UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

ISABELLA JESKE ROSA

Uma análise comparativa de métricas de generalização de redes neurais profundas em imagens médicas

Campinas
24/11/2023

ISABELLA JESKE ROSA

Uma análise comparativa de métricas de generalização de redes neurais profundas em imagens médicas

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. João Florindo.

Resumo

A tarefa de compreender e prever a performance e generalização de modelos de rede neural profunda tem sido explorada amplamente, em especial, através de métricas de complexidade. Observa-se ainda que dentro desse grande problema existem subtemas que apresentam os seus próprios desafios e precisam ser estudados à parte, como é o caso de modelos feitos para aplicações em imagens médicas. Após a análise de alguns estudos que avaliam múltiplas métricas em diferentes contextos, como Jiang et al. [2019] e Vakanski and Xian [2021], criou-se a concepção de um estudo inspirado e adaptado, com a finalidade de enxergar de forma controlada as diferenças entre a relação métrica-generalização para os contextos gerais e médico. Como resultado da experimentação, muitas métricas de complexidade demonstram ter comportamento diferente quando se tratam de modelos com imagens médicas, e poucas obtiveram índices de correlação com a generalização altos e relevantes; em especial, destacaram-se em ambos os casos as métricas baseadas em *PAC-Bayes* e nas características do ponto ótimo local encontrado, embora apontem para direções diferentes, o que implica que as individualidades de cada problema impactam tanto a performance quanto o uso.

Abstract

Comprehending and predicting the performance and generalization of deep neural network models has been widely explored, especially, through complexity measures. The subject shows that there are subtopics with different behaviors that must be studied separately, such as medical image models. After the review of existing articles about the topic, such as Jiang et al. [2019] and Vakanski and Xian [2021], the conception of an adapted and inspired study was created, to find the differences between the relationship of complexity measures and generalization for the generic e medical images, in a controlled way. By the end of the experiment, many complexity measures had shown different behaviors when using medical images, and few had correlation indices with generalization high or relevant enough; Especially, the PAC-Bayes and Flatness-based metrics stood-out, although they pointed to different directions, implying that each individual scenario may impact in both performance and usage of metrics.

Conteúdo

1 Introdução

O universo do aprendizado de máquinas tem se tornado indiscutivelmente popular. Em especial, as redes neurais profundas surpreendem com seu alto desempenho. Ao se falar de performance, entretanto, é crucial incluir quão bem o modelo preditivo generaliza, ou seja, se adapta a novos dados e mantém seu desempenho estável; algoritmos usuais de *machine learning* mais complexos podem, por exemplo, "decorar" os dados em que foram treinados, fazendo com que pequenas oscilações nos dados tornem sua performance instável. Com o crescente uso de algoritmos de *deep learning* modernos, a difícil explicabilidade do porquê redes neurais profundas generalizam bem, apesar de sua enorme complexidade, é uma área de pesquisa ainda em aberto (Dinh et al. [2017]).

Neste cenário, diversas medidas têm sido apresentadas com o objetivo de compreender a capacidade de generalização de um modelo, chamados de métricas de complexidade. Com propostas teóricas e empíricas, as métricas podem depender das propriedades do modelo treinado, como parâmetros ou estrutura da rede, otimizador escolhido e até dos dados de treino. Por descreverem de formas diversas uma rede neural treinada, supõe-se que as métricas tenham uma relação monotônica com a generalização, relação esta que vem sido estudada em diferentes âmbitos. Em Jiang et al. [2019] são realizados experimentos de larga escala, ao criar um grande conjunto de modelos com pequenas diferenças, e explorar afundo a relação entre a generalização e as métricas calculadas em cada um eles, utilizando na análise bases bem fundamentadas e com imagens variadas e diversas, como a CIFAR-10 e a SVHN (Krizhevsky and Hinton [2009], Yuval et al. [2011]), tendo sido avaliadas mais de 40 métricas diferentes.

Há ainda, subtemas mais específicos e com características próprias, como os modelos voltados a imagens médicas. Esse tipo de conjunto de dados costuma ter menor volume e imagens de maior resolução; isso implica em um comportamento de convergência possivelmente diferente durante a otimização, segundo Vakanski and Xian [2021]. Além disso, as imagens coletadas podem ser de diferentes hospitais, regiões, condutas de procedimento etc., interferindo na distribuição dos dados e fazendo com que a generalização se torne mais complicada. Vakanski and Xian [2021] explora o tema inspirando-se nas análises de Jiang et al. [2019] e readaptando em menor escala alguns experimentos, utili-

zando dados de ultrassom de mamas como o BUSIS (Xian et al. [2018]) e BUS-Combined (Xian et al. [2018];Al-Dhabyani et al. [2020];Hoon et al. [2017];Ult).

