



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

**Avaliação do uso de bibliotecas de aprendizagem de
máquina para integração com simuladores de redes
ópticas elásticas**

Mikael Marques Mello

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. André Costa Drummond

Brasília
2020



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

**Avaliação do uso de bibliotecas de aprendizagem de
máquina para integração com simuladores de redes
ópticas elásticas**

Mikael Marques Mello

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. André Costa Drummond (Orientador)
CIC/UnB

Prof. Dr. Donald Knuth Dr. Leslie Lamport
Stanford University Microsoft Research

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 5 de dezembro de 2020

Dedicatória

Na *dedicatória* o autor presta homenagem a alguma pessoa (ou grupo de pessoas) que têm significado especial na vida pessoal ou profissional. Por exemplo (e citando o poeta): *Eu dedico essa música a primeira garota que tá sentada ali na fila. Brigado!*

Agradecimentos

Nos *agradecimentos*, o autor se dirige a pessoas ou instituições que contribuíram para elaboração do trabalho apresentado. Por exemplo: *Agradeço aos gigantes cujos ombros me permitiram enxergar mais longe. E a Google e Wikipédia.*

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Atualmente, pesquisadores estão cada vez mais realizando estudos acerca das aplicações de aprendizagem de máquina (ML) na área de redes ópticas elásticas (EONs). Entretanto, simuladores de EONs existentes atualmente não permitem a integração de ML em suas simulações. Este trabalho apresenta uma análise qualitativa e quantitativa de bibliotecas de ML populares na literatura com o objetivo de definir direções e recomendações para futuras implementações, além de realizar uma contribuição ao Optical Network Simulator (ONS) com uma implementação inicial de funcionalidades para dar suporte ao uso de modelos de ML em simulações.

Palavras-chave: redes ópticas elásticas, aprendizagem de máquina

Abstract

Researchers are increasingly studying the applications of machine learning (ML) in Elastic Optical Networks (EONs). However, existing EON simulators do not currently support the integration of ML models in their simulations. This work presents a qualitative and quantitative analysis of popular ML libraries aiming to define directions and recommendations for future implementations, and a contribution to the Optical Network Simulator (ONS) implementing initial features to support the use of ML models.

Keywords: elastic optical networks, machine learning

Sumário

1	Introdução	1
1.1	Problema	3
1.2	Objetivos	4
1.3	Contribuições	4
1.4	Organização do trabalho	4
2	Fundamentação Teórica	6
2.1	Aprendizagem de Máquina	6
2.1.1	Aprendizado supervisionado	7
2.1.2	Aprendizado não-supervisionado	7
2.1.3	Aprendizado por reforço	7
2.2	Aprendizagem profunda	8
2.3	Simulações de EONs	8
3	Avaliação Qualitativa	9
3.1	OpenCV	11
3.2	PyTorch	11
3.3	Scikit Learn	12
3.4	TensorFlow	12
3.5	Keras	13
3.6	TensorFlow Lite	13
3.7	ONNX Runtime	14
3.8	Deeplearning4j	14
3.9	MXNet	15
3.10	Outras tecnologias	15
4	Avaliação de Desempenho	16
4.1	Motivação	16
4.2	Metodologia	16
4.2.1	Ambiente de Simulação	17

4.2.2 Modelos e Cargas de Trabalho	17
4.2.3 Métricas	17
4.2.4 Execução	18
4.3 Resultados	18
5 Resultados	21
6 Conclusão	22
Referências	23

Lista de Figuras

3.1 Porcentagem de cientistas de dados e desenvolvedores de inteligência artificial que usam ou priorizam cada linguagem [1]	10
4.1 Tempo de execução de execuções sequenciais do modelo para cada programa de simulação	20

Lista de Tabelas

4.1 Estatísticas descritivas acerca do tempo de execução (microsegundos) do modelo em todas as combinações de simulação, descartadas as primeiras 1000 execuções	19
--	----

Capítulo 1

Introdução

Historicamente, o tráfego de internet global cresce de forma exponencial, possuindo uma taxa composta de crescimento anual de 45% nos anos 2000 [2] e aproximadamente 30% nos anos 2010 [3, 4, 5, 6, 7, 8, 9, 10]. Uma fatia significativa deste crescimento anual se deve ao tráfego de dados em redes móveis, cujas taxas de crescimento anuais na última década tem variado entre 50% a 60% e são motivadas pelo crescente número assinaturas de *smartphones* e o volume de dados consumido por assinatura, este alavancado principalmente pelo crescente consumo de conteúdos de vídeo [11]. Além disso, circunstâncias especiais podem incentivar a população a aumentar ainda vez mais seu uso de internet: Em Abril de 2020, a empresa Akamai relatou um crescimento de 30% do tráfego global em apenas um mês, aproximadamente dez vezes a taxa de crescimento esperada, atribuindo o pico de crescimento às mudanças de estilo de vida causadas pela pandemia do COVID-19 [12].

A infraestrutura responsável por lidar com tamanho tráfego é composta por redes ópticas que têm sido tradicionalmente *rígidas* e *homogêneas*, isto é, redes baseadas em *Wavelength-Division Multiplexing* (WDM), ou multiplexação por divisão de comprimento de onda. Redes baseadas em WDM oferecem a possibilidade de estabelecer conexões com comprimentos de onda fixos e com uma taxa de *bits* fixa, em que os canais são modulados com um formato comum e espaçados por uma distância fixa de 50 GHz [13, 14]. O crescimento exponencial de demanda motivou a indústria a focar esforços em aumentar a capacidade destas redes, resultando na evolução das taxas de *bits* permitidas de inicialmente 10 Gb/s para 40 Gb/s e por fim 100 Gb/s [15]. Entretanto, este tipo de rede conta com alguns problemas:

- **Baixa adaptabilidade.** A flexibilidade destas redes é limitada pela configuração do *hardware*, tornando o processo de atualizar ou modificar a rede para se adaptar a mudanças de demanda ou de condições de rede desafiador [13].

- **Baixa eficiência espectral.** O desenho da rede deve garantir que o caminho óptico mais longo (pior caso) seja transmitido com qualidade suficiente. Como os comprimentos de onda são fixos e homogêneos, a maioria das conexões irá possuir comprimentos muito menores do que o pior caso, gerando um problema de ineficiência onde há faixas de comprimentos de onda não utilizadas [14, 15].
- **Limite de futuros avanços.** Com taxas de bits maiores que 100 Gb/s, melhorias na eficiência espectral ao aumentar o número de *bits* por símbolo se torna cada vez mais difícil devido ao limite de Shannon, além do fato de que aumentar a taxa para além de 100 Gb/s é um desafio por si só [14].

