IF867 - Introdução à Aprendizagem Profunda

February 13, 2025

2º Lista Teórica

Discente: - Gabriel D'assumpção de Carvalho - gdc2@cin.ufpe.br

Curso: - Ciências Atuariais - 7º Período

Vista Teórica

(1) As redes VGG são consideradas redes densas, onde foi proposto a estruturação de blocos, contendo um padrão de utilização de n filtros de kernel (3x3), seguido de max-pooling (2x2) e stride de 2, mas ao longo dos blocos o número de filtros vão duplicando, aumentando o custo computacional.

O uso de kernel (3x3) é vantajoso pois com 2 filtros vamos ter que estimar apenas 18 parâmetros que é menos custoso que os 25 e 49 parâmetros dos kernel (5x5) e (7x7) respectivamente. Além disso, quando utilizamos filtros menores conseguimos ter uma diversificação de características melhor, devido a diminuição mais lenta das dimensões da imagem e utilização de mais filtros para aprender características diferentes.

(2) Um dos tópicos mais abordados nos modelos de CNN era a magnitude de parâmetros que devem ser estimados pelas camadas totalmente conectadas. Por exemplo, se no último bloco da convolução temos um tensor (16 x 16 x 512) e na primeira camada conectada temos 4096 neurônios vai ser preciso estimar 131.073.000 parâmetros. Por conta disso foi proposto o modelo network in network (NiN) que implementa uma convolução de 1x1 pixels com isso é introduzido não linearidade a rede, e diminuição de canais enquanto a output continua com a mesma resolução, fazendo com que não seja necessário camadas totalmente conectadas e diminuindo o custo computacional por só estimar n parâmetros

O bloco inception proposta pela googlenet possui 4 ramificações, a primeira ramificação contendo apenas um kernel (1x1), a segunda e a terceira iniciam com um filtro (1x1) seguido de uma convolução (3x3), (5x5) e padding de 1 e 2 respectivamente e na última ramificação é feito um max-pooling

(3x3) o padding de 1, seguidos com uma convolução de (1x1), depois todos os output das convoluções são agrupados. Portanto, essa abordagem contribui para a detecção de diferentes características ao utilizar diferentes tamanhos de kernels e filtros, tudo isso sem ter um alto custo computacional devido as convoluções dos kernel (1x1) que diminui os canais de input.

③ O modelo ResNet vem com a proposta de blocos residuais, que é um bloco que tem duas convoluções e o seu output é somado com o input pela conexão residual, isso faz com que a rede só tenha que modelar $g(x)$, dado que $y = x + g(x)$. Com essa estrutura a ResNet consegue construir uma rede mais profunda sem que aconteça o degradation problem que é o esquecimento da informação e vanishing gradient.

Com base nisso o modelo da ResNet consegue ter estrutura mais profunda, geralmente tendo uma precisão melhor que o Inception, mais por ser mais profunda o custo computacional é maior.

As funções mais avançadas para ambos os modelos são a relu que ajuda no problema do vanishing gradient e como mostrado no paper da ResNet tem uma convergência mais rápida, também pode ser utilizado a LeakyReLU que mitiga o problema de neurônios mortos

(4) Quando temos poucas observações para treinar o nosso modelo podemos aplicar técnicas de data argumentation, pegando as observações e criando dados sintéticos com implementação de ruído, espelhamento, aumentando ou diminuindo a intensidade de um canal de cor, dá zoom na imagem e dependendo do problema podemos quebrar o tensor em x partes e misturar suas ordens.

A estrutura mais adequada para esse tipo de problema é a VGG, ViT, Inception ou ResNet, mas deve ser implementado técnicas para diminuir o risco de overfitting ao usar modelos profundos.

Para evitar o overfitting podemos utilizar técnicas como transfer learning, que é quando pegamos um modelo que já foi treinado e alteramos a sua saída para ser compatível com o problema a ser tratado, também pode usar o fine-tuning que é o retreinamento dessa rede adquirida, geralmente as primeiras camadas são congeladas por serem mapas de características simples, que detectam borda e texturas. A aplicação de dropout é amplamente utilizada no treinamento para introduzir ruído e evitar o overfitting.

(5)

1) O modelo 2 é o que possui a melhor generalização pois possui a loss de teste mais baixa, enquanto a sua acurácia de treinamento e teste são 92% e 90% respectivamente.

3) O modelo 1 possivelmente está com overfitting, pois a sua loss de treinamento foi de 0,2 e a do teste foi 0,6, mostrando que o modelo não consegue generalizar para novas observações. Se aplicarmos a exponencia na loss temos ($e^{-0,2} \approx 0,82$) e ($e^{-0,6} \approx 0,55$), portanto podemos que a probabilidade média dos acertos é 1,49 vezes maior, tendo um modelo quase aleatório para problemas binário, dado que só acerta em média 55% das novas observações

3) O modelo 2 apresenta fortes indícios de overfitting, pois a acurácia de treinamento é 100% e 68% e tenho uma probabilidade média de classe correta de aproximadamente 45% e 47%, indicando que o modelo não estima a classe correta com confiança.

