# Deep Learning - LeCun, Bengio e Hinton – Resumo

Deep-learning são métodos de machine learning (ML) que representam conhecimento em múltiplos níveis, desde o dado bruto, até níveis mais abstratos, por meio de módulos de transformação não-linear. Ao contrário de métodos tradicionais de ML, permitem o aprendizado de funções muito complexas, prescindindo assim de extração prévia, por especialistas humanos, de características dos dados voltadas para a discriminação da solução. Isso é conseguido automaticamente pelas camadas de sua arquitetura, fornecendo assim um procedimento de aprendizado de propósito geral.

O aprendizado supervisionado é realizado com um enorme conjunto de dados de entrada, cada um relacionado a um rótulo, que é a saída esperada da rotina de ML. Parte desses dados é empregada para treinamento, no qual cada dado de entrada é processado, e a saída é comparada ao rótulo fornecido, por meio de uma função alvo, que fornece o erro ou distância entre o produzido e o esperado. A máquina então modifica seus parâmetros internos, ou pesos, de forma a minimizar a função alvo. Processada uma quantidade suficiente de dados de entrada, os pesos estabilizarão com valores baixos da função alvo, e assim terá aprendido a produzir saídas adequadas. Parte dos dados de entrada é empregada como dados de teste, que servem para verificar a generalidade do modelo produzido, ou seja, sua habilidade de produzir respostas boas para dados de entrada que não foram vistos durante o treinamento.

A multiplicidade de camadas não lineares permite produzir funções intrincadas, sensíveis a pequenos detalhes relevantes e insensíveis a grandes variações irrelevantes para a resolução do problema. O método backpropagation é uma aplicação da regra da cadeia, na qual os gradientes são calculados e propagados desde a função alvo até as primeiras camadas, permitindo um fácil ajuste dos pesos. O uso de camadas neurais compostas de funções lineares com funções de ativação, entre as quais a ReLU, permitiu a produção de modelos muito bons para resolução de problemas. Com as GPUs veio o rápido processamento paralelo que permitiu o treinamento em um tempo relativamente curto.

Redes neurais convolucionais possuem arquitetura estruturada em vários estágios. Os primeiros são compostos por camadas convolucionais e de pooling. Camadas convolucionais aplicam filtros que permitem detectar padrões locais nos dados recebidos da camada anterior. As camadas de pooling combinam padrões semanticamente similares, reduzindo a massa de dados a ser enviada para chamadas posteriores. Os estágios finais são formados por camadas neurais convencionais. A combinação de arquiteturas convolucionais, GPUs e técnicas de regularização como drop-out (desligamento aleatório de neurônios) e técnicas de deformação de imagens permitiu alcançar o melhor nível de classificação de imagens, como nas competições ImageNet.

Redes neurais recorrentes (RNNs) se aplicaram a tarefas que envolvem entradas sequenciais, como fala, texto, vídeos. Processam uma sequência de entrada um elemento por vez, mantendo unidades intermediárias contendo informação histórica dos elementos já processados da sequência. Long short-term memory (LSTM) são evoluções de RNNs com estruturas especiais de memórias (gates), que conseguem reter melhor informações de sequências muito longas.

As tendências futuras apontadas são o desenvolvimento de técnicas de aprendizado não supervisionado no longo prazo e o desenvolvimento do aprendizado de reforço. Além disso, a área de entendimento de linguagem natural causando um grande impacto nos anos subsequentes, com sistemas usando RNNs para entender sentenças ou documentos inteiros se tornando muito melhores ao aprender estratégias de atenção seletiva a partes do texto a cada momento.