Identificação de Períodos Críticos na Regularização de Redes Neurais

Prof. Dr. Artur Jordão Lima Correia

Departamento de Engenharia de Computação e Sistemas Digitais

Escola Politécnica da Universidade de São Paulo

Resumo

Modelos modernos de aprendizado de máquina vêm adotando redes neurais como o consenso atual para resolver tarefas cognitivas. Apesar dos resultados sem precedentes, modelos profundos compartilham o problema de como garantir generalização a novos dados. Técnicas de regularização mitigam essa questão impondo restrições durante o processo de aprendizado. Nessa direção, estudos recentes confirmaram a existência do fenômeno denominado *período críticos*: fases iniciais do treinamento onde técnicas de regularização possuem efeito positivo na generalização do modelo. Após tais fases, técnicas de regularização promovem pouco ou nenhum benefício. Enquanto esses estudos fornecem teoria para compreender a efetividade da regularização durante o treino, existe uma lacuna no sentido de identificar *quando* períodos críticos emergem. Este projeto de pesquisa pretende preencher tal lacuna identificando sistematicamente e interrompendo a regularização após períodos críticos. Para este propósito, o projeto planeja aplicar métricas relativamente recentes desenvolvidas para capturar alterações na dinâmica do treinamento. Particularmente, este projeto concentra-se na regularização por meio do aumento de dados que, apesar de fornecer vários benefícios, aumenta o tempo de treinamento, a quantidade de emissão de CO_2 e o consumo de energia devido à maior quantidade de dados. Desta forma, dentre as contribuições esperadas com o projeto, destacamos uma maior eficiência em ambas as fases de treinamento e ajuste fino de modelos modernos.

Palavras-chave: Redes Neurais, Aprendizado Profundo, Períodos Críticos, *Green AI*.

1 Introdução

Redes neurais modernas têm levado a resultados sem precedentes em diversas atividades cognitivas como visão computacional e processamento de linguagem natural. Seguindo o consenso atual, redes neurais cada vez mais profundas e super parametrizadas tendem a fornecer melhores resultados devido às suas habilidades em capturar padrões discriminativos a partir os dados. Apesar dos resultados promissores, frequentemente, modelos desta família estão confinados a uma alta sobrecarga computacional, o que impõe diversos desafios tecnológicos, finan-

ceiros e organizacionais para seu desenvolvimento e estudo. Dentre esses obstáculos, o tempo de treinamento destaca-se como um dos maiores limitadores para o desenvolvimento desses modelos em cenários com recursos computacionais escassos.

Além das questões de custo computacional, redes neurais modernas compartilham um problema central envolvendo sua fase de treinamento: garantir um bom desempenho não apenas nos dados de treinamento, mas também em novos dados (dados de teste, na prática, nunca vistos). Essa questão torna-se fundamental uma vez que redes super parametrizadas (*overparameterized* – regime no qual o número de parâmetros do modelo supera o número de amostras de treinamento) possuem a "habilidade"de memorizar dados aleatórios (mais concretamente, dados com rótulos aleatórios) [1]. Nessa direção, mecanismos de regularização surgem como estratégias eficazes para mitigar o problema. Em linhas gerais, regularização consiste em qualquer modificação para reduzir o erro de generalização (conjunto de teste) sem afetar o erro de treinamento.