Ambos os artigos citados acima apresentam as métricas mais promissoras em seus contextos e, embora Vakanski and Xian [2021] realize uma comparação entre seus resultados e os de Jiang et al. [2019], as metodologias dos experimentos são distantes. Isso torna difícil entender se os diferentes resultados de métricas são devidos a características do aparato experimental ou se de fato funcionam melhor ou pior no caso médico; no projeto anterior a este (Rosa and Florindo [2023]) é possível ver uma análise comparativa aprofundada sobre estudos existentes.

2 Desenvolvimento

2.1 Objetivos e metodologia

Dado o objetivo de comparar o desempenho das métricas entre os casos geral e médico de forma justa, foi realizado um experimento inspirado em Jiang et al. [2019] e Vakanski and Xian [2021]. A relação métricas de complexidade-generalização é avaliada alterando apenas os dados utilizados, mantendo em cada caso (genérico e médico) uma metodologia que, apesar de simples, é idêntica em ambos, possibilitando a comparação de resultados de maneira adequada.

A base de dados utilizada para aferir o desempenho das métricas no caso geral é a CIFAR-10, mesma utilizada por Jiang et al. [2019]. Ela contém 60000 imagens 32x32 (pequenas) RGB com 10 classes balanceadas. As 10 classes correspondem a temas diversos, como meios de transporte e animais, como no exemplo abaixo:

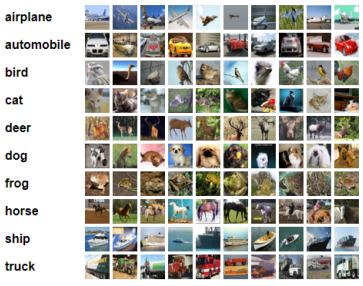

Figura 1: Amostra de imagens dos conjuntos de dados CIFAR-10

Já para o caso médico, optou-se pela utilização da BreastMNIST, uma base de

dados de ultrassom de mamas da biblioteca MedMINIST (Yang et al. [2021]), que reúne e pré-processa diversos conjuntos de dados de caráter médico, facilitando o acesso e uso. A BreastMNIST conta com 780 imagens em tons de cinza 28×28, que é um tamanho pequeno levando-se em conta que as imagens originais - assim como grande parte das imagens em contexto médico - têm tamanho e resolução maiores.

Figura 2: Amostra de imagens dos conjuntos de dados BreastMNIST

A fim de comprovar o desempenho de uma métrica de complexidade ao "prever" a generalização de um modelo, é necessário compreender a relação entre os dois valores, evitando possíveis vieses ao tentar simular casos reais. É importante definir-se uma boa configuração inicial para os modelos, e como de fato é avaliada essa relação; neste estudo, são utilizados os mesmos conceitos de generalização que em Jiang et al. [2019] e Vakanski and Xian [2021], caracterizados da seguinte forma:
Seja h o preditor formado pela rede neural treinada, $X_i \in \chi$ os dados de entrada, $y_i \in 1, ...\kappa$ as classes resposta correspondentes, e n o número de amostras, definimos:

1. Risco do Modelo $L(h)$ como o risco do preditor, aproximado pelo erro de teste.

$$L(h) \approx \frac{1}{n} \sum_{i=1}^{n} h(X_i) \neq y_i$$

2. Risco Empírico $\hat{L}(h)$ como o erro observado no treino.

3. *Gap* de Generalização $\Delta L(h)$ como a diferença entre o Risco Empírico observado no treino e o Risco do modelo, aproximado pelo teste.

$$\Delta L(h) = L(h) - \hat{L}(h)$$

Desta forma, calculamos a Generalização como uma aproximação baseada no erro do modelo observado no treino e no teste - simulando o erro em dados não vistos.

Para a realização dos experimentos, são gerados N modelos variando hiperparâmetros comumente relacionados à performance e generalização, como número de blocos de convolução, tamanho do *batch* e taxa de aprendizado. Foram testados três quantidades de blocos de convolução (3, 4, 5), três taxas de aprendizado (0.00032, 0.0001, 0.000032) e dois tamanhos de *batch* (128 e 64), criando assim 18 preditores diferentes. Cada modelo foi treinado por uma rede com arquitetura VGG, a mesma rede utilizada por Vakanski and Xian [2021], ao longo de 2000 épocas ou até o custo no treino atingir 0,01; modelos que não atingiram o critério de custo não foram utilizados na análise.

