Deep Residual Learning for Image Recognition

É apresentado um framework de aprendizado residual que é capaz de utilizar estruturas com camadas mais profundas e melhores resultados do que visto até o momento deste trabalho. Para isto, realizou-se múltiplos testes capazes de provar a capacidade de treinamento e otimização do modelo, tanto em acurácia quanto em performance. Sabe-se através da literatura, que redes neurais profundas quebram paradigmas em relação a classificação de imagens, redes cada vez mais profundas e com diferentes configurações.

Porém, uma das dúvidas existentes é a capacidade de treinamento e generalização do modelo em redes muito profundas e se basta realmente apenas acrescentar camadas em tais estruturas. Um dos problemas já estudados é o de desaparecimento do gradiente, porém este é muito estudado e combatido através de normalização de entradas e camadas. Outro problema que pode ser encontrado e é abordado neste trabalho é o de saturação da acurácia, que não é causada por sobreajuste, visto que várias técnicas também podem auxiliar neste problema. Isto indica que nem toda rede, principalmente as profundas são tão facilmente treináveis, e que redes mais rasas sem um devido tratamento, podem exibir resultados melhores que redes realmente profundas.

Para solucionar este problema, o trabalho sugere um framework de aprendizado residual, o que significa que, ao invés de empilhar sequencialmente cada camada, passa-se a ter uma arquitetura de empilhamento residual. Este resíduo tem este nome devido às informações das camadas anteriores que são passadas para as camadas à frente, ou seja, dado um mapa H(x) que deseja-se otimizar, têm-se camadas não lineares empilhadas de forma que cria-se um novo mapa F(x) := H(x) - x, assim o mapa original passa a ser F(x) + x. Acredita-se então que é mais fácil otimizar este novo mapa que o original.

Em casos extremos, onde o mapeamento de identidade é ótimo, é mais fácil fazer com que o mapa residual torna-se zero que ajustar o mapeamento de identidade por toda a pilha de camadas não lineares. Neste trabalho estudou-se conexões simples de identidade que também apresentam uma baixa complexidade computacional e pode continuar sendo utilizada junto com outras técnicas já implementadas, como SGD com backpropagation ou bibliotecas já muito difundidas, sem apresentar muita dificuldade. Outros trabalhos também utilizam esta ideia de vetores residuais e também apresentam versões mais efetivas que suas versões básicas, o que serve de motivação. A hipótese é que se múltiplas camadas não lineares podem aproximar funções assintoticamente, então pode-se também aproximar funções residuais, basta que a dimensão de x seja a mesma que a de F, porém caso contrário pode-se utilizar projeções lineares. Estas funções F são bem flexíveis, podendo conter múltiplas camadas. Neste trabalho utilizou-se funções de duas camadas.

Para provar sua eficácia, múltiplas redes e configurações foram propostas. Uma Plain network, que propõe uma rede profunda sem resíduo e uma com resíduo para comparações, chamada de rede residual. Também testou-se diferentes tipos dos chamados atalhos ou formas de calcular os resíduos. Como pré-processamento, utilizou-se a aumentação por escala, cores e normalização dos lotes.

Observou-se que, redes sem resíduos apresentam melhores resultados quando não são tão profundas, devido ao problema da saturação da acurácia, porém redes que utilizaram a estratégia de resíduos apresentam melhores resultados para redes mais profundas, dentro de um limite, que acredita-se está relacionado com o sobre aprendizado e tamanho do conjunto de dado de treinamento. Nos resultados obtidos verificou-se uma melhora de até 4.49% se comparados com redes do estado da arte em top-5, primeiro lugar na tarefa de classificação do ILSVRC 2015, uma melhora de 28% na detecção de objetos do conjunto de dados COCO, primeiro lugar nas tarefas de detecção de imagens e localização de imagens no ImageNet, detecção e segmentação para o COCO.