4) O batch normalization é uma técnica usada para controlar a desestabilização da ativação de uma rede neural, transformando a sua média zero e variância 1. Essa manipulação ajuda a rede a não ter gradient explosion e vanish gradient, e também ajuda na velocidade de convergência do modelo, por permitir uma taxa de aprendizagem mais rápida.

Em modelos de CNN o batch-normalization vai atuar em cada canal de todo o mini-batch.

Como a normalização é feita usando o mini-batch, essa distribuição introduz ruído na rede, tornando o modelo mais generalizado, reduzindo o overfitting.

5) As CNN são modelos propostos para a área de visão computacional, onde vários casos são empregados em problemas de classificação com isso o modelo é projetado para estimar a probabilidade da observação pertencer a cada categoria, sendo escolhida a classe de maior probabilidade.

Quando utilizado a função de loss vamos ter valores de 0 e o modelo acertar a classe independente do nível de confiança e 1 caso a previsão esteja errada. Entretanto ao utilizar a cross-entropy loss passa a mostrar a probabilidade da classe certa. Por exemplo, use la classe correta é 1 e o modelo estimou uma probabilidade de 0,8, mesmo acertando a loss vai ser de aproximadamente 0,195. Por conta disso o modelo tem uma convergência e melhora na confia

vigência das suas estimativas.

b) A função de ativação sigmoid é considerada uma função não saturada, isso implica que a sua derivada em muitos pontos é igual a zero e o valor máximo é 0,25. Por conta disso a aplicação dessa função em redes profundas pode causando o vanish gradient e como mostrado no paper do ALEXNET a sua convergência é mais lenta.

c) Devido a uma possível overfitting do modelo um poderia utilizar técnica de data augmentation e dropout para adicionar ruído ao modelo, também pode ser proposto uma estrutura menos profunda.

O modelo 2 que apresentou underfitting pode ser revisto se a função de ativação não está causando vanish gradient, também é recomendado tomar mais learning que tiver uma base de dados pequena, podendo aumentar a profundidade da rede ou aumento uma quantidade de épocas. É importante implementar métodos como ADAM que ajusta a taxa de aprendizado ao decorrer do treinamento, começando com taxas mais elevadas.

O modelo 3 apresentou um bom ajuste para ambos os conjuntos de dados, mas ainda pode ser proposto implementações como dropout ou data augmentation para tornar o modelo mais estável ao longo do tempo.

(6) O TRANSFER LEARNING É UMA TÉCNICA QUE CONSISTE EM PEGAR O ALGORITMO E OS PARÂMETROS QUE SÃO FORAM CONSTRUÍDO E ESTIMADO EM UMA TAREFA MAIS POBRE E APLICAR O MODELO PARA OUTRA TAREFA EM ESPECÍFICO.

QUANDO ESTAMOS COM UM PROBLEMA DE CLASSIFICAÇÃO DE IMAGEM DEVEMOS ALTERAR A ÚLTIMA CAMADA DO ALGORITMO PARA TER UM SOFTMAX COM A MESMA SAÍDA DAS CLASSES DO NOSSO PROBLEMA. OUTRO PONTO A SER LEVADO EM CONTA É QUE MUITAS VEZES AS ÚLTIMAS CAMADAS APRENDE CARACTERÍSTICAS ESPECÍFICAS, SENDO RECOMENDADO RETREINAR A REDE MAIS CONGELANDO AS CAMADAS INICIAIS, PERCEBENDO COM QUE O TREINAMENTO FIQUE MAIS EFICIENTE.

QUANDO SE TEM MUITAS OBSERVAÇÕES DE TREINAMENTO É RECOMENDADO UTILIZAR O TRANSFER LEARNING APENAS PARA A INICIALIZAÇÃO DOS PESOS, SENDO FEITO UM FINE-TUNING, QUE É O RETREINO EM TODOS OS PARÂMETROS DO MODELO.

(7) O AUTOENCODERS (AES) É UM MODELO QUE REDUZ A DIMENSIONALIDADE DE INPUT PARA DETECTAR AS VARIÁVEIS LATENTES, REMOVENDO RUÍDO E REDUNDÂNCIAS DOS DADOS, ISSO AJUDA PARA MELHORAR A EFICIÊNCIA EM PROBLEMAS MODELADOS NÃO LINEARES, PORTANTO AES SÃO MODELOS DETERMINÍSTICOS E INCAPAZ DE GERAR NOVAS OBSERVAÇÕES.

O MODELO VARIATIONAL AUTOENCODERS (VAES) ADICIONA UM ALEATORIEDADE NA VARIÁVEL LATENTE, PARA QUE QUANDO ATA A SUA DECODIFICAÇÃO SEJA GERADA UMA NOVA OBSERVAÇÃO, TAMBÉM PODE FAZER PERTURBAÇÕES NA VARIÁVEL LATENTE E VERIFICAR QUAL ALTERAÇÃO É FEITA.

