GUIA
PROGRAMA: ANÁLISE
NÚMERO DO DOCUMENTO: T147

NUCLEUS RESEARCH

GUIA APRENDIZAGEM PROFUNDA NA AWS

ANALISTAS

Daniel Elman

Nucleus Research, Inc.
100 State Street, Boston, MA, 02109
+1 (617) 720-2000
www.nucleusresearch.com

PONTOS IMPORTANTES

A aprendizagem profunda continua sendo um dos tópicos mais falados na área de inteligência artificial (IA) e está migrando rapidamente da academia e da teoria para cargas de trabalho operacionalizadas com valor agregado. Os avanços na infraestrutura de dados e no poder de computação, bem como as novas classes de redes neurais, ajudaram a viabilizar a aprendizagem profunda para as empresas modernas, enquanto isso ainda era praticamente uma ideia improvável a apenas um ano atrás. A parceria com fornecedores de nuvem, como a Amazon Web Services (AWS), e o aproveitamento de serviços de machine learning hospedados na nuvem, como o Amazon SageMaker, foram essenciais para empresas que procuram acelerar os projetos de aprendizagem profunda do conceito à produção. No passado, quando a infraestrutura não era tão avançada e os serviços de machine learning eram menos desenvolvidos ou ainda não estavam disponíveis, organizações sem os orçamentos e a experiência para desenvolver internamente do zero sistemas de IA eram deixadas de lado enquanto empresas maiores com fundos disponíveis e bancos de desenvolvedores pesquisavam e implementavam esses recursos de IA.

Para entender o estado da adoção e do uso da aprendizagem profunda atualmente, e como ela mudou desde o relatório do ano passado (Nucleus Research *s180 - Guia: TensorFlow na AWS* - Novembro de 2018), a Nucleus realizou entrevistas e analisou as experiências em 316 projetos únicos. Pela segunda vez em dois anos, o número de projetos de aprendizagem profunda em produção mais do que dobrou. Descobrimos que 96% da aprendizagem profunda está em execução em um ambiente de nuvem, com o TensorFlow sendo a estrutura mais popular, usado em 74% dos projetos de aprendizagem profunda. O PyTorch também foi usado em 43% dos projetos (observe que a maioria dos projetos aproveitam várias estruturas; MXNet, Keras e Caffe2 também apareceram, praticamente sempre em conjunto com o TensorFlow, o PyTorch ou ambos), um aumento significativo na adoção desde o ano anterior. Do total de 316 projetos, somente 9% foram criados com uma estrutura singular. Mais notadamente, dos projetos de aprendizagem profunda hospedados na nuvem, 89% estão em execução na AWS. Um dos principais motivadores disso é a amplitude de opções de estrutura na AWS aliada ao próprio investimento contínuo dela em serviços novos e existentes.

Também descobrimos que 85% das cargas de trabalho baseadas em nuvem do TensorFlow estão em execução na AWS, e 83% dos projetos PyTorch baseados em nuvem estão na AWS. No ano passado, cerca de um terço dos clientes entrevistados usavam ou estavam avaliando a possibilidade de usar o SageMaker, o serviço gerenciado da Amazon para criação, treinamento, implantação e orquestração de modelos de aprendizagem profunda em escala. Dos usuários entrevistados neste ano, 63% dos clientes da Amazon começaram a usar o SageMaker.

A SITUAÇÃO

O machine learning engloba a tecnologia em que um computador analisa os dados para "aprender" empiricamente sem envolvimento humano. A aprendizagem profunda é um subconjunto do machine learning. No machine learning, o computador recebe dados com um conjunto estabelecido de recursos para analisar, enquanto que na aprendizagem profunda o computador recebe dados não estruturados como texto, áudio ou vídeo e determina, por conta própria, quais recursos são relevantes para a análise. Em termos simples, o computador recebe pares e amostras de entrada e as saídas correspondentes e é capaz de fazer o processo inverso para descobrir quais operações são necessárias para transformar a entrada na saída.