Por estas dificuldades, o conceito de *Elastic Optical Networking* (EON), ou redes ópticas elásticas, foi introduzido como um modo de oferecer uma utilização eficiente dos recursos ópticos disponíveis, sendo capaz de acomodar taxas de *bits* que variam desde alguns Gb/s até a magnitude de *terabits* por segundo [13, 16], além da alocação adaptável de recursos de *hardware* e espectrais de acordo com a demanda do tráfego [14]. Na literatura, os termos "flexível", "elástico", "flexgrid ou flexigrid", "gridless" e "adaptável" são usados intercambiavelmente. Esta alocação flexível é permitida graças ao uso de tecnologias como *Orthogonal Frequency Division Multiplexing* (OFDM), *Nyquist WDM* (N-WDM) e *Optical Arbitrary Waveform Generation* (OAWG) [13]. EONs têm sido amplamente aceitas como uma das melhores soluções com arquiteturas de rede flexíveis e capazes de alocar recursos de forma flexível [17].

A nível de design e otimização de *Elastic Optical Networks* (EONs), um dos principais desafios em seu desenvolvimento se trata da alocação eficiente de recursos. Algoritmos e ferramentas de planejamento de rede convencionais (WDM) não podem ser aplicados devido à natureza flexível das redes. Assim, de modo a aproveitar completamente a flexibilidade disponibilizada pelos avanços de tecnologias na camada física, novos algoritmos de alocação de recursos têm sido explorados [13, 18, 19, 20, 21]. Por exemplo, com flexibilidade apenas no número de subportadoras disponíveis para alocação técnicas de roteamento e alocação de espectro, *Routing and Spectrum Assignment* (RSA), devem ser usadas. Caso haja flexibilidade na seleção do formato de modulação, então técnicas de roteamento e atribuição de espectro com modulação adaptativa, *Routing, Modulation Level and Spectrum Allocation* (RMLSA), devem ser escolhidas.

A avaliação de desempenho de sistemas de comunicação óptica é um desafio para os pesquisadores. A dinamicidade e complexidade, especialmente em redes ópticas elásticas, torna inviável uma modelagem analítica precisa e o uso de ambientes reais para medições torna a avaliação bastante custosa, devido principalmente aos equipamentos e ferramentas envolvidos. Assim, a simulação é a alternativa disponível para atividades de teste, validação e avaliação de novos protocolos de controle para o ambiente de redes ópticas [22].

Diferentes ferramentas de simulação de redes ópticas elásticas foram desenvolvidas para auxiliar os pesquisadores a implementar, testar e analisar novos algoritmos ou soluções de problemas diferentes da área. Como por exemplo, *Optical Network Simulator* (ONS) [22], *ElasticO++* [23], *Complex Elastic Optical Network Simulator* (CEONS) [24] e *Net2Plan* [25].

Recentemente, com a crescente popularização do uso de *Machine Learning* (ML), ou aprendizagem de máquina, na academia, estudos sobre a aplicação de ML para solução de problemas relacionados a EONs também têm se popularizado.

Machine Learning é o estudo de algoritmos de computação que se auto-otimizam de acordo com um critério de desempenho, usando dados de exemplo ou a própria experiência [26, 27]. Em sua forma mais básica, o método de ML constitui-se em coletar dados relevantes ao domínio do problema a ser resolvido para serem usados por um algoritmo de aprendizagem, o modelo. Este modelo é definido com parâmetros iniciais que são otimizados automaticamente a partir do consumo dos dados coletados. O objetivo do modelo pode ser tanto preditivo, para realizar previsões no futuro sobre dados potencialmente desconhecidos; descritivo, para obter-se conhecimento novo acerca dos dados; ou ambos [27, 28]. O processo de realizar previsões é chamado de inferência.

Dentre recentes usos de ML em pesquisas relacionadas a EONs, pode-se citar a pesquisa de Yu et al., que desenvolveu uma estratégia RSA baseada em *Deep Learning* (DL) [17], ou aprendizagem profunda, um subgrupo de ML. Guilherme et al. desenvolveram um modelo de DL capaz de identificar estratégias RSA em EONs com 98% de acurácia [29]. Outras pesquisas acerca do uso de ML em problemas na área de EONs podem ser encontradas em [30] e [31].

Devido ao recente crescimento do uso de ML em pesquisas no campo de EONs, os simuladores de redes ópticas elásticas disponíveis na literatura ainda não possuem integração com modelos de ML para uso durante as simulações. O presente trabalho visa analisar na literatura atual as bibliotecas de ML capazes de importar modelos pré-treinados, especificamente no campo de aprendizagem profunda como o desenvolvido por Yu et al. [17], no contexto de integrá-las à execução de simulações de EONs.

1.1 Problema

Devido ao recente crescimento do uso de ML em pesquisas no campo de EONs, os simuladores de redes ópticas elásticas disponíveis na literatura ainda não possuem suporte nativo à execução de modelos de ML para uso durante as simulações. Assim, pesquisas que envolvem o uso de ML para a solução de problemas como RSA não possuem um *framework* definido para analisar o desempenho de seus algoritmos ou soluções.

1.2 Objetivos

O objetivo geral deste trabalho é a análise das bibliotecas de ML disponíveis na literatura, com foco na capacidade de execução de modelos de ML pré-treinados em simulações de EONs executadas por simuladores conhecidos na literatura, como o ONS [22]. Sendo atingido por meio dos seguintes objetivos específicos:

- Estudar as bibliotecas existentes na literatura capazes de executar modelos de ML;
- Selecionar bibliotecas apropriadas para o caso de uso de integração com simuladores de EONs;
- Avaliar de forma quantitativa o desempenho das bibliotecas com métricas e parâmetros pré-definidos.
- Analisar os resultados das avaliações e elencar as bibliotecas recomendadas para diferentes casos de uso.
- Realizar uma implementação de integração de ML com o ONS como prova de conceito e contribuição para a comunidade acadêmica.

1.3 Contribuições

A partir das análises qualitativas e quantitativas das bibliotecas de *machine learning* populares na literatura, este trabalho contribui com diversas recomendações acerca de como realizar a integração de um simulador de EONs com modelos de ML, considerando diversos casos de uso de acordo com o tipo de pesquisa sendo realizada.

Adicionalmente, é feita uma contribuição para o repositório do simulador ONS, onde são realizadas as seguintes implementações: configuração do projeto para utilização do Maven como ferramenta de gerenciamento de projetos; implementação de classes abstratas para implementação de modelos de ML utilizando as bibliotecas ONNX e Deeplearning4j; implementação do uso de um modelo de ML para classificar a topologia da rede e imprimir a avaliação durante a execução de um algoritmo RSA. É esperado que esta implementação sirva de base para futuros estudos da aplicação de ML em EONs.