2.2 Métricas de Complexidade

Foram avaliadas 25 métricas de complexidade derivadas de grandes grupos relacionados a qual informação do modelo é utilizada para o cálculo, como dimensão, saída, parâmetros etc., como explicitado abaixo. Todas as anotações e equações são retiradas do Apêndice de Vakanski and Xian [2021], sendo uma descrição completa presente em Jiang et al. [2019].

2.2.1 Métricas *VC-Dimension-based*

Baseada no número de parâmetros de uma rede convolucional profunda. Dada uma rede com d camadas, para uma camada i, com c_i filtros de convolução e tamanho de *kernel* $k_i X k_i$, temos:

$$\mu_{num\_params} = \sum_i k_i^2 c_{i-1}(c_i + 1) \tag{1}$$

2.2.2 Métricas *Output-based*

A margem inversa é "*output-based*" pois depende dos dados de treino e saídas do modelo. A margem γ é o 10 percentil da distribuição de diferenças nas logits das saídas do modelo sob os dados de treino.

$$\mu_{inverse\_margin} = \frac{1}{\gamma^2} \tag{2}$$

2.2.3 Métricas *Norm-based*

Seja o tensor de pesos de um modelo treinado na i- ésima camada chamado $\mathbf{W}_i$, e o tensor de pesos iniciais na mesma camada $\mathbf{W}_i^0$. A notação $||.||_F$ indica a norma de Frobenius (ou Euclidiana).

$$\mu_{frob\_dist} = \sum_i ||\mathbf{W}_i - \mathbf{W}_i^0||_F^2 \tag{3}$$

$$\mu_{param\_norm} = \sum_i ||\mathbf{W}_i||_F^2 \tag{4}$$

Seja $\mathbf{w}^2$ os pesos quadráticos em que $\mathbf{w} = vec(\mathbf{W}_1, ..., \mathbf{W}_d)$, e as entradas vetores de uns, $h_{\mathbf{w}^2}(\mathbf{1})$. Temos:

$$\mu_{path\_norm} = \sum_i h_{\mathbf{w}^2}(\mathbf{1})[i] \tag{5}$$

$$\mu_{path\_norm\_over\_margin} = \sum_i \frac{h_{\mathbf{w}^2}(\mathbf{1})[i]}{\gamma^2} \tag{6}$$

2.2.4 Métricas *Flatness-based*

O conceito de "*flatness*" investiga o quão "plano" o ótimo local encontrado é. Fundado na Teoria PAC-Bayes, ao se adicionarem perturbações Gaussianas $\mathbf{u} \sim \mathcal{N}(0, \sigma^2\mathcal{I})$ aos pesos do modelo treinado, observa-se a variância σ^2 a fim de encontrar o valor que faz com que o risco empírico reduza em pelo menos 10%. As métricas *Magnitude-aware* aplicam perturbações ajustadas de acordo com a magnitude dos pesos $\mathbf{u} \sim \mathcal{N}(0, \sigma'^2|w_i|^2 + \epsilon^2)$. Sendo ϵ fixado em 10^{-3} e ω denotando o número de pesos, temos:

$$\mu_{pacbayes\_flatness} = 1/\sigma^2 \tag{7}$$

$$\mu_{pacbayes\_mag\_flatness} = 1/\sigma'^2 \tag{8}$$

$$\mu_{pacbayes\_mag\_init} = \frac{1}{4}\sum_i log(\frac{\epsilon^2 + (\sigma'^2+1)\frac{||\mathbf{w}-\mathbf{w}^0||_2^2}{\omega}}{(\epsilon^2 + \sigma'^2|\omega_i - \omega_i^0|^2)}) + log((m+2)/\delta) \tag{9}$$

$$\mu_{pacbayes\_mag\_orig} = \frac{1}{4}\sum_i log(\frac{\epsilon^2 + (\sigma'^2+1)\frac{||\mathbf{w}||_2^2}{\omega}}{(\epsilon^2 + \sigma'^2|\omega_i - \omega_i^0|^2)}) + log((m+2)/\delta) \tag{10}$$

