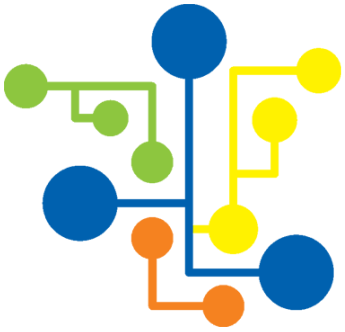

Data Science Academy

www.datascienceacademy.com.br

Deep Learning II

Dataset Augmentation

A melhor maneira de fazer um modelo de aprendizagem de máquina generalizar melhor é treiná-lo em mais dados. Na prática, a quantidade de dados que temos é limitada. Uma maneira de resolver esse problema é criar dados falsos (fake) e adicioná-los ao conjunto de treinamento. Para algumas tarefas de aprendizado de máquina, essa é uma atividade relativamente simples.

Essa abordagem é mais fácil em tarefas de classificação. Um classificador precisa receber uma entrada x com alta dimensionalidade e complicada e resumi-la com uma identidade de categoria única y. Isto significa que a tarefa principal que enfrenta um classificador é ser invariante para uma grande variedade de transformações. Podemos gerar novos pares (x, y) facilmente simplesmente transformando as entradas x em nosso conjunto de treinamento.

Mas esta abordagem não é tão simples para muitas outras tarefas. Por exemplo, é difícil administrar novos dados falsos para uma tarefa de estimativa de densidade, a menos que já resolvamos o problema de estimativa de densidade.

O aumento do conjunto de dados (Dataset Augmentation) tem sido uma técnica particularmente eficaz para um problema de classificação específica: reconhecimento de objetos. As imagens são de alta dimensão e incluem uma enorme variedade de fatores de variação, muitos dos quais podem ser facilmente simulados. As operações como a tradução das imagens de treinamento alguns pixels em cada direção geralmente podem melhorar a generalização, mesmo que o modelo já tenha sido projetado para ser invariante de tradução parcial, usando as técnicas de convolução e pooling descritas nos capítulos 7 e 8 do curso Deep Learning I. Muitas outras operações, como girar a imagem ou dimensionar a imagem, também provaram ser bastante eficientes.

Devemos ter cuidado para não aplicar transformações que alterem a classe correta. Por exemplo, as tarefas de reconhecimento de caracteres ópticos exigem o reconhecimento da diferença entre 'b' e 'd' e a diferença entre '6' e '9', de modo que flips horizontais e rotações 180◦ não são formas apropriadas de aumentar os conjuntos de dados para estes datasets. Há também transformações às quais gostaríamos que nossos classistas fossem invariantes, mas que não são fáceis de executar. Por exemplo, a rotação fora do plano não pode ser implementada como uma operação geométrica simples nos pixels de entrada. O aumento do conjunto de circuitos também é eficaz para tarefas de reconhecimento de fala.

Injetar ruído na entrada para uma rede neural também pode ser visto como uma forma de aumento de dados. Para muitas classificações e até mesmo algumas tarefas de regressão, a tarefa ainda deve ser possível, mesmo que seja adicionado pouco ruído aleatório à entrada. As redes neurais revelam-se não muito robustas ao ruído, no entanto. Uma maneira de melhorar a robustez das redes neurais é simplesmente treiná-las com o ruído aleatório aplicado às suas entradas. A injeção de ruído de entrada é parte de alguns algoritmos de aprendizado sem supervisão, como o Autoencoder Denoising (que estudaremos aqui no curso Deep Learning II). A injeção de ruído também funciona quando o ruído é aplicado às unidades ocultas, o que pode ser visto como um aumento de conjunto de dados em vários níveis de abstração. Esta abordagem pode ser altamente eficaz, desde que a magnitude do ruído seja cuidadosamente sintonizada. Dropout, uma poderosa estratégia de regularização, pode ser vista como um processo de construção de novas entradas, multiplicando por ruído.

Ao comparar resultados de benchmark de aprendizado de máquina, é importante levar em consideração o foco do aumento de conjunto de dados. Muitas vezes, os esquemas de levantamento de dados projetados à mão podem reduzir drasticamente o erro de generalização de uma técnica de aprendizado de máquina. Para comparar o desempenho de um algoritmo de aprendizagem de máquina com outro, é necessário realizar experimentos controlados. Ao comparar o algoritmo de aprendizagem da máquina A e o algoritmo de aprendizado da máquina B, é necessário garantir que ambos os algoritmos sejam avaliados usando os mesmos esquemas de aumento de conjunto de dados projetados à mão. Suponha que o algoritmo A seja executado de forma fraca sem o aumento do conjunto de dados e o algoritmo B funcione bem quando combinado com numerosas transformações sintéticas da entrada. Nesse caso, é provável que as transformações sintéticas tenham causado o desempenho melhorado, em vez do uso do algoritmo de aprendizado de máquinas B. Às vezes, decidir se um experimento foi devidamente controlado requer julgamento subjetivo.

Os algoritmos de aprendizagem que injetam ruído na entrada estão realizando uma forma de aumento de conjunto de dados. Normalmente, as operações que são geralmente aplicáveis (como a adição de ruído gaussiano à entrada) são consideradas parte do algoritmo de aprendizagem da máquina, enquanto as operações que são específicas para um domínio de aplicação (como

cortar aleatoriamente uma imagem) são consideradas como pré-etapas de processamento.

Referência:

Deep Learning Book – Capítulo 7 – Regularization for Deep Learning
http://www.deeplearningbook.org/