1.4 Organização do trabalho

O trabalho está organizado em 6 capítulos com os seguintes propósitos:

- Capítulo 2 - Introdução de conceitos considerados essenciais para o entendimento do trabalho. São descritos de forma mais profunda o uso de simuladores para pesquisas no campo de redes ópticas elásticas e conceitos de ML usados ao longo do trabalho.
- Capítulo 3 - Descrição das bibliotecas de ML populares na literatura, sendo explicadas suas principais características e o racional para a consideração delas ou não na análise de desempenho.
- Capítulo 4 - Serão descritas as propriedades da análise de desempenho como métricas, parâmetros, fatores, carga de trabalho, e a máquina em que os testes foram realizados, detalhes da implementação do ambiente de simulação e os resultados das simulações.
- Capítulo 5 - Serão avaliados os resultados quantitativos das simulações e os aspectos qualitativos de cada possibilidade para a definição de propostas e recomendações de integração de simuladores com ML.
- Capítulo 6 - Será apresentado qual foi o resultado obtido, o conjunto de recomendações que melhor satisfazem os objetivos da integração de modelos de ML com simuladores de redes ópticas elásticas.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta conceitos básicos fundamentais para o melhor entendimento do trabalho. Na seção 2.1 são introduzidos conceitos fundamentais de aprendizagem de máquina e uma breve descrição de suas principais categorias.

São introduzidos detalhes do funcionamento de simulações de EONs e a relevância deles para o trabalho, aprofundados são aprofundados conceitos de machine learning como suas aplicações, seu funcionamento e sua relevância para EONs. Por fim, na seção ??, são discutidos detalhes da execução de modelos de *deep learning* importantes para o entendimento de considerações feitas neste trabalho.

2.1 Aprendizagem de Máquina

Machine Learning (ML), ou aprendizagem de máquina é o estudo de algoritmos de computação que se auto-otimizam de acordo com um critério de desempenho, usando dados de exemplo ou a própria experiência [26, 27].

Em sua forma mais básica, o método de ML é a coleta de um grande número de dados do domínio do problema a ser resolvido e o uso deles em um algoritmo de aprendizagem, o modelo. Este modelo é definido com parâmetros iniciais que são otimizados com o consumo dos dados. O objetivo do modelo pode ser tanto preditivo, para realizar previsões no futuro sobre dados potencialmente desconhecidos, descritivo, para obter-se conhecimento novo acerca dos dados, ou ambos. O processo de obter resultados de um modelo é chamado de inferência, sejam eles preditivos ou descritivos.

Há dois grandes desafios no campo de ML: primeiramente, são necessários algoritmos eficientes para o problema de otimização do modelo inicial, de modo que a fase de treinamento seja completada em tempo viável; segundamente, uma vez que um modelo tenha sido aprendido, sua representação e solução algorítmica para a inferência também devem ser eficientes. [27] [28]

Normalmente, o campo de ML é dividido em três principais categorias: *supervised learning*; *unsupervised learning*; e *reinforcement learning*.

2.1.1 Aprendizado supervisionado

Em inglês, *supervised learning* é a categoria em que o conjunto de dados de entrada (*training set*) possui um mapeamento para o comportamento esperado, rotulado por um "supervisor" [27]. O objetivo é aprender uma regra geral que mapeie os valores de entrada para os respectivos valores esperados de saída. Os valores de saída podem ser contínuos (problemas de regressão) ou discretos (problemas de classificação) [32].

2.1.2 Aprendizado não-supervisionado

Em inglês, *unsupervised learning* é a categoria em que existe apenas o conjunto de dados de entrada e o objetivo é encontrar regularidades presentes nos mesmos [27]. Este tipo de aprendizagem é capaz de desempenhar várias tarefas, porém a mais comum é *clustering* [32].

Clustering é o processo de agrupar dados de modo que a similaridade de dados nos grupos (*clusters*) é alta, porém a similaridade de dados entre grupos diferentes é baixa. Esta similaridade é tipicamente expressada como uma função de distância, que depende do tipo de dados presente no conjunto [32].

Dentre os usos de aprendizagem não-supervisionada, pode-se destacar análise de redes sociais, agrupamento de genes e pesquisa de mercado como aplicações bem-sucedidas [32].

2.1.3 Aprendizado por reforço

Em inglês, *reinforcement learning* é a categoria em que o algoritmo de aprendizagem tem como objetivo aprender uma política de ações que maximizem a recompensa em um dado ambiente [27].

O paradigma de *reinforcement learning* permite que agentes explorem possíveis ações e refinem seu comportamento utilizando apenas uma avaliação, conhecida como recompensa, tendo como objetivo maximizar seu desempenho de longo prazo [32].

Esta técnica é comumente usada em aplicações como robótica, área de finanças como decisões de investimentos e gerenciamento de estoque [32].

2.2 Aprendizagem profunda

2.3 Simulações de EONs

Capítulo 3

Avaliação Qualitativa

Pela visão de auxiliar pesquisadores em suas integrações de modelos de ML com simuladores de redes ópticas elásticas, as análises aqui feitas devem buscar o aumento de produtividade do pesquisador. Para isto, as tecnologias de ML analisadas serão avaliadas de forma qualitativa, sendo esta avaliação guiada pelas seguintes questões:

- Há algum custo para usar a tecnologia?
- O código da tecnologia é aberto?
- A tecnologia é ativamente mantida por *maintainers* e/ou pela comunidade?
- O uso da tecnologia é amplamente documentado?
- A instalação e uso da tecnologia é simples?
- A tecnologia permite a execução de modelos pré-treinados?
- A tecnologia permite a execução de modelos pré-treinados com outras tecnologias?

Nesta análise, apenas tecnologias com licenças de código aberto serão consideradas de modo que os pesquisadores tenham livre acesso às recomendações e sejam capazes de manipulá-las em seus projetos, caso seja necessário. Dentre essas, apenas tecnologias disponíveis para uso em programas Java ou Python serão consideradas.

A linguagem Java é amplamente utilizada para a implementação de simuladores de redes ópticas elásticas, como em [22], [24] e [25], por isto, a integração de modelos de ML de forma embutida no simulador permite uma integração performática ao excluir-se a necessidade de comunicação com outros serviços ou processos para executar os modelos.

A linguagem Python é a mais usada e priorizada para desenvolvimento em ML entre trabalhadores da área, como evidenciado na figura 3.1. Assim, uma integração de simuladores com um processo ou serviço independente escrito em Python, responsável por

MACHINE LEARNING LANGUAGES SHOOT OUT – HOW TO PICK THE RIGHT LANGUAGE

% of machine learning developers / data scientists who use or prioritise each language (n = 2,022)



Source: State of the Developer Nation Q1 2017 | <http://vmob.me/DE1Q17> | Copyright VisionMobile | Licensed under CC BY ND | All rights reserved

Figura 3.1: Porcentagem de cientistas de dados e desenvolvedores de inteligência artificial que usam ou priorizam cada linguagem [1]

executar os modelos, pode ser mais fácil de ser desenvolvida e mantida por um pesquisador de ML na área de EONs, apesar do custo de tempo da execução graças ao tempo necessário para comunicação entre os processos ou serviços.