Os modelos de aprendizagem profunda modernos exigem enormes quantidades de computação e armazenamento, impedindo que a maioria das organizações criem esses sistemas sozinhas. Assim, como descobrimos no decorrer desta pesquisa, as empresas buscam esmagadoramente a nuvem para projetos de aprendizagem profunda. Essa abordagem permite que as empresas comprem as quantidades de armazenamento de dados e poder de computação de que precisam para os projetos sem a necessidade de adquirir, configurar e manter a infraestrutura internamente, produzindo uma economia significativa no decorrer do tempo.

Com uma base de usuários diversificada e em rápido crescimento devido à grande repercussão e ao potencial da IA, o cenário da tecnologia muda rapidamente. Novas ferramentas e metodologias se tornam disponíveis constantemente. Portanto, permitir que os usuários desenvolvam na plataforma com a máxima flexibilidade é essencial para o sucesso de longo prazo no mercado de machine learning baseado na nuvem. Resumindo, as plataformas de nuvem precisam dar suporte a milhares de ferramentas e estruturas de desenvolvimento utilizadas hoje e amanhã, com a segurança obrigatória e a disponibilidade de aderir às normas de processamento e privacidade de dados.

Este é o terceiro ano consecutivo em que este estudo é realizado e o que vimos nesse período foram mudanças transformadoras nos recursos dos modelos, no poder de computação e nas ferramentas do desenvolvedor que permitem resultados novos e empolgantes. No primeiro ano foi difícil encontrar organizações que tinham ido além do desenvolvimento preliminar e dos projetos de prova de conceito com a aprendizagem profunda. Em 2018, vimos um forte progresso para essa finalidade com 14% dos projetos sendo classificados como em produção, processando dados dinâmicos. Este ano trouxe outro salto em direção a esse efeito com organizações variando de startups de 20 pessoas a empresas globais listadas na Fortune 100 implantando a aprendizagem profunda na produção, com 38% dos projetos em produção. Outros aspectos discutidos nas entrevistas incluem:

- As metas e motivações por trás dos projetos de aprendizagem profunda
- A estratégia de implantação e os benefícios associados
- As estruturas, os métodos e outras ferramentas de desenvolvimento em utilização
- Os pontos fortes e fracos relativos dos diferentes modelos e estruturas

- O número de pessoas envolvidas e as respectivas funções nas equipes de projeto

No total, a Nucleus realizou entrevistas detalhadas com 32 especialistas em aprendizagem profunda, muitos deles responsáveis por vários projetos simultaneamente, representando 316 projetos únicos.

APRENDIZAGEM PROFUNDA NA NUVEM

As cargas de trabalho de aprendizagem profunda em escala de produção envolvem o processamento de milhares ou milhões de dados de exemplo para treinar o modelo. Do ponto de vista computacional, isso é extremamente caro, especialmente para dados de entrada complexos como imagens ou vídeo, e a maioria das organizações não pode arcar com a criação e a manutenção de sistemas de computação de alta performance com CPUs ou GPUs paralelizadas para executar os cálculos. Como resultado, as organizações buscam a nuvem para acessar os recursos e a infraestrutura de que precisam. Este ano, descobrimos que 96% dos projetos de aprendizagem profunda estão sendo executados em um ambiente de nuvem. Isso reflete a descoberta do ano passado, mas o aumento de 177 projetos em 2018 para 316 em 2019 ainda demonstra o forte movimento do cliente para a nuvem em relação à aprendizagem profunda. Das cargas de trabalho em produção sobre dados dinâmicos, 98% estão na nuvem. Para organizações que não estão totalmente na nuvem, uma estratégia de implantação comum envolve o desenvolvimento de um modelo de pequena escala no hardware no local antes de migrar a produção para a nuvem.

BENEFÍCIOS DA NUVEM

Com 96% dos projetos de aprendizagem profunda em execução na nuvem, os clientes reconhecem claramente o valor da abordagem. Pedimos aos entrevistados para identificarem os principais benefícios da execução da aprendizagem profunda na nuvem. As respostas se agruparam em três temas principais:

- Redução de custos evitando despesas com hardware, pessoal e energia. Essa foi a resposta mais comum, citada por praticamente todos os entrevistados. A aprendizagem profunda exige enormes quantidades de computação. Criar e manter sistemas de hardware capazes de executar a aprendizagem profunda em escala requer profissionais de TI dedicados. Colocar o hardware fisicamente em operação para treinar modelos de aprendizagem profunda consome milhares de horas de tempo de CPU e GPU. Só o custo da eletricidade em si costuma ser proibitivo. Com a nuvem, os usuários pagam pelos recursos que usam sem os custos associados.
- Capacidade de colaborar e trabalhar em equipes distribuídas. Os modelos implantados na nuvem podem ficar acessíveis a todos os usuários com permissões, independentemente da localização física. Isso acelera o modelo de desenvolvimento, especificamente entre equipes remotas, o que vem se tornando cada vez mais comum.