Para, por fim, quantificar a relação existente entre cada métrica e a generalização dos modelos, há diversas formas discutidas por Jiang et al. [2019], enquanto a escolhida aqui é o índice de correlação, por apontar mais claramente a natureza da relação. Como resultado dos experimentos realizados, tem-se um conjunto de modelos gerados através de um espaço de hiper-parâmetros Θ. Cada preditor $h(\theta)$ tem seu respectivo *gap* de generalização $\{g(\theta)|\theta \in \Theta\}$ e métricas $\{\mu(\theta)|\theta \in \Theta\}$; criando-se um conjunto $\mathcal{T}$, em que cada elemento é um par formado por

$$\mathcal{T} \triangleq \cup_{\theta\in\Theta}\{(\mu(\theta), g(\theta))\}$$

Como índice de correlação, é utilizado o **Kendall's Rank-Correlation Coefficient**, ou Coeficiente de Kendall τ (KENDALL [1938]), que considera o grau de consistência entre elementos do conjunto $\mathcal{T}$; considera-se consistência quando para qualquer par de $\mathcal{T}$, se $\mu(\theta_1) > \mu(\theta_2)$ então $g(\theta_1) > g(\theta_2)$, da seguinte forma:

$$\tau(\mathcal{T}) = \frac{1}{|\mathcal{T}|(|\mathcal{T}|-1}\sum_{(\mu_1,g1)\in\mathcal{T}}\sum_{(\mu_2,g2)\in\mathcal{T}/(\mu_1,g1)} sign(\mu_1-\mu_2)sign(g_1-g_2)$$

O coeficiente pode variar entre -1 e 1, e se baseia no número de concordâncias e discordâncias entre os pares de elementos de $\mathcal{T}$; 1 representa perfeita concordância e -1 o oposto.

2.3 Resultados e Discussão

Durante o treino dos modelos, devido ao critério de convergência do modelo até o custo de $0,01$, obtiveram-se ao fim, 16 modelos para a BreastMNIST e 17 para a CIFAR, com acurácias no teste em torno de 70 a 85%, variando principalmente com o número de blocos de convolução e taxa de aprendizado. Na tabela abaixo, é possível verificar quais foram os índices de correlação de Kendall entre cada métrica de complexidade e a generalização dos modelos treinados na base médica (BreastMNIST) e na base geral (CIFAR).

Métricas	BreastMNIST	Cifar
VC dimension		
PARAMS	-0,147	0,322
Output Measure		
INVERSE_MARGIN	0,000	0,226
Spectral-norm Measures		
LOG_PROD_OF_SPEC	-0,088	-0,466
LOG_PROD_OF_SPEC_OVER_MARGIN	-0,123	-0,391
LOG_SUM_OF_SPEC	-0,035	-0,587
LOG_SUM_OF_SPEC_OVER_MARGIN	-0,035	-0,496
LOG_SPEC_INIT_MAIN	-0,141	-0,316
LOG_SPEC_ORIG_MAIN	-0,123	-0,346
FRO_OVER_SPEC	-0,088	0,180
Frobenius-norm Measures		
FRO_DIST	-0,105	-0,211
LOG_PROD_OF_FRO	-0,123	-0,045
LOG_PROD_OF_FRO_OVER_MARGIN	-0,105	-0,030
LOG_SUM_OF_FRO	-0,123	-0,511
LOG_SUM_OF_FRO_OVER_MARGIN	-0,105	-0,466
log_DIST_SPEC_INIT	0,053	-0,196
PARAM_NORM	-0,105	-0,271
Path-based measures		
PATH_NORM	-0,070	-0,617
PATH_NORM_OVER_MARGIN	0,035	-0,557
Flatness-based measures		
PACBAYES_FLATNESS	-0,353	0,621
PACBAYES_INIT	-0,281	0,511
PACBAYES_ORIG	-0,299	0,466
PACBAYES_MAG_FLATNESS	-0,436	0,062
PACBAYES_MAG_INIT	-0,281	0,361
PACBAYES_MAG_ORIG	-0,299	0,256

Tabela 1: Correlações de Kendall entre a métrica de complexidade e a generalização dos modelos treinados na base BreastMNIST e na CIFAR. Os números dentro da coluna têm cores que vão do vermelho (valores baixos) ao verde (valores altos).

É evidente certa discordância entre os modelos treinados na BreastMNIST e na CIFAR, principalmente nas métricas *Flatness-based*, em que todas demonstram valores invertidos, e em casos em que a métrica mostra-se mais relevante em um, mas pouco no outro. Em geral, varias métricas parecem mais relevantes no caso geral (CIFAR), em especial a *Log-sum-of-spec*, a *Log-sum-of-fro*, ambas as *Path-based*, a *Pacbayes-Flatness* e a *Pacbayes-Init*, todas acima de $0,5$ (independente do sinal), o que indica uma correlação relevante; as mesmas métricas são destacadas nos experimentos com a CIFAR em Jiang et al. [2019]. Já na base *BreastMNIST*, muitas métricas têm valores próximos de zero, indicando uma relação mais fraca entre elas e a generalização dos modelos, além de poucas métricas com valores positivos e nenhuma relevante. As métricas de complexidade com mais destaque são também as *Flatness-based*, embora seja de forma oposta às da CIFAR; em geral, os índices são mais fracos do que os presentes no experimento 1 de Vakanski and Xian [2021], e não concordam sempre em sinal.