As tecnologias também devem ser fáceis de serem instaladas, configuradas e manipuladas de acordo com as necessidades de cada pesquisador. Para isto, é fundamental que suas APIs sejam bem documentadas e que a instalação exija o mínimo de modificações de configurações da máquina, externas ao simulador.

Por fim, é ideal que a análise considere a flexibilidade das bibliotecas à respeito dos diversos tipos de modelos de *machine learning*, uma vez que pesquisas sobre o uso de ML em EONs utilizam diversas técnicas.

3.1 OpenCV

OpenCV (Open Source Computer Vision Library) é uma biblioteca de código aberto voltada para visão computacional e aprendizagem de máquina, construída para fornecer uma infraestrutura comum para aplicações de visão computacional e acelerar o uso de percepção de máquina em produtos comerciais [33].

Apesar do foco principal de OpenCV ser visão computacional, a biblioteca possui um módulo de redes neurais profundas e interfaces para as linguagens Python e Java. Adicionalmente, também é possível realizar a importação de modelos serializados em diversos formatos, como Darknet [34], Torch7 [35], ONNX [36] e TensorFlow [37].

A API da biblioteca é extensamente documentada, porém com poucos tutoriais sobre o uso do módulo de DNNs. Entretanto, no quesito facilidade de instalação e configuração os resultados foram variados:

Em Python, para a execução de modelos com apenas o uso da CPU a instalação se resume a instalar o pacote `opencv-python-headless`, a versão sem dependências de bibliotecas de interfaces de usuário gráficas, e está pronto para uso. Porém, se há interesse em utilizar a GPU na execução dos modelos, o processo de instalação se torna bastante complexo, sendo necessário compilar manualmente a biblioteca considerando diversas configurações do ambiente da máquina, de modo que a criação de uma solução generalizada se torne inviável.

Em Java, o processo de instalação é complexo independente do uso ou não de GPU, sendo necessário o mesmo processo de compilação manual do projeto considerando configurações da máquina, não havendo nenhuma integração com gerenciadores de pacotes populares como *Maven* ou *Gradle*.

Assim, pelas dificuldades presentes na instalação, o único uso de OpenCV considerado é o de Python com as execuções sendo realizadas apenas pela CPU da máquina, sem uso de GPU.

3.2 PyTorch

PyTorch é uma biblioteca de ML de código aberto que provê uma plataforma de pesquisas em aprendizagem profunda, oferecendo máxima flexibilidade e velocidade [38], sendo considerada a biblioteca de aprendizagem profunda que mais cresce no mundo [39].

A biblioteca possui suporte para Python e conta com ampla documentação de diversas práticas. Na análise da biblioteca, não foram encontrados métodos nativos de conversão de modelos de outros formatos para serem inferidos com o uso de PyTorch e por este motivo a biblioteca foi descartada.

3.3 Scikit Learn

Scikit-learn é uma biblioteca de código-aberto desenvolvida em Python que integra diversos algoritmos de ML para problemas supervisionados e não-supervisionados de média escala com ênfase em trazer ML para não-especialistas [40]. É uma biblioteca extremamente popular, principalmente entre iniciantes na área de ML, com desenvolvimento ativo na comunidade.

A biblioteca, com suporte apenas para Python, possui extensa documentação de modo a melhor auxiliar iniciantes. Pelo seu foco em simplicidade, a biblioteca não possui suporte para uso de GPU em treinos ou execuções de modelos. Para importação e exportação de modelos, a biblioteca conta apenas com serialização nativa de Python por meio da biblioteca *pickle*, não sendo possível importar modelos de outros formatos.

Apesar da facilidade de se desenvolver modelos de ML com *Scikit Learn*, o uso da biblioteca foi descartado pela falta de suporte ao uso de GPU e de importação de modelos pré-treinados salvos em formatos diversos.

3.4 TensorFlow

TensorFlow é uma interface para a expressão de algoritmos de ML e uma implementação para a execução de tais algoritmos [41]. Além da biblioteca principal (TensorFlow Core), a organização possui vários produtos para a execução de algoritmos de ML em larga-escala (TensorFlow Extended), execução de modelos em sistemas embarcados, mobile ou de baixa potência (Tensorflow Lite) e execução de modelos em JavaScript, para ser usado em browsers ou em Node.js (TensorFlow.js).

TensorFlow é uma das bibliotecas mais populares na área de ML, sua biblioteca principal possui ampla documentação além de diversos guias e tutoriais desenvolvidos pela comunidade. Está em ativo desenvolvimento e tem o apoio de diversas empresas de tecnologia reconhecidas.

O uso de TensorFlow Extended está fora do escopo desta pesquisa pelo objetivo de integração com a execução de um simulador e não um sistema de larga-escala. TensorFlow Lite será avaliado de forma mais profunda na seção 3.6 e TensorFlow.js também está fora do escopo por ser exclusivo para JavaScript. Restando apenas considerar a biblioteca principal, que possui suporte para Python, C, Go e Java.

O suporte para Java está em uma fase de transição, onde a biblioteca antiga está depreciada para uso e a biblioteca nova ainda está em fase experimental. O uso da biblioteca em fase experimental é complexo e não existe quase nenhuma documentação, sendo encontrada até uma recomendação de ler os testes unitários da implementação para aprender

como usar certos elementos da API [42]. Por estes motivos, o uso de TensorFlow em Java foi descartado.

O suporte para Python é a base da implementação da biblioteca, logo a biblioteca de Python satisfaz todos os requisitos de nossa análise qualitativa, pelo que mencionado anteriormente nesta seção, sendo considerada para execução de modelos com e sem GPU, utilizando a biblioteca Keras como a API de alto nível para implementação do programa de simulação.

3.5 Keras

Keras [43] é uma biblioteca de *deep learning* de alto nível, desenvolvida usando a plataforma Tensorflow como base, sendo seu foco principal foco a capacidade de fácil experimentação.

A biblioteca é uma interface de alto nível para a criação e execução de modelos, instalada por meio da biblioteca do TensorFlow, sendo então utilizada como interface para a implementação do programa de simulação responsável por avaliar o desempenho da biblioteca TensorFlow.

3.6 TensorFlow Lite

TensorFlow Lite [44] é uma *framework* de aprendizagem profunda para execução de modelos em dispositivos. Se trata da versão da popular *framework* TensorFlow que é projetada para execuções em dispositivos com menor poder computacional.

