treinamento sejam uma sequência de sentença natural.