- Capacidade de aproveitar recursos e ferramentas suplementares da plataforma. Segurança e disponibilidade foram os aspectos mais mencionados. Os sistemas em nuvem adequadamente configurados se beneficiam dos investimentos em segurança do cliente e do provedor da nuvem. Além disso, a capacidade de executar modelos nos datacenters locais para manter a conformidade com as leis de proteção de dados, como o GDPR, é importante. Com o armazenamento e a computação, os provedores de nuvem oferecem ferramentas e recursos de plataforma para melhorar a experiência do desenvolvedor. Ferramentas como o Amazon SageMaker oferecem grande valor para os clientes da nuvem, com fluxo de trabalho de machine learning totalmente gerenciado de ponta a ponta - desde a limpeza dos dados até o treinamento, a criação e a implantação de modelos.

A APRENDIZAGEM PROFUNDA É REAL

O ano passado demonstrou um avanço no estado do machine learning com 14% dos projetos em produção. Vimos um salto similar em 2019, com 38% dos 316 projetos de aprendizagem profunda classificados como em produção. 89% dos projetos de aprendizagem profunda em produção estão em execução na AWS. 66% dos projetos em produção utilizam o TensorFlow e 28% dos projetos usam o PyTorch. O Keras e o Apache MXNet também foram vistos nas configurações de produção, pois a maioria dos projetos tem componentes criados com várias estruturas. Apenas 9% dos projetos foram criados com apenas uma estrutura. Como as empresas reconheceram que a aprendizagem profunda e outros recursos de IA estão atingindo um nível de maturidade que lhes permite oferecer um legítimo valor de negócios, elas se esforçaram para implementar alternativas mais fáceis ou casos de uso mais simples, bem demonstrados e rápidos de implementar. Os exemplos comuns incluem interfaces de voz para sites e aplicativos e mecanismos de recomendação para sites de compras online.

As empresas ainda estão explorando aplicativos de aprendizagem profunda mais complexos também, mas muitos ainda estão em teste. Os resultados do estudo deste ano sugerem que as empresas expandiram seus investimentos gerais em aprendizagem profunda desde o ano passado, continuando o progresso em projetos aspiracionais mais complicados e plurianuais. Ao mesmo tempo, adicionaram aplicativos de valor agregado de rápida implementação, como recomendações, análise de sentimentos em bots de chat e interfaces de voz, para acompanhar as tendências do mercado e demonstrar o valor e a viabilidade contínuos da aprendizagem profunda para uso no mundo real.

APRENDIZAGEM PROFUNDA NA AWS

A Nucleus descobriu que os principais motivos para a escolha da AWS – a amplitude dos recursos da plataforma, o relacionamento com a Amazon e o investimento contínuo em serviços de aprendizagem profunda da AWS – permanecem inalterados desde o ano passado.

AMPLITUDE DOS RECURSOS DA AMAZON

A Amazon oferece suporte ao processo de aprendizagem profunda de ponta a ponta, portanto, não há necessidade de reunir os melhores componentes de fornecedores distintos. Os clientes podem armazenar dados, criar e implantar modelos e criar aplicativos que aproveitam as saídas do modelo, tudo na plataforma. Os clientes têm a flexibilidade de selecionar o hardware especializado otimizado para suas cargas de trabalho. Por exemplo, eles podem acessar instâncias com GPUs poderosas otimizadas para aprendizagem profunda, como a Amazon EC2 P3 e G4, sob demanda. A Amazon conta com datacenters regionais em todo o mundo, para que os clientes possam localizar seus dados e operações conforme necessário e cumprir os regulamentos regionais de compartilhamento de dados. A segurança no nível da plataforma adiciona outra camada de proteção aos dados e aplicativos. Outras ferramentas e serviços integrados à plataforma, como serviços de contêiner, servem bem a projetos de grande escala. Os usuários disseram:

- *"Outros provedores de nuvem não estão à altura da maturidade e abundância da plataforma da AWS. Recebemos muito mais do que tecnologia da parceria. Eles têm a experiência e os serviços adicionais como S3 e CloudFormation para nos ajudar a levar projetos de tecnologia do conceito à produção. "*
- *"Estamos avaliando constantemente outros fornecedores de nuvem, mas até agora ninguém se aproxima da AWS. Ninguém mais fornece valor comparável em serviços de nuvem, sem mencionar a melhor segurança e a disponibilidade da plataforma."*

Perfil de usuário – empresa de biotecnologia

Uma empresa de biotecnologia está usando a aprendizagem profunda para desenvolver classificadores em nível de produção para prever o câncer a partir de dados genômicos. Como ainda está em fase de pesquisa, está criando modelos para testar diferentes hipóteses em paralelo, de modo que a flexibilidade para implementar e testar rapidamente novas ideias é muito importante.

Há quatro abordagens diferentes sendo testadas em paralelo: duas são criadas no TensorFlow, uma no PyTorch e a outra foi desenvolvida com código totalmente personalizado. As estruturas foram escolhidas estrategicamente. O TensorFlow foi escolhido para dois projetos de estimadores pré-criados e funções de perda. O PyTorch foi escolhido para o outro porque a abordagem foi inspirada em um projeto na academia baseado no PyTorch.

A empresa escolheu a AWS para seus projetos de aprendizagem profunda por vários motivos. A Amazon oferece diferentes opções de computação que são mais adequadas para diferentes casos de uso. A aprendizagem profunda sobre dados genômicos requer pré-processamento extensivo, e a empresa precisava de máquinas com otimização de memória para pré-processar os dados de treinamento e máquinas aceleradas por GPU para treinar o modelo. Trabalhando com petabytes de dados de treinamento genômicos, a capacidade do parceiro de nuvem de apoiar um projeto nessa escala era crítica. Como os projetos usam estruturas diferentes, era importante que o provedor de nuvem oferecesse suporte a várias

estruturas como cidadãos de primeira classe em sua infraestrutura, sem a necessidade de uma configuração abrangente. A Amazon foi uma opção natural pelo desempenho que oferece, mesmo em escala de petabytes, e pela flexibilidade das máquinas compatíveis e seu trabalho para otimizar todas as estruturas para um bom funcionamento.

RELACIONAMENTO COM A AMAZON

Os clientes entrevistados também mencionaram seu relacionamento com a AWS como um fator determinante para a decisão de negócios. Muitos clientes já estavam usando a AWS em outras áreas da empresa e aumentaram seus investimentos para incluir a aprendizagem profunda. A aprendizagem profunda exige muitos dados, e as organizações já confiam seus dados à AWS. A Amazon também fornece aos clientes a tecnologia e as melhores práticas para concluir seus projetos sem prendê-los a qualquer solução específica. Os clientes foram rápidos em enfatizar a flexibilidade da plataforma: eles podem usar as estruturas e as bibliotecas que acharem mais confortáveis, sem se preocupar com a incompatibilidade com a tecnologia da AWS. Um cliente mencionou o fato de a Amazon estar interessada em estudos como este, como uma demonstração de seu compromisso em maximizar o valor do cliente. Outros usuários disseram:

- *"Criamos todo o nosso negócio na AWS, por isso seria preciso muito para nos motivar a escolher outro provedor. Nossa arquitetura da AWS cresceu conosco e o suporte foi sólido ao longo do caminho. Realmente sentimos que eles investiram em nosso sucesso, o que é fundamental para qualquer parceria de longo prazo."*
- *"Com a Amazon, temos acesso a casos de uso de clientes e experiência no setor para nossos projetos de machine learning. A equipe de suporte foi fundamental para nossos esforços. Eles foram capazes de responder a todas as nossas perguntas sobre estruturas, escolha de modelo e requisitos de infraestrutura."*

Perfil do usuário – empresa de mídia digital e jogos para dispositivos móveis

Uma empresa de mídia digital que projeta e vende jogos para dispositivos móveis usa a aprendizagem profunda para equilibrar a dificuldade do jogo com a receita projetada gerada. Isso é feito usando a aprendizagem profunda de reforço para treinar um bot para jogar cada jogo. Em seguida, ele pode monitorar o desempenho do bot para estimar a dificuldade de cada nível e prever quantos usuários deixarão de jogar em cada nível. Obviamente, quando um usuário para de jogar, ele para de gerar receita de publicidade para o fornecedor do jogo. Portanto, a empresa usa esse processo para equilibrar a dificuldade do jogo com a receita projetada.

