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

ISABELLA JESKE ROSA

Avaliação de métricas de generalização de redes neurais profundas aplicadas a imagens médicas

Campinas
28/06/2023

ISABELLA JESKE ROSA

Avaliação de métricas de generalização de redes neurais profundas aplicadas a 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

Prever a boa ou má generalização de um modelo de rede neural profunda não é uma tarefa simples, tendo gerado diversos estudos procurando comprovar a relação entre a generalização e métricas de complexidade baseadas nos modelos. O problema se torna ainda mais complexo quando se trata de imagens médicas, com particularidades na distribuição e características dos dados. Esta monografia se propõe a discutir e comparar principalmente dois estudos relacionados ao tema, e publicados recentemente (Jiang et al. [2019] e Vakanski and Xian [2021a]), compreendendo melhor o problema médico e onde se encontram as principais barreiras para seu avanço. Em geral, no panorama genérico (i.e., não médico) os testes são extensos, enquanto no panorama médico os estudos têm menor escala e ainda há espaço para maiores experimentações; dado isso, o artigo mais recente e focado em imagens médicas é baseado no primeiro em questão de avaliação das métricas, embora traga testes mais simples específicos ao contexto. Isso faz com que, ao se comparar os dois panoramas, seja difícil entender o que de fato é efeito da mudança de cenário, e não de outras características dos estudos. Ainda que ambos destaquem o potencial das métricas baseadas em *PAC-Bayes*, ainda há abertura para maiores experimentações envolvendo a investigação do caso geral e o caso médico.

Abstract

Predicting a good or bad generalization of a Deep Learning Model is not a simple task, generating several studies searching for better proof of the correlation between generalization and complexity measures. The problem becomes even more complex when it involves medical images, with data properties and distribution particularities. This monography has the purpose of gathering and discussing mostly two related and recent articles (Jiang et al. [2019] and Vakanski and Xian [2021a]), better understanding the medical case problem and the main barriers to the theme's evolution. In general, when looking at the generic landscape (i.e. non-medical) the studies are extensive, while at the medical landscape, the searches are smaller, opening room for further experimentation; given that the most recent article mentioned above and focused on medical images is based on the first one on evaluation's matter, although it makes more simple and specific testing for its context. It makes it difficult to understand what is resulting from the generic/medical changes when comparing both scenarios, and what is just experimental particularities. Although both articles highlight the potential of PAC-Bayes-based measures, there are still opening for further experimentation involving the generic and medical scenarios.

Conteúdo

1 Introdução

As Redes Neurais Profundas estão por trás de toda a revolução dos algoritmos de *Deep Learning* modernos e são indiscutivelmente populares em diversas aplicações. Porém, o que está por trás da sua capacidade de generalização após o treinamento permanece um mistério (Zhang et al. [2017]). Chamamos de generalização o quão bem um modelo se adapta e performa na "vida real", ou seja, em dados nunca vistos; é crucial entender a razão por trás da mesma em modelos modernos de *Deep Learning*, dando benefícios como garantias em cenários críticos, e o *design* de modelos melhores em geral (Jiang et al. [2019]).

As métricas de complexidade podem ser um recurso para entender a profundidade de um modelo, e são conhecidas por terem uma relação supostamente monotônica com a generalização. Existem muitas propostas teóricas e empíricas de métricas diferentes, dependendo das propriedades do modelo treinado - como parâmetros ou estrutura da rede, otimizador escolhido e até os dados de treino. Atualmente existem diversos estudos que se propõem a explorar o tema, além de propor adaptações e novos métodos para testar hipóteses, sendo muitos desses mencionados nos artigos citados. A problemática explorada em Jiang et al. [2019] é a maioria dos estudos serem feitos em problemas de ensaio e não em *datasets* maiores e mais desafiadores, além de explorar pouco possíveis vieses criados ao encontrar a correlação entre métrica e generalização estimada.

Por fim, tem-se um sub-tema explorado em Vakanski and Xian [2021a], que são os problemas relacionados a imagens médicas. Baseado no artigo de Jiang et al. [2019], que utilizam grandes conjuntos de dados de imagens não médicas, alega-se que não há garantia que seus resultados se mantêm na mudança de panorama. Conjuntos de dados de imagens médicas são caracterizados por menor volume de dados e imagens de maior resolução; segundo Vakanski and Xian [2021a], o volume de dados é importante, pois o comportamento de convergência dos modelos pode ser diferente. Além disso, a variância nas imagens coletadas pode interferir na hipótese de i.i.d. (independência e idêntica distribuição entre os dados), por serem de diferentes hospitais, regiões, condutas de procedimento etc., fazendo com que a generalização se torne mais complicada.