A importação de gráficos é limitada apenas a arquivos do tipo TFLite, porém existem ferramentas para realizar a conversão de formatos comuns em Tensorflow, como Keras e SavedModel. Atualmente, a biblioteca não fornece suporte a execução com uso de GPUs NVIDIA [45].

Tendo os fatores acima em consideração, a instalação e uso da biblioteca são simples e amplamente documentados para programas Python. Para a instalação, é necessário instalar a versão do pacote específica para a versão do interpretador Python instalado na máquina, sendo possível configurar uma detecção automática. A biblioteca não possui versões para uso em programas Java que não sejam voltados para Android.

Apesar da limitação de importação de modelos e a impossibilidade de uso em programas Java, TensorFlow Lite será avaliado de forma quantitativa pelo seu foco específico de execução rápida de modelos em dispositivos de borda.

3.7 ONNX Runtime

A ONNX Runtime [46] se trata de um acelerador de treinamento e execução de modelos de ML salvos em formatos ONNX [36]. *Open Neural Network Exchange* (ONNX) é um formato aberto construído para ser uma representação comum de modelos de ML, possuindo amplo suporte em suas bibliotecas de ML.

A ONNX Runtime é desenvolvida com foco em suporte cross-plataforma, fornecendo uma API comum para diversas linguagens como C, C++, Java e Python. Seu desenvolvimento é apoiado por diversas empresas como Microsoft, Facebook, IBM, Intel, NVIDIA, entre outras.

Graças ao foco de implementações cross-plataforma, a instalação das bibliotecas para Java e Python são simples e exigem pouca ou nenhuma configuração. Em ambos os casos, é necessário apenas instalar o respectivo pacote caso a intenção seja usar apenas a CPU para inferências. Se houver interesse em usar uma GPU NVIDIA, é necessário instalar a versão da *runtime* compatível com a versão de CUDA instalada.

A documentação da biblioteca possui poucos guias ou tutoriais mas conta com diversos exemplos de código em Python para operações comuns e um exemplo de código em Java, além da completa referência dos pacotes de ambas versões da biblioteca.

Apesar da documentação limitada, a ONNX Runtime será considerada na avaliação de desempenho pela alta portabilidade da biblioteca, sendo ideal para atender aos diferentes casos de uso de pesquisas de ML em EONs.

3.8 Deeplearning4j

Eclipse Deeplearning4j [47] é uma biblioteca de código-aberto para aprendizagem de máquina distribuída, disponível para Java e Scala. Seu desenvolvimento está ativo e é conduzido pela empresa Konduit.

A biblioteca possui ampla documentação da API e diversos tutoriais que explicam os diferentes usos da biblioteca. É possível importar modelos no formato ONNX, HDF5 e TensorFlow, além de seu próprio formato.

A instalação da biblioteca é extremamente simples para uso de CPU e se dá por meio do gerenciador de pacotes Maven, sendo necessário apenas adicionar apenas algumas linhas na configuração do projeto para que seja possível executar o Deeplearning4j. Para uso de GPU, apenas máquinas com placas de vídeo NVIDIA e com CUDA configurado são suportadas. Neste caso, o processo de instalação é quase o mesmo, tendo como diferença o identificador da biblioteca que depende da versão da GPU instalada na máquina.

A biblioteca Deeplearning4j será avaliada em um programa Java com uso de ambas CPU e GPU.

3.9 MXNet

A biblioteca MXNet [48] é uma *framework* de código-aberto para aprendizagem profunda que permite a definição, o treinamento e a execução de redes neurais profundas em diversas plataformas, estando atualmente em desenvolvimento, em processo de incubação pela Apache Incubator. Se trata de uma biblioteca bastante versátil, com suporte à diversas linguagens, incluindo Java e Python, e importação de modelos em formato ONNX para execução.

Entretanto, o suporte para Java é limitado primeiramente no quesito disponibilidade de bibliotecas. Alguns arquivos binários presentes no pacote possuem licenças incompatíveis com a licença Apache 2, resultado na retirada dos pacotes de Java do repositório Maven. Assim, para utilizar uma versão atualizada da biblioteca é necessário compilar o código-fonte. Além disso, a documentação para Java também é quase vazia, tendo apenas dois tutoriais simples e a publicação da referência da API. Pela dificuldade de instalação e a falta de tutoriais e suporte da comunidade, o uso de MXNet para Java foi descartado.

A versão em Python da biblioteca MXNet é bastante completa, contando com diversos tutoriais para diferentes casos de uso, além de prover uma fácil instalação por meio do gerenciador de pacotes *pip*. Entretanto, em testes iniciais para avaliar a viabilidade de uso da biblioteca, foi descoberto que a biblioteca dispara erros ao tentar executar o modelo usado pelo autor [29] para a comparação de desempenho entre as bibliotecas, logo o uso de MXNet para Python também foi descartado.

3.10 Outras tecnologias

Além das bibliotecas previamente mencionadas, outras bibliotecas foram brevemente avaliadas mas descartadas por motivos desclassificatórios, como Paddle Lite, cuja documentação está disponível apenas em chinês, a biblioteca Caffe2 disponível para Python que foi descontinuada em favor de PyTorch, entre outras não disponíveis para Python ou Java ou sem atual expressão na literatura.

Capítulo 4

Avaliação de Desempenho

Este capítulo tem como objetivo apresentar a metodologia da avaliação de desempenho das diferentes bibliotecas de ML selecionadas.

4.1 Motivação

Avaliações são importantes na busca pelo máximo desempenho de um sistema com os recursos disponíveis. Seus resultados auxiliam tanto nas decisões de escolhas entre diferentes sistemas ou simplesmente entender o funcionamento de um sistema já existente. Devido à grande diversidade de sistemas, não existe um procedimento padrão comum em que seja possível analisar eficientemente um sistema qualquer, sendo necessário conhecer o sistema a ser avaliado e escolher as métricas, carga de trabalho e técnicas de avaliação apropriadas. [49]

Uma simulação executada por simuladores de EONs costuma envolver dezenas de milhares de eventos como requisições de conexões. Tamanha magnitude do número de eventos representa a importância de um bom desempenho na execução de modelos, como por exemplo em propostas de soluções para problemas de alocação de recursos (e.g. RSA) cujos modelos seriam executados em cada evento.

Assim, uma análise quantitativa do desempenho de bibliotecas de ML é importante na busca por uma solução de integração de ML com simuladores que seja flexível de acordo com as necessidades de cada pesquisa e que possua um bom desempenho de modo a acelerar a obtenção de resultados.