A AWS é o provedor de nuvem da empresa desde 2008, por isso usar a AWS para a aprendizagem profunda foi um ato automático. Ambos os projetos são criados no PyTorch, o padrão do setor para a aprendizagem profunda de reforço, e grande parte da pesquisa atual é publicada com a implementação do PyTorch.

A empresa usou o SageMaker para implantar o modelo responsável por prever o desgaste de jogadores na produção. Foi escolhido porque proporciona à liderança do projeto uma

"visão panorâmica" do modelo e oferece à liderança do projeto um local centralizado para visualizar e controlar todos os modelos na implantação. Com tecnologia como o SageMaker combinada com os recursos flexíveis de computação oferecidos pela Amazon, o especialista entrevistado disse: "seria necessário um argumento comercial convincente para nos persuadir a deixar a AWS e até agora nenhum outro provedor de nuvem pôde nos oferecer as ferramentas específicas de machine learning, além de armazenamento e computação por um [melhor valor]."

INVESTIMENTO DA AWS EM APRENDIZAGEM PROFUNDA

Os clientes sabem que a Amazon está desenvolvendo e usando sua própria tecnologia de aprendizagem profunda. Os especialistas em aprendizagem profunda mencionaram as melhorias contínuas na documentação, o suporte da estrutura e os serviços de nuvem, como o Amazon SageMaker, como fatores principais na escolha da AWS em relação a outros provedores de nuvem. O SageMaker ficou disponível em 2017. É um serviço em nuvem totalmente gerenciado que cobre todo o fluxo de trabalho de machine learning, desde a criação, o treinamento e a implantação de modelos de machine learning.

A adoção do SageMaker está crescendo rapidamente à medida que os desenvolvedores percebem como ele pode reduzir a complexidade e acelerar a implantação do modelo. No ano passado, aproximadamente um terço dos entrevistados estava usando ou explorando o uso do SageMaker para automatizar aspectos de seus projetos de aprendizagem profunda. Neste ano, esse número quase dobrou, com 63% dos clientes usando ou avaliando a possibilidade de usar o SageMaker. No decorrer deste estudo, encontramos clientes que estavam migrando suas implantações internas do TensorFlow para um serviço gerenciado na AWS por meio do SageMaker, bem como clientes que criaram seus sistemas totalmente do zero com o SageMaker. Os usuários disseram:

- *"Não precisamos adquirir hardware dedicado para executar projetos de aprendizagem profunda em larga escala. Sem a AWS e, especificamente, o SageMaker, precisaríamos comprar hardware, treinar o modelo localmente e, em seguida, armazenar e hospedar o modelo em um servidor interno para que ele estivesse acessível quando eu quisesse fazer previsões. Só para começar, isso pode levar semanas e vem com uma tonelada de custos adicionais de hardware, eletricidade e pessoal, sem mencionar todo o tempo perdido na criação de recursos internos pré-criados no SageMaker."*
- *"Nosso foco é implementar o machine learning de conversação para nossos clientes e o SageMaker simplifica a criação de modelos. Em média, temos cinco projetos em andamento para um cliente em uma determinada semana. Desde que começamos a usar o SageMaker no ano passado, descobrimos que economizamos cerca de duas horas por projeto ao automatizar o treinamento do modelo distribuído."*

Perfil do usuário – empresa de software empresarial

Uma empresa global de software corporativo que produz aplicativos principalmente para equipes de vendas e serviços adotou o Amazon SageMaker para gerenciar sua implantação do TensorFlow. Seus esforços de aprendizagem profunda abrangem principalmente a análise de sentimentos e a classificação das interações com os clientes, a fim de entender efetivamente como os diferentes tipos de divulgação afetam a probabilidade de o cliente se interessar ou comprar novamente.