2 Desenvolvimento

2.1 Sobre os artigos estudados

Este estudo se baseia principalmente em dois artigos, um de larga escala em *datasets* de imagens gerais *'Fantastic Generalization Mearsures and Where to fing Them'* de Jiang et al. [2019], e outro de menor escala com foco em imagens médicas, *'Evaluation of Complexity Measures for Deep Learning Generalization in Medical Image Analysis'* de Vakanski and Xian [2021a].

O primeiro, feito em 2019 pela Google, se propôs a realizar um teste mais extenso e cauteloso sobre a correlação entre 40 métricas de complexidade diferentes e a generalização, utilizando redes *Network-in-Network* (NiN) (Lin et al. [2014]) nos *datasets* CIFAR-10 (Krizhevsky and Hinton [2009]) e Street View House Numbers (SVHN) (Yuval et al. [2011]).

O segundo, realizado em 2021, é baseado no artigo anterior e em outro estudo menor (Dziugaite et al. [2020]), e reuniu propostas de avaliação de correlação de ambos a fim da verificação de 25 métricas de complexidade em duas configurações de modelo diferentes - Single-task e Multi-task, utilizando os *datasets* BUSIS (Xian et al. [2018]) e BUS-Combined (Xian et al. [2018];Al-Dhabyani et al. [2020];Hoon et al. [2017];Ult).

2.2 Métricas de Complexidade

Nesta seção vamos explicitar as fórmulas utilizadas para o cálculo de algumas das métricas de complexidade avaliadas - em especial, as que serão discutidas mais a frente. Todas as notações e equações foram retiradas do Apêndice de Vakanski and Xian [2021a], 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 diferentes nos logs 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 da 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 adicionar 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 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:

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

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

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

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

2.3 Conceitos e Métodos de Avaliação

Provar a relação métrica de complexidade-generalização envolve muitas considerações e cuidado, devido a vieses facilmente inseridos no processo, além da necessidade de torná-lo o mais fidedigno à realidade possível. Nesta seção serão discutidos alguns conceitos, definições, e quais foram os métodos de avaliação abordados em cada um dos artigos. Alguns conceitos e notações importantes são:
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.

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

Logo, a Generalização avaliada é uma aproximação baseada no erro do modelo observado no treino e no teste - este que configura uma simulação do erro em dados não vistos. O próximo passo é compreender quais testes são satisfatórios para analisar a relação entre o *Gap* definido e as métricas de complexidade. Jiang et al. [2019] discute no tópico *'Generalization: What is the goal and how to evaluate?'* três principais abordagens:

- A rigidez dos limites de generalização - ou *'Tightness of Generalization Bounds'*, que, brevemente, é um limite probabilístico para o *gap* de generalização (Rosasco and Poggio [2009]). Os autores escolhem não usar esta abordagem por ter pouca utilidade prática para a maioria das atuais demandas em Deep Learning.

- Utilizar a métrica de complexidade como regularização no modelo e avaliar a melhora na generalização. O artigo cita duas principais razões para a falha do método; primeiro, utilizar a métrica como regularização na otimização pode alterar o custo de forma não trivial, dificultando a convergência e fazendo com que nenhuma conclusão sobre causalidade possa ser tomada. Em segundo, a regularização implícita (e inevitável) do próprio otimizador, o que faz com que o experimento não seja controlado e não se possa garantir uma conclusão confiável.

- Por fim, o Índice de correlação com a Generalização é o método escolhido em ambos os estudos. Jiang et al. [2019] destaca o valor da correlação na avaliação das métricas, mas mantém que também podem ocorrer vieses, sendo importante produzir diversos modelos ao variar a arquitetura e algoritmos de otimização, com cuidado ao considerar hiper-parâmetros já conhecidamente correlacionados com a generalização. Em síntese, o experimento deve ser controlado.

Estabelecida a forma de avaliação, existem diversos índices de correlação que podem ser usados. Dado um conjunto de modelos formados no experimento, gerados através de um espaço de hiper-parâmetros Θ, seus respectivos *Gap*'s de Generalização

$\{g(\theta)|\theta \in \Theta\}$ e métricas $\{\mu(\theta)|\theta \in \Theta\}$, cria-se um conjunto $\mathcal{T}$, em que cada elemento é um par formado por

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

Disso, uma das opções de uso é o **Kendall's Rank-Correlation Coefficient**, ou Coeficiente de Kendall τ (KENDALL [1938]), que captura o grau de consistência entre os elementos do conjunto $\mathcal{T}$; define-se consistência quando para qualquer par de $\mathcal{T}$, se $\mu(\theta_1) > \mu(\theta_2)$ então $g(\theta_1) > g(\theta_2)$.