4.2 Metodologia

De acordo com Raj Jain [49], há três métodos de avaliação de desempenho: modelagem analítica, simulação e medição. O sistema a ser avaliado, devido à presença de modelos

de redes neurais profundas, é complexo o suficiente para tornar a modelagem analítica inviável. A medição foi descartada pelo fato de não buscar-se uma solução para apenas um simulador e haver um alto custo de implementação para cada possibilidade de integração. Por estes motivos, os resultados serão obtidos por meio de simulações.

4.2.1 Ambiente de Simulação

Para a realização da simulação, 10 programas foram desenvolvidos considerando todas as combinações de linguagens (Java e Python), bibliotecas (ONNX Runtime, Tensorflow Lite, Tensorflow, OpenCV e Deeplearning4j) e o uso ou não da GPU para execução de modelos. Estes programas realizam o mesmo conjunto de tarefas: 1. carregar o modelo e inicializar procedimentos necessários para futuras execuções; 2. carregar a carga de trabalho usada como entrada do modelo; 3. executar o modelo com as entradas carregadas.

Cada programa trata de medir apenas o intervalo de tempo em que a execução do modelo ocorre, sem considerar outros fatores como o tempo de carregamento da carga de trabalho ou do modelo em si. A execução de cada instância de simulação foi orquestrada por *scripts* auxiliares feitos em Python, responsáveis por instalar dependências, compilar e executar os programas de simulação.

As simulações foram realizadas em uma máquina com processador Intel Core i3-8100, placa de vídeo GeForce RTX 2060 e memória RAM de 32 GB (2x16GB 3000Mhz DDR4). O código-fonte dos programas de simulação e de programas auxiliares pode ser encontrado na url tcc.mikaelmello.com.

4.2.2 Modelos e Cargas de Trabalho

Para a simulação, um classificador de estratégias de RSA em EONs, baseado em *deep learning*, será utilizado para comparação. Este modelo recebe como entrada o estado de uma EON e tem como saída a classificação da estratégia RSA em utilização, de acordo com o estado. O modelo possui como saída uma classificação da estratégia de alocação identificada pelo estado como ruim, média ou boa.

A carga de trabalho consiste de 97301 diferentes estados de rede, sendo cada estado representado por uma matriz de 86 linhas e 320 colunas, a representação da topologia USANet com 24 nós e 86 enlaces em que cada enlace contém 320 *slots*.

4.2.3 Métricas

Devido à necessidade de diminuir o tempo de inferência dos modelos como elencado na seção 4.1, esta será a métrica de desempenho avaliada, sendo definida como o tempo total

percorrido desde a chamada do serviço de execução de modelos, o método `classify` da classe `Classifier` na implementação das simulações, até o retorno da chamada.

Além disso, como a execução dos modelos em diferentes bibliotecas exigem algumas conversões do formato de serialização, será avaliado se em alguma das simulações, o resultado de execuções do modelo difere de outras simulações.

4.2.4 Execução

Foram testadas 10 combinações de linguagens, bibliotecas, uso ou não de GPU e formato do modelo executado:

- Java, Deeplearning4j, sem GPU, modelo original no formato HDF5 (TensorFlow);
- Java, Deeplearning4j, com GPU, modelo original no formato HDF5 (TensorFlow);
- Java, ONNX Runtime, sem GPU, modelo convertido para o formato ONNX;
- Java, ONNX Runtime, com GPU, modelo convertido para o formato ONNX;
- Python, OpenCV, sem GPU, modelo convertido para o formato ONNX;
- Python, Tensorflow Lite, sem GPU, modelo convertido para o formato TFLite;
- Python, Tensorflow, sem GPU, modelo original no formato HDF5 (TensorFlow);
- Python, Tensorflow, com GPU, modelo original no formato HDF5 (TensorFlow);
- Python, ONNX Runtime, sem GPU, modelo convertido para o formato ONNX;
- Python, ONNX Runtime, com GPU, modelo convertido para o formato ONNX.

Cada elemento da carga de trabalho foi executado 5 vezes por todas as combinações de simulação, sendo selecionada a mediana de cada uma destas 5 execuções com o objetivo de remover ruídos arbitrários causados pela máquina durante a execução.

4.3 Resultados

Programas Java são popularmente conhecidos pelo seu alto tempo de inicialização que causa a desaceleração da execução de outras tarefas, como as execuções da simulação. Por este motivo, as primeiras 1000 execuções foram descartadas nesta análise geral.

De acordo com a tabela 4.1, a execução do modelo com a biblioteca Deeplearning4j em Java é a mais rápida em média, de modo que a execução sem uso de GPU é em média 300 microsegundos mais rápida que a execução com GPU, porém possui um desvio padrão muito maior devido a diversas limitações deste caso de uso: execuções em CPU estão mais

ling.	biblioteca	GPU		média	desvio padrão	mínimo	máximo
Java	Deeplearning4j	sem	765.89	121.34	722	14620	
		com	1027.66	35.53	990	2180	
	ONNX Runtime	sem	2365.80	153.39	2322	23961	
		com	2067.71	12.62	2041	2383	
Python	ONNX Runtime	sem	2105.99	12.30	2060	2276	
		com	1767.82	10.62	1729	1902	
	OpenCV	sem	1773.15	12.41	1729	1972	
	TensorFlow	sem	2988.95	19.50	2940	3332	
		com	2885.59	18.68	2827	3219	
	TensorFlow Lite	sem	2083.73	12.20	2043	2218	

Tabela 4.1: Estatísticas descritivas acerca do tempo de execução (microsegundos) do modelo em todas as combinações de simulação, descartadas as primeiras 1000 execuções

sujeitas a interrupções externas, como por exemplo o coletor de lixo da JVM ou trocas de contexto do sistema operacional.

A simulação com a biblioteca ONNX Runtime, sendo a única presente em programas Java e Python, foi importante para observar o melhor desempenho de programas Python em versões que utilizam e não utilizam a GPU para inferência. Além disso, a alta variância do tempo de execução na versão em Java com uso da CPU para inferência repete o comportamento visto na execução do Deeplearning4j.

As simulações com bibliotecas da família TensorFlow mostram um desempenho expressivamente maior da biblioteca TensorFlow Lite, uma biblioteca otimizada para inferência de modelos, comparada com TensorFlow.

As simulações de ambas bibliotecas ONNX Runtime e TensorFlow mostram uma melhora de desempenho significativa ao utilizar a GPU da máquina para a execução dos modelos, com ganhos que variam de 100 a 300 microsegundos em média, ao contrário de Deeplearning4j que se mostrou mais devagar com o uso da CPU. Apesar da GPU ser em geral mais rápida na execução de modelos pelas suas características, um dos motivos para este cenário se trata da sobrecarga adicional de tempo ao transferir dados da CPU para a GPU que se torna mais significativa quando o tempo de execução do Deeplearning4j na CPU é em média de duas a três vezes menor comparado com as outras bibliotecas.