Dentre as formas populares de regularização em aprendizado profundo, podemos destacar a penalização dos pesos dos neurônios (*weight decay*), que diretamente modifica a função de perda, e normalizações na distribuição das respostas dos neurônios nas camadas [2, 3]. Outra forma poderosa e efetiva de regularização corresponde à técnica de aumento de dados (*data augumentation*). Tal técnica consiste em expandir o conjunto de treinamento criando réplicas dos dados a partir de modificações em seu conteúdo [4, 5]. Particularmente, além de questões de generalização, técnicas recentes de aumento de dados produzem modelos mais robustos a ataques adversariais [5]. Devido a essas vantagens, diversos mecanismos de aumento de dados vêm sendo propostos na literatura, em particular, no contexto de visão computacional e processamento de linguagem natural [5, 7]. Adicionalmente, avanços no paradigma de *foundation models* – grandes modelos de aprendizagem capazes de efetivamente transferir seu conhecimento para novas tarefas – intensificaram ainda mais pesquisas em aumento de dados, uma vez que o sucesso por trás desses modelos reside no treinamento usando uma ampla gama de dados [6, 7]. Apesar desses benefícios, o tempo de treinamento aumenta proporcionalmente em razão do maior volume de dados. Além disso, outros fatores também se intensificam como, por exemplo, o custo financeiro e a quantidade de emissão de CO_2 devido ao consumo de energia.

Independentemente da forma de regularização, estudos recentes observaram que somente nas fases iniciais do treinamento mecanismos de regularização possuem efeito positivo na habilidade preditiva do modelo [8, 9, 10]. Especificamente, estes trabalhos confirmam a existência de *períodos críticos* durante as fases iniciais do aprendizado, para o qual mecanismos de regularização são mais efetivos. Após essas fases, a regularização fornece pouco ou nenhuma vantagem. Enquanto esses estudos fornecem arcabouços teóricos fundamentais para compreender a efetividade da regularização durante a dinâmica do treino, nenhum deles fornece uma maneira sistemática para identificar períodos críticos. Portanto, uma questão natural que surge é *como identificar períodos críticos na regularização*?

Responder o questionamento acima possibilita reduzir notavelmente a sobrecarga computacional durante o treinamento imposta pela regularização por aumento de dados. Este projeto pretende explorar tal questão identificando períodos críticos aplicando métricas relativamente recentes desenvolvidas para capturar alterações na dinâmica do treinamento (detalhes na Seção 5) [11, 12, 13]. Especificamente, o projeto planeja reduzir o problema de identificar períodos críticos a uma questão de mensurar alterações na qualidade da dinâmica do treinamento. Acreditamos que essas alterações podem capturar *quando* o efeito da regularização começa a perder eficácia.

2 Objetivos

De uma perspectiva teórica e metodológica, os objetivos deste projeto são os seguintes. Pesquisar, elaborar e desenvolver soluções capazes de sistematicamente identificar períodos críticos. Nesta direção, vale reforçar que o objeto de pesquisa deste projeto não consiste em criar mecanismos novos de regularização e sim encontrar períodos críticos onde a regularização gera efeitos positivos no aprendizado do modelo. Analisar a dinâmica do treinamento e o comportamento da habilidade preditiva do modelo ao aplicar as soluções desenvolvidas. Apontar e mitigar eventuais desvantagens das soluções desenvolvidas.

De uma perspectiva prática, os objetivos deste projeto correspondem aos seguintes pontos. Reduzir o custo computacional no treinamento inerente à regularização por aumento de dados. Reduzir o custo financeiro e consumo energético (consequentemente a quantidade de CO_2 emitida) para treinar redes neurais em diferentes tarefas de reconhecimento de padrões.

3 Contribuições e Resultados Esperados

As contribuições esperadas com o desenvolvimento deste projeto são as seguintes. Promover maior eficiência no custo de treinamento de redes neurais reduzindo o tempo de treinamento requerido por modelos modernos. Tornar mais eficiente o processo de ajuste fino de redes neurais em cenários com restrições de infraestrutura e ambientes com recursos limitados. Reduzir a quantidade de dióxido de carbono decorrente do desenvolvimento de redes neurais. Estimular estudos rumo à IA sustentável (Green AI).

Do ponto de vista teórico, o projeto espera contribuir com os seguintes tópicos. Fornecer maior entendimento das fases críticas da regularização, isto é, as fases onde formas de regularização (particularmente, aumento de dados) são mais efetivas. Apontar quais técnicas são mais adequadas para capturar a qualidade do aprendizado das redes neurais e alterações na dinâmica do treinamento. Demonstrar como a remoção de regularização afeta a generalização e robustez adversarial das redes neurais.