Surpreendentemente, a métrica com maior valor absoluto para a *BreastMNIST*, a *Pacbayes-Mag-Flatness* com índice de $-0,436$, é também a métrica de menor valor absoluto dentre as *Flatness-based* para a *CIFAR*. As métricas *Flatness-based* são baseadas no quão plano ou largo é o ótimo local encontrado, como explicado na Seção *2.2.4.*. Isso trás sentido para a discordância entre o caso geral e o médico, pois sendo problemas diferentes, intuitivamente entende-se que as características de seus pontos ótimos sejam também distintos.

Dado que ambos os números foram gerados de maneira idêntica, alterando apenas os dados, é possível ver que de fato as métricas podem funcionar melhor ou pior, afim de prever a generalização, dependendo do contexto em que são usadas, e dentre as analisadas, muitas aparentam ter menor aproveitamento no caso médico. É importante destacar entretanto que embora o fato de os conjuntos de dados da *MedMNIST* serem de fácil acesso e uso tenham incentivado sua escolha, justamente por serem preprocessados com esse fim - como diminuir as imagens de 500x500 para 28x28 - existem diferenças em relação a modelos "reais" de imagens médicas, que costumam manter a resolução e o tamanho grande das imagens. Além disso, como as imagens na MedMNIST são em tons de cinza (apenas 1 canal) e as imagens na CIFAR são coloridas (3 canais), também existe essa diferença que pode gerar, ou não, oscilações nas correlações.

3 Conclusão

Embora existam diversos estudos explorando o tema, existe ainda muito espaço para maiores investigações desses subtemas e suas particularidades, como é o caso médico. Este estudo, assim como o de Vakanski and Xian [2021], avalia pequenas partes do subtema e desperta tanto hipóteses quanto dúvidas. De maneira geral, as métricas com *PAC-Bayes* e *Flatness-based* apareceram com destaque nos dois casos avaliados, e tanto neste estudo quanto nos de Jiang et al. [2019] e Vakanski and Xian [2021], mostraram-se promissoras para o maior entendimento prévio da generalização de modelos.

Ainda que o estudo aqui apresentado seja de média escala, tornou-se mais clara aqui a diferença que o contexto médico tem em relação a contextos mais genéricos. A comparação entre os casos demonstrou tanto métricas que não aparentam ter tanto sucesso no caso médico quanto métricas que, apesar de serem boas nos dois contextos, têm sinais diferentes, impactando diretamente na conclusão observada. Isso confirma a necessidade desse tipo de estudo, afim de facilitar e possibilitar um uso mais correto das métricas de complexidade no desenvolvimento de modelos médicos no futuro.

Referências

Site ultrasoundcases, último acessado em 27/06/2023. URL https://www.ultrasoundcases.info/cases/breast-and-axilla/.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2019.104863. URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *CoRR*, abs/1703.04933, 2017. URL http://arxiv.org/abs/1703.04933.

Yap Moi Hoon, Pons Gerard, Marti Joan, Ganau Sergi, Sentís Melcior, Zwiggelaar Reyer, Davison Adrian, and Martí Robert. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, PP, 2017. doi: 10.1109/JBHI.2017.2731873.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *CoRR*, abs/1912.02178, 2019. URL http://arxiv.org/abs/1912.02178.

M. G. KENDALL. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Isabella Jeske Rosa and João Batista Florindo. Avaliação de métricas de generalização de redes neurais profundas aplicadas a imagens médicas, 2023. URL https://www.ime.unicamp.br/ mac/db/2023-1S-237040.pdf.

Aleksandar Vakanski and Min Xian. Evaluation of complexity measures for deep lear-

ning generalization in medical image analysis. *CoRR*, abs/2103.03328, 2021. URL https://arxiv.org/abs/2103.03328.

Min Xian, Yingtao Zhang, Heng-Da Cheng, Fei Xu, Kuan Huang, Boyu Zhang, Jianrui Ding, Chunping Ning, and Ying Wang. A benchmark for breast ultrasound image segmentation (BUSIS). *CoRR*, abs/1801.03182, 2018. URL http://arxiv.org/abs/1801.03182.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021.

Netzer Yuval, Wang Tao, Coates Adam, Bissacco Alessandro, Wu Bo, and Ng Andrew Y. Reading digits in natural images with unsupervised feature learning. 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011$_h$*ousenumbers.pdf*.