Também é possível observar que o tempo de execução para as combinações "Python, ONNX Runtime, sem GPU", "Python, TensorFlow Lite, sem GPU" e "Java, ONNX Runtime, com GPU" é virtualmente igual. Assim como os tempos das combinações "Python, OpenCV, sem GPU" e "Python, ONNX Runtime, com GPU", apesar deste último par possuir tempos menores em 300 microsegundos em média. Após estudos e experimentos

adicionais, não foi possível achar uma explicação satisfatória para a inesperada similaridade entre combinações tão diversas.

Por fim, pode-se concluir que o uso de CPU em programas Java é suscetível a diversas interrupções, gerando uma alta variância no tempo de execução do modelo. O uso de GPU não necessariamente representa um ganho de desempenho, sendo importante observar fatores como a sobrecarga de transferência de dados. Programas Python ou que utilizam a GPU para inferência possuem uma baixa variância de tempos de execuções.

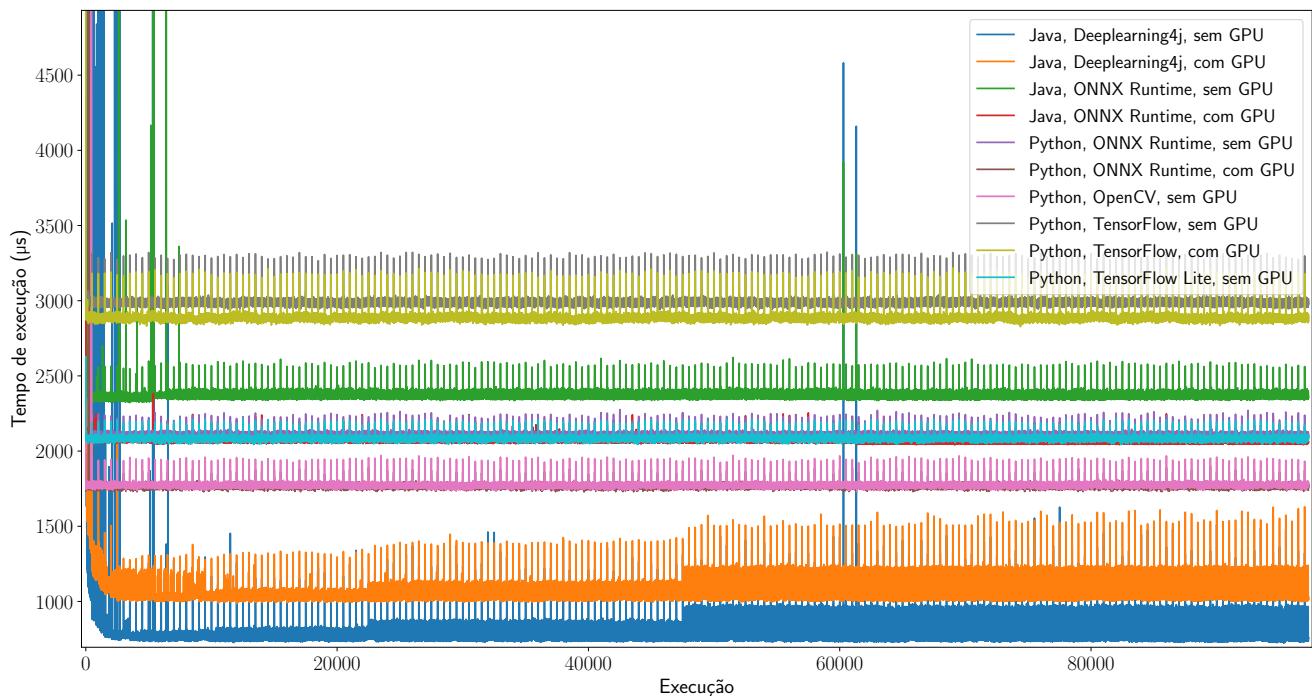


Figura 4.1: Tempo de execução de execuções sequenciais do modelo para cada programa de simulação

Capítulo 5

Resultados

Capítulo 6

Conclusão

Trabalhos futuros: avaliação do treinamento

Referências

- [1] Wilcox, Mark, Stijn Schuermans, Christina Voskoglou e Alexandre Sabolevski: *State of the developer nation 12th edition - q1 2017*, Mar 2017. <https://www.developereconomics.com/resources/reports/state-of-the-developer-nation-q1-2017>. ix, 10
- [2] Le Rouzic, E.: *Network evolution and the impact in core networks*. Em *36th European Conference and Exhibition on Optical Communication*, páginas 1–8, 2010. 1
- [3] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2010–2015*. Growth Lakeland, 2011. 1
- [4] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2011–2016*. Growth Lakeland, 2012. 1
- [5] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2012–2017*. Growth Lakeland, 2013. 1
- [6] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2013–2018*. Growth Lakeland, 2014. 1
- [7] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2014–2019*. Growth Lakeland, 2015. 1
- [8] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2015–2020*. Growth Lakeland, 2016. 1
- [9] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2016–2021*. Growth Lakeland, 2017. 1
- [10] Cisco, T: *Cisco visual networking index: Forecast and methodology, 2017–2022*. Growth Lakeland, 2018. 1
- [11] Ericsson: *Q2 2020 update*. Ericsson Mobility Report, Sep 2020. <https://www.ericsson.com/4a4e5d/assets/local/mobility-report/documents/2020/emr-q2-update-03092020.pdf>. 1
- [12] McKeay, Martin: *The building wave of internet traffic*. Akamai Security Intelligence & Threat Research Blog, Apr 2020. <https://blogs.akamai.com/sitr/2020/04/the-building-wave-of-internet-traffic.html>. 1