4 Questões de Pesquisa

Durante o desenvolvimento deste projeto, pretende-se explorar as seguintes questões de pesquisa. Quais são as técnicas mais promissoras para mensurar a qualidade do treinamento e alterações na dinâmica do treinamento? Técnicas agnósticas a dados (ex. as que consideram somente os parâmetros do modelo) são as melhores escolhas? Como a remoção de regularização afeta a generalização das redes neurais? Existe uma relação entre a complexidade (ex. profundidade) da rede e as épocas de períodos críticos? Qual é o papel da regularização na robustez a ataques adversariais em modelos de aprendizado profundo? Mais diretamente relacionado ao escopo deste projeto, como a remoção da regularização afeta os modelos em termos de ataques adversariais. Existe um compromisso entre reduzir o custo computacional (eliminando a regularização) e a habilidade preditiva final do modelo?

5 Formalismos e Definição do Problema

Considere $X \in \mathbb{R}^{n \times m}$ um conjunto de n amostras descritas a partir de m atributos, e $Y \in \mathbb{R}^{n \times 1}$ as categorias (rótulos) associadas a cada amostra de X. Seja $\mathcal{F}(\cdot, \cdot)$ uma rede neural que será treinada usando o paradigma supervisionado sob os dados X e os rótulos Y. Defina θ_i os parâmetros de $\mathcal{F}(\cdot, \cdot)$ que serão ajustados por um processo iterativo de otimização (ex. *Stochastic Gradient Descent* – SGD), onde i representa uma iteração (época) deste processo.

Durante o processo iterativo de otimização, para melhorar a generalização de um modelo $\mathcal{F}$ uma forma comumente aplicada na literatura são mecanismos de regularização e aumento de dados. Particularmente, este projeto prioriza explorar técnicas de aumento de dados, formalizada da seguinte maneira. Seja $T(\cdot)$ uma função que recebe um conjunto de amostras X e modifica seu conteúdo produzindo um novo conjunto do mesmo tamanho. Tipicamente, métodos recentes de aumento de dados possuem elementos estocásticos [4, 5], possibilitando criar conjuntos de dados arbitrariamente grandes ao se aplicar a função T várias vezes. Formalmente, podemos aumentar o conjunto original de dados aplicando k vezes a função T, isto é, $\{(T(X), Y)\}^k$. Visando simplificar a nomenclatura, defina $\mathcal{D} = (X, Y)$ como o par de amostras e seus respectivos rótulos; assim, podemos reescrever o aumento de dados k vezes como $\{\mathcal{D}\}^k$.

Aplicando técnicas de aumento de dados, podemos formalizar a atualização dos parâmetros θ seguindo

$$\theta_{i+1} = \theta_i - \eta \frac{1}{B} \sum_{b=1}^{B} \nabla \mathcal{L}(\{\mathcal{D}\}_b^k, \theta_i), \tag{1}$$

onde $\mathcal{D}_b^k$ indica um *batch* de b amostras dos dados aumentados, $\nabla \mathcal{L}$[1] corresponde ao gradiente da função de

[1] É possível reescrever o gradiente da seguinte maneira: $\mathcal{L}(\{(Y_b, F(T(X_b)))\}^k, \theta_i)$, onde X_b e Y_b são *batches* de dados e seus correspondentes rótulos.

erro em relação ao parâmetros θ_i e η indica a magnitude da atualização (taxa de aprendizado).

A partir do formalismo acima, o objeto de pesquisa deste projeto consiste em encontrar uma época de treinamento i de tal forma que seja possível reduzir a quantidade de aumento de dados, isto é, diminuir o valor k (idealmente utilizar $k = 1$). Vale reforçar que a literatura confirmou a possibilidade de reduzir o valor de k; entretanto, existe uma lacuna de pesquisas para definir quando reduzi-lo. Para este propósito, o projeto planeja explorar métricas relativamente recentes que capturam a dinâmica do treinamento e alterações dos pesos [11, 12, 13]. Tais métricas são descritas a seguir.

