Resumo – Deep Residual Learning for Image Recognition

Redes neurais convolucionais profundas levaram a diversas descobertas para classificação de imagens. Evidências de trabalhos anteriores revelam que profundidade tem importância crucial, e os melhores resultados anteriores do ImageNet empregavam modelos de 16 a 30 camadas. Além dessa profundidade, ocorre degradação do desempenho do modelo.

O artigo introduz um elemento arquitetural, o framework de aprendizado residual profundo, em que a saída de uma camada é somada à saída de uma camada anterior, um atalho, ou uma função desta saída. Mostra que modelos com função residual otimizam mais facilmente que modelos sem função residual, cujo erro é muito maior em redes mais profundas. Mostra ainda que redes residuais facilmente ganham acurácia com profundidades bem maiores.

Resultados similares foram demonstrados sobre a base CIFAR-10, sugerindo que as dificuldades de otimização e os efeitos do método não se restringem a uma base particular.

A rede base, não residual, inspirada nas redes VGG, contém redes convolucionais com filtros 3x3, complexidade constante entre as camadas, balanceando o tamanho dos feature maps com o número de filtros e o stride, encerrando com uma camada de average pooling global e uma camada densa de 1000 neurônios e softmax, totalizando 18 ou 34 camadas com pesos. A rede residual é uma variação da rede base, acrescendo atalhos.

A rede usa as técnicas de pre-processamento e aumento de dados da AlexNet, assim como taxa de aprendizagem, mudanças, momento e decaimento de pesos. Batch normalização após cada convolução, sem pré-treino, SGD, mini-batch de 256, sem dropout.

Comparativamente, as redes residuais de 18 e 34 camadas apresentaram quedas mais rápidas de erro de validação, taxas top-1 menores. Redes residuais de 34 camadas com menor erro que redes de 18, enquanto em redes não residuais ocorre o contrário. O artigo alega que isso não deve ocorrer por conta de vanishing gradients, por conta de batch normalization.

O artigo testa ainda diferentes funções de atalho, além da função identidade, não observando diferenças relevantes. Redes mais profundas são testadas, com 56, 101 até 1202 camadas. Redes extremamente profundas apresentam resultados piores que rede de 110 camadas, argumentando-se que isso se deve a overfitting, sendo necessário maior base de treinamento ou uso maior de técnicas de regularização.

Apresentou uma rede residual de 152 camadas no torneio ILSVRC 2015 do ImageNet, sendo a rede mais profunda já apresentada até então, ainda assim com menor complexidade que as redes VGG. O ensemble apresentado alcançou erro top-5 de 3,57%, ganhando o primeiro lugar na competição de classificação, além de várias categorias de detecção, localização e segmentação em ILSVRC e COCO. Isso evidencia fortemente que o princípio de aprendizado residual é geral, e a expectativa é de que seja aplicável em outros problemas de visão computacional ou não.