Como uma grande empresa de tecnologia, ela se antecipou com seus esforços de aprendizagem profunda em comparação com o mercado maior. Antes do SageMaker, ela criou a maior parte de sua infraestrutura de aprendizagem profunda baseada no TensorFlow. No início do ano, decidiu migrar a implantação autogerenciada do TensorFlow para o SageMaker, no qual poderia ser gerenciada como serviço. Enquanto o esforço está em andamento e ainda não está totalmente completo, a organização conseguiu reatribuir 3 ETIs até agora, que foram os principais responsáveis pelo gerenciamento do ecossistema do TensorFlow. Além disso, a velocidade dos modelos de treinamento aumenta drasticamente, pois o SageMaker distribui automaticamente a carga de computação em várias CPUs ou GPUs em paralelo. A empresa informou que a implantação de um novo modelo com o SageMaker leva menos de 50% do tempo necessário para fazê-lo em um ambiente autogerenciado.

Perfil do usuário – empresa de desenvolvimento de aplicativos

Uma empresa de desenvolvimento de aplicativos especializada na criação de jogos integrados por voz que rodam no Amazon Alexa, a caixa de som inteligente, criou um projeto de aprendizagem profunda inteiramente no Amazon SageMaker que recomenda jogos para manter o envolvimento do usuário elevado. O sistema usa dados de jogos anteriores jogados no sistema para recomendar, de forma previsível, outros jogos semelhantes ao usuário.

A empresa criou os negócios dela na plataforma da AWS, por isso escolheu utilizar o SageMaker para a integração nativa com a arquitetura existente na AWS, principalmente o Amazon S3 e o AWS Lambda, para acessar dados armazenados e computação sem servidor. Além disso, como a empresa já estava na AWS, as permissões de usuário e os procedimentos de DevOps já haviam sido formalizados, permitindo evitar a duplicação desse esforço. O treinamento do modelo em paralelo e a interface do usuário no nível de controle tornaram o treinamento e a avaliação muito mais rápidos do que seriam manualmente. O cliente estima que o uso do SageMaker torna o treinamento do modelo três vezes mais rápido e a implantação do modelo quatro vezes mais rápida do que o gerenciamento manual do sistema.

CONCLUSÃO

No ano passado, vimos projetos de aprendizagem profunda saindo *em massa* da sala de aula para a sala de reuniões, à medida que os avanços no machine learning, na tecnologia de computação, na infraestrutura de nuvem e nos conjuntos de dados adequados ajudaram a tornar a aprendizagem profunda comercialmente viável. Neste ano, a tendência continuou com mais e mais empresas buscando implementar a aprendizagem profunda em escala para resolver problemas do mundo real. Pela segunda vez em dois anos, o número de projetos de aprendizagem profunda em produção mais que dobrou, principalmente em função dos mesmos fatores-chave do ano passado.

- Maior disponibilidade de infraestrutura e serviços de nuvem de fornecedores como a AWS para oferecer suporte a processos que envolvem muitos dados e computação, como a aprendizagem profunda.
- Avanços para o estado da arte em aprendizagem profunda com técnicas aprimoradas, arquiteturas de rede e conjuntos de dados que tornam as redes neurais mais precisas e capazes.
- Investimento contínuo na comunidade para compartilhar experiências e capacitar outros pesquisadores de aprendizagem profunda por meio de fóruns e documentação on-line, bibliotecas e estruturas de código aberto e ofertas em nuvem, como modelos pré-criados e hardware especializado otimizado para machine learning.

À medida que o custo inicial para explorar a aprendizagem profunda diminui, vemos mais e mais empresas procurando entrar na briga. Em vez de procurar reinventar a roda, a estratégia mais eficiente para esse fim é fazer parceria com um fornecedor de nuvem que tenha a infraestrutura, a experiência e os serviços adicionais para levar a aprendizagem profunda do conceito à conclusão. De nossa análise, descobrimos que a reputação da Amazon como o provedor de tecnologia em nuvem empresarial mais maduro e sofisticado, juntamente com seus investimentos específicos em campo em serviços de machine learning e flexibilidade de plataforma, para oferecer suporte à escolha do cliente pela arquitetura de rede, estrutura de desenvolvimento ou fontes de dados, o tornam a plataforma de nuvem preferida para profissionais de aprendizagem profunda.