Confusão no Gradiente (CG). Sejam A e B o gradiente de dois *batches* de dados distintos de determinada época de treinamento i. Sankararaman et al. [11] define a confusão do gradiente entre esses *batches* em termos de

$$CG(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \in [-1, 1]. \tag{2}$$

Previsibilidade do Gradiente. A confusão no gradiente descrita acima leva em consideração *batches* de dados da mesma época i de treinamento. Diferentemente, a métrica de previsibilidade do gradiente, explorada no trabalho de Chen et al. [13], compara o gradiente de um *batch* da época i com o gradiente de um passo posterior (futuro) $i + 1$.

Rotação de Camadas. Esta métrica calcula a distância cosseno entre os parâmetros de uma determinada época i, θ_i, com sua inicialização original (aleatória) θ_0. Portanto, em termos da Equação 2, podemos descrever a rotação de camadas atribuindo $A = \theta_0$ e $B = \theta_i$, com $i \neq 0$.

Neste projeto, a ideia por trás das métricas acima consiste em capturar estabilidade na convergência da dinâmica do treinamento. Especificamente, quando determinada métrica não apresentar variações acima de um limiar pre-definido entre duas (ou mais) épocas, pretendemos reduzir o valor de k na Equação 1; desta forma, eliminando/reduzindo a regularização por aumento de dados.

Finalmente, vale destacar que as métricas descritas são agnósticas a dados, isto é, consideram somente os pesos, θ_i, do modelo. Acreditamos que essa é uma linha mais promissora de pesquisa, uma vez que não existe dependência de um conjunto de validação com amostras de qualidade (ex. que seguem a mesma distribuição do treinamento, isto é, independentes e identicamente distribuídas – i.i.d) como requerido, por exemplo, pela técnica de *early-stopping*.

6 Metodologia

Para alcançar os objetivos e responder a questões de pesquisa, o projeto adota o seguinte planejamento metodológico. Explorar métricas capazes de quantificar a qualidade do aprendizado de redes neurais. A partir da aplicação dessas técnicas, quantificar o impacto da remoção de regularização analisando o compromisso entre

performance computacional e habilidade preditiva da rede obtida. Desenvolver algoritmos capazes de identificar equilíbrios adequados entre sustentabilidade, eficiência e desempenho preditivo. Mensurar o tempo de treinamento necessário após reduzir o aumento de dados, estimando a emissão de CO_2, o consumo energético e financeiro.

As questões de pesquisa do projeto poderão ser validadas em diferentes áreas do conhecimento que utilizam redes neurais para apoiar tomadas de decisão: visão computacional, geologia (predição da produção de óleo), medicina diagnóstica (reconhecimento de atividades), dentre outras. As bases de dados para essas aplicações estão publicamente disponíveis e não envolvem questões éticas para uso e desenvolvimento.

As redes neurais consideradas serão modelos populares da literatura de aprendizado profundo, tais como redes residuais (ResNet) e os Transformers. Para mensurar a habilidade preditiva dos modelos, serão utilizadas as métricas padrões da aplicação em questão. Em relação às métricas que quantificam o desempenho computacional, serão utilizadas algumas bem estabelecidas na literatura do tema e que seguem recomendações definidas por estudos anteriores. Acompanhando tendências recentes [14, 15], também será considerada a emissão de CO_2 como uma métrica quantitativa da qualidade dos resultados obtidos. Nesta direção, existem ferramentas onlines e gratuitas.