- [13] Tomkos, I., S. Azodolmolky, J. Solé-Pareta, D. Careglio e E. Palkopoulou: *A tutorial on the flexible optical networking paradigm: State of the art, trends, and research challenges*. Proceedings of the IEEE, 102(9):1317–1337, 2014. 1, 2
- [14] Jinno, M.: *Elastic optical networking: Roles and benefits in beyond 100-gb/s era*. Journal of Lightwave Technology, 35(5):1116–1124, 2017. 1, 2
- [15] Jorge López Vizcaíno, Yabin Ye e I. T. Monroy: *Energy efficiency in elastic-bandwidth optical networks*. Em *2011 International Conference on the Network of the Future*, páginas 107–111, 2011. 1, 2
- [16] Tomkos, I., E. Palkopoulou e M. Angelou: *A survey of recent developments on flexible/elastic optical networking*. Em *2012 14th International Conference on Transparent Optical Networks (ICTON)*, páginas 1–6, 2012. 2
- [17] Yu, J., B. Cheng, C. Hang, Y. Hu, S. Liu, Y. Wang e J. Shen: *A deep learning based rsa strategy for elastic optical networks*. Em *2019 18th International Conference on Optical Communications and Networks (ICOON)*, páginas 1–3, 2019. 2, 3
- [18] Christodoulopoulos, K., I. Tomkos e E. A. Varvarigos: *Elastic bandwidth allocation in flexible ofdm-based optical networks*. Journal of Lightwave Technology, 29(9):1354–1366, 2011. 2
- [19] Klinkowski, M. e K. Walkowiak: *Routing and spectrum assignment in spectrum sliced elastic optical path network*. IEEE Communications Letters, 15(8):884–886, 2011. 2
- [20] Costa, Lucas R. e André C. Drummond: *Novo algoritmo rmlsa com roteamento multihop em redes Ópticas elásticas*. Em *Anais do XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. SBC, 2016. <http://www.sbrc2016.ufba.br/downloads/SessoesTecnicas/152171.pdf>. 2
- [21] Costa, Lucas R. e André C. Drummond: *Novo esquema para provisão de modulação adaptativa em redes Ópticas elásticas*. Em *Anais do XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Porto Alegre, RS, Brasil, 2017. SBC. <https://sol.sbc.org.br/index.php/sbrc/article/view/2662>. 2
- [22] Costa, Lucas R, Léia S de Sousa, Felipe R de Oliveira, Kaio A da Silva, Paulo JS Júnior e André C Drummond: *Ons: Simulador de eventos discretos para redes ópticas wdm/eon*. Salão de Ferramentas do Simpósio Brasileiro de Redes de Computadores-SBRC, 2016. 2, 3, 4, 9
- [23] Tessinari, Rodrigo Stange, Bart Puype, Didier Colle e Anilton Salles Garcia: *Elastico++: An elastic optical network simulation framework for omnet++*. Optical Switching and Networking, 22:95 – 104, 2016, ISSN 1573-4277. <http://www.sciencedirect.com/science/article/pii/S1573427716300571>. 3
- [24] Aibin, M. e M. Blazejewski: *Complex elastic optical network simulator (ceons)*. Em *2015 17th International Conference on Transparent Optical Networks (ICTON)*, páginas 1–4, 2015. 3, 9

- [25] Pavon-Marino, P. e J. Izquierdo-Zaragoza: *Net2plan: an open source network planning tool for bridging the gap between academia and industry*. IEEE Network, 29(5):90–96, 2015. 3, 9
- [26] Mitchell, Thomas M.: *Machine Learning*. McGraw-Hill, Inc., USA, 1^a edição, 1997, ISBN 0070428077. 3, 6
- [27] Alpaydin, E.: *Introduction to Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2020, ISBN 9780262043793. <https://books.google.com.br/books?id=tZnSDwAAQBAJ>. 3, 6, 7
- [28] Simeone, O.: *A very brief introduction to machine learning with applications to communication systems*. IEEE Transactions on Cognitive Communications and Networking, 4(4):648–664, 2018. 3, 6
- [29] Silva, Guilherme, Lucas R. Costa e André C. Drummond: *Classificador baseado em aprendizado profundo para identificação de estratégias de alocação de espectro em redes Ópticas elásticas*. Em *Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. SBC, 2020. <https://sol.sbc.org.br/index.php/sbrc/article/download/12293/12158>. 3, 15
- [30] Choudhury, P. D. e T. De: *Recent developments in elastic optical networks using machine learning*. Em *2019 21st International Conference on Transparent Optical Networks (ICTON)*, páginas 1–3, 2019. 3
- [31] Zhang, Yongjun, Jingjie Xin, Xin Li e Shangguo Huang: *Overview on routing and resource allocation based machine learning in optical networks*. Optical Fiber Technology, 60:102355, Dec 2020. 3
- [32] Musumeci, F., C. Rottundi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini e M. Tornatore: *An overview on application of machine learning techniques in optical networks*. IEEE Communications Surveys Tutorials, 21(2):1383–1408, 2019. 7
- [33] *Opencv - about.* <https://web.archive.org/web/20201104030111/https://opencv.org/about/>, Accessed: 2020-11-08. 11
- [34] *Darknet: Open source neural networks in c.* <https://web.archive.org/web/20201101011701/https://pjreddie.com/darknet/>, Accessed: 2020-11-08. 11
- [35] *Torch / scientific computing for luajit.* <https://web.archive.org/web/20201101135030/https://torch.ch/>, Accessed: 2020-11-08. 11
- [36] *Onnx.* <https://web.archive.org/web/20201105074026/https://onnx.ai/>, Accessed: 2020-11-08. 11, 14
- [37] *Tensorflow.* <https://web.archive.org/web/20201108012745/https://www.tensorflow.org/>, Accessed: 2020-11-08. 11
- [38] PyTorch: *What is pytorch?* https://pytorch.org/tutorials/beginner/blitz/tensor_tutorial.html#what-is-pytorch. 11

- [39] Gugger, Sylvain e Jeremy Howard: *Practical deep learning for coders*. <https://web.archive.org/web/20201103183225/https://course.fast.ai/>, Accessed: 2020-11-08. 11
- [40] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay: *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12:2825–2830, 2011. 12
- [41] Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu e Xiaoqiang Zheng: *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. <https://www.tensorflow.org/>, Software available from tensorflow.org. 12
- [42] *Getting start tutorial - issue #68 - tensorflow/java*, Oct 2020. <https://github.com/tensorflow/java/issues/68#issuecomment-704622036>. 13
- [43] Chollet, François *et al.*: *Keras*. <https://keras.io>, 2015. 13
- [44] *Tensorflow lite / ml for mobile and edge devices*. <https://web.archive.org/web/20201104084527/https://www.tensorflow.org/lite>, Accessed: 2020-11-08. 13
- [45] *Is tf lite optimized for nvidia gpu's and intel cpus?* <http://web.archive.org/web/20201108062538/https://github.com/tensorflow/tensorflow/issues/34536>, Accessed: 2020-11-08. 13
- [46] *Onnx runtime*. <https://web.archive.org/web/20201107081602/https://microsoft.github.io/onnxruntime/>, Acessed: 2020-11-10. 14
- [47] *Deeplearning4j*. <https://web.archive.org/web/20201103204920/https://deeplearning4j.org/>, Accessed: 2020-11-08. 14
- [48] *Apache mxnet / a flexible and efficient library for deep learning*. <https://web.archive.org/web/20201108093437/https://mxnet.apache.org/>, versions/1.7.0/, Accessed: 2020-11-10. 15
- [49] Jain, R.: *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1991, ISBN 978-0-471-50336-1. 16