O MODELO GANS VAI ATUAR COM DUAS REDES PRONUNCIAS UMA VAI SER A GERADORA DE DADOS APARTIR DE UM RUÍDO E ESSA OBSERVAÇÃO DEVE ENGANAR A OUTRA REDE QUE VAI DISCRIMINAR A OBSERVAÇÃO DE PARA

COM A VERDADEIRA. EM RESUMO, O MODELO TENTA IDENTIFICAR A DISTRIBUIÇÃO DE UMA VARIÁVEL PARA NÃO SER POSSÍVEL IDENTIFICAR A ORIGEM VERDADEIRA DA AMOSTRA. PORÉM, ESSES MODELOS TEM UM CUSTO COMPUTACIONAL GRANDE E INSTABILIDADE NO TREINAMENTO.

(8) MODELOS DISCRIMINATIVOS SÃO MÉTODOS UTILIZADOS PARA DISTINGUIR QUAL O VALOR DE UMA VARIÁVEL Y PARTIR DE UM CONJUNTO DE VARIÁVEIS PREDITORAS X, SENDO UM TIPO DE MODELO UTILIZADO PARA REGRESSÃO E CLASSIFICAÇÃO. TENDO EM VISTA ISSO, O MODELO TENTA ESTIMAR A PROBABILIDADE DE Y APARTIR DE X, SENDO $P(Y|X)$.

OS MODELOS GENERATIVOS SÃO MÉTODOS QUE SÃO APLICADOS PARA GERAÇÃO DE NOVOS DADOS. O SEU ALGORITMO RECEBE COMO ENTRADA UMA VARIÁVEL X E O MODELO TENTA ESTIMAR A SUA PROBABILIDADE E GERAR UM $\hat{X}$, SENDO ESTIMANDO OS MODELOS GENERATIVOS VÃO MODELAR $P(X|\hat{X})$, AO MAXIMIZAR ESSA PROBABILIDADE VAMOS ESTAR APROXIMANDO $\hat{X}$ DOS DADOS REAIS.

AS VANTAGENS DOS MODELOS DISCRIMINATIVOS É QUE MUITOS POSSUEM UM PODER MAIS EXPLICATIVOS E CONSEGUEM SER AJUSTADO COM UMA BASE DE DADOS MENOR, PORÉM MUITOS MODELOS NECESSITAM QUE SUAS SUPOSIÇÕES SEJAM ATENDIDAS.

SÓ OS MODELOS GENERATIVOS É VANTAJOSO QUANDO NECESSITAMOS GERAR SIMULAÇÕES E NÃO PRECISA QUE SUPOSIÇÕES SEJAM ATENDIDAS, MAS A SUA GRANDE DESVANTAGEM É QUE É PRECISO DE UMA FORÇA COMPUTACIONAL MAIOR E UM BANCO DE DADOS GRANDE.

(9) A REDE U-NET FOI PROPOSTA PARA FAZER SEGMENTAÇÃO SEMÂNTICA, ONDE CADA PIXEL PERTENCE A UMA CLASSE.

POR CONTA DESSE TIPO DE SEGMENTAÇÃO AO APLICAR CNNs PROFUNDAS AS CONVOLUÇÕES E POOLING FAZEM COM QUE A CADA CAMADA O TENSOR DE ENTRADA DIMINUA, PERDENDO INFORMAÇÕES DO INPUT DEVIDO AO DOWNSAMPLING.

PREVINDO A ESSA PERDA DE INFORMAÇÃO, A U-NET ELABOROU A CONEXÃO DE SALTO QUE CONCATENA O TENSOR DA ETAPA DE DOWNSAMPLING QUE ESTÁ EM PARALELO AO TENSOR DE UPSAMPLING, FAZENDO COM QUE AS CAMADAS MAIS PROFUNDAS TENHAM INFORMAÇÃO COM UMA RESOLUÇÃO MAIS ALTA, AJUDANDO NA PRESERVAÇÃO DE BORDAS E CARACTERÍSTICAS ESPACIAIS.

(10)

1) SUPONDO QUE $W = 5 \times 5$, $K_1 = 3 \times 3$ E $K_2 = 2 \times 2$. O TAMANHO DA IMAGEM APÓS A PRIMEIRA CONVOLUÇÃO É DADO POR $(5 - 3 + 2 + 1) = (5 \times 5)$. COM A SEGUNDA CONVOLUÇÃO PASSA A SER $(5 - 2 + 2 + 1) = (6 \times 6)$.

2) A CONVOLUÇÃO É UMA OPERAÇÃO MATEMÁTICA QUE LEVA EM CONTA A APLICAÇÃO DE DUAS FUNÇÕES PASSANDO POR t PONTOS, PORÉM TENDO A PRIMEIRA FUNÇÃO INVERTIDA, SENDO DADO POR $(f * g)(t) = \sum_{i}^{n} f(i)\, g(t - i)$. MAS EM CNN AS CONVOLUÇÕES SÃO APLICADAS SEM A INVERSÃO, SENDO APLICADAS DIRETAMENTE, ISSO ACONTECE PORQUE OS KERNELS JÁ VEM TREINADOS A PRÉ-UMA ESTIMAR OS MELHORES VALORES INCREMENTE DA ROTAÇÃO FEITA NO FILTRO.