As atividades deste projeto serão desenvolvidas na linguagem de programação Python que fornece suporte simples e fácil para o desenvolvimento de redes neurais. Além disso, como a literatura de redes neurais utiliza predominantemente Python, torna-se fácil a reprodutibilidade e aprimoramento de métodos modernos e do estado da arte. Por fim, não existem restrições ou questões éticas relevantes que devam ser consideradas para o compartilhamento de dados. Todos os dados envolvidos podem ser compartilhados sem nenhuma restrição.

7 Pré-requisitos e Outras Informações

Para o desenvolvimento do projeto, o aluno precisa ter conhecimento prévio na linguagem de programação Python. Não é necessário que o candidato tenha experiência em reconhecimento de padrões, aprendizado de máquina ou redes neurais.

Os dados que serão utilizados neste projeto estão disponíveis publicamente e as considerações éticas sobre os dados são de responsabilidade dos autores/pesquisadores que coletaram as bases de dados. Desta forma, este projeto não envolve questões éticas e está em conformidade legal. Finalmente, a utilização desses dados disponíveis publicamente permite transparência e reprodutibilidade na pesquisa.

8 Cronograma de Execução

A Tabela 1 detalha o cronograma, em bimestres, das atividades envolvidas no projeto.

Tabela 1: Cronograma das atividades em bimestres.

Atividade	1	2	3	4	5	6
Leitura de artigos científicos	X	X	X	X	X	X
Seleção das aplicações e base de dados	X					
Implementação dos Métodos (Identificar períodos críticos)	X	X	X			
Treinamento das redes neurais	X	X	X			
Execução dos experimentos	X	X	X	X	X	
Documentação dos códigos			X			X
Escrita de relatório parcial			X			
Escrita de artigo científico			X	X	X	X
Relatório Final						X

Referências

[1] P. Maini, M. C. Mozer, H. Sedghi, Z. C. Lipton, J. Z. Kolter, and C. Zhang, "Can neural network memorization be localized?" in *International Conference on Machine Learning (ICML)*, 2023.

[2] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015.

[3] R. Burkholz, "Batch normalization is sufficient for universal function approximation in CNNs," in *International Conference on Learning Representations (ICLR)*, 2024.

[4] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Neural Information Processing Systems (NeurIPS)*, 2020.

[5] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "Pixmix: Dreamlike pictures comprehensively improve safety measures," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] R. Bommasani, D. A. Hudson, E. Adeli, and et al., "On the opportunities and risks of foundation models," *ArXiv*, 2021.

[7] X. Amatriain, A. Sankar, A. Sankar, P. K. Bodigutla, T. J. Hazen, and M. Kazi, "Transformer models: an introduction and catalog," *arXiv*, 2023.

[8] A. Golatkar, A. Achille, and S. Soatto, "Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence," in *Neural Information Processing Systems (NeurIPS)*, 2019.

[9] M. Kleinman, A. Achille, and S. Soatto, "Critical learning periods for multisensory integration in deep networks," in *Conference on Computer Vision and Pattern Recognition (CVPR*, 2023.

[10] ——, "Critical learning periods emerge even in deep linear networks," in *International Conference on Learning Representations (ICLR)*, 2024.

[11] K. A. Sankararaman, S. De, Z. Xu, W. R. Huang, and T. Goldstein, "The impact of neural network overparameterization on gradient confusion and stochastic gradient descent," in *International Conference on Machine Learning (ICML)*, 2020.

[12] S. Carbonnelle and C. D. Vleeschouwer, "Layer rotation: a surprisingly simple indicator of generalization in deep networks?" in *International Conference on Machine Learning (ICML) Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

[13] Y. Chen, A. L. Yuille, and Z. Zhou, "Which layer is learning faster? A systematic exploration of layer-wise convergence rate for deep neural networks," in *International Conference on Learning Representations (ICLR)*, 2023.

[14] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," in *NeurIPS*, 2019.

[15] A. Faiz, S. Kaneda, R. Wang, R. C. Osi, P. Sharma, F. Chen, and L. Jiang, "LLMCarbon: Modeling the end-to-end carbon footprint of large language models," in *ICLR*, 2024.