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

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.

Entendendo as limitações do Coeficiente de Kendall, Jiang et al. [2019] propôs outra métrica derivada, o **Granulated Kendall's Coeficient** Ψ, com o objetivo de mitigar o efeito de coeficientes de correlação fortes que não necessariamente discriminam a causa da generalização. A ideia é que as métricas de complexidade podem performar melhor, ou pior, dependendo dos hiper-parâmetros, e então o coeficiente ψ deve ser calculado separadamente ao longo de cada *set* de hiper-parâmetros Θ_i, e após computar a média Ψ com base no espaço Θ. Isso implica que mesmo a métrica performando bem em algum *set* Θ_i, se o mesmo não ocorrer em outros, o valor final pode ser baixo.

Não obstante, em Vakanski and Xian [2021a] também é utilizado o **Robust Kendall's** (Dziugaite et al. [2020]), baseado em robustez distribucional. Neste aplicam-se pesos maiores em elementos com *Gap's* menores, afim de evitar o uso de média simples. Varia ente 0 e 1, em que menores coeficientes indicam maior correlação.

2.4 Experimentos

2.4.1 Fantastic Generalization Measures and Where to find them

Sobre os conjuntos de dados utilizados, temos o CIFAR-10 e o SVHN (*Street View House Numbers*). O CIFAR-10 contém 60000 imagens 32x32 (pequenas) RGB com 10 classes balanceadas, sendo 50000 treino e 10000 teste. Já o SVHN contém imagens 32x32 RGB de números de casas, com 600 mil amostras, sendo 73257 treino, 26032 teste e de 530 mil imagens adicionais. Uma pequena amostra de ambos os conjuntos pode ser vista na Figura 1 abaixo.

Figura 1: Amostra de imagens dos conjuntos de dados CIFAR-10 (à esquerda) e SVHN (à direita)

O Design Experimental de Jiang et al. [2019] foi baseado em gerar um conjunto de modelos treinados ao variar 7 hiper-parâmetros diferentes. Foram gerados $3^7 = 2187$ preditores, inspirados na arquitetura de rede *Network-in-Network* (NiN) e treinados no conjunto CIFAR-10, além da repetição do mesmo para o conjunto SVHN; cada experimento foi repetido 5 vezes, garantindo um resultado robusto em relação à aleatoriedade. Os modelos foram treinados até a convergência, com o critério de *cross-entropy loss* igual a $0,01$, garantindo que todos os modelos estejam no mesmo patamar de convergência - que pode afetar a generalização. O espaço de hiper-parâmetros Θ foi definido por 3 escolhas para cada um (i.e $|Theta_i| = 3$), sendo eles: *weight decay, width, mini-batch size, learning rate, dropout, depth* e o algoritmo de otimização. Menos de 5% dos modelos não atingiu o critério. Na Figura 2 é possível ver a distribuição do *Gap* de Generalização.

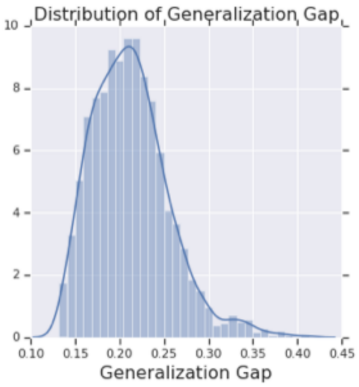

Figura 2: Distribuição do *Gap* de Generalização nos modelos de experimentação de Jiang et al. [2019]

2.4.2 Evaluation of C. Measures for D. Learning Generalization in Medical Image Analysis

Iniciando novamente com os conjuntos de dados utilizados, temos o BUSIS e o BUS-combined. Ambos são compostos por imagens de ultrassom de mamas; O BUSIS, fruto de um *benchmark* de segmentação de imagens do BUS Xian et al. [2018], tem 562 imagens 250x250 (46% com tumores malignos). Já o BUS-combined é uma combinação feita pelos autores, contendo 3574 imagens (48% malignos) de quatro outros conjuntos (BUSIS (562 images) Xian et al. [2018], BUSI (647 images) Al-Dhabyani et al. [2020], B (163 images) Hoon et al. [2017], e HMSS (2,202 images) Ult).

Na Figura 3 é possível ver algumas imagens do conjunto original BUS retiradas de um *Review* da base de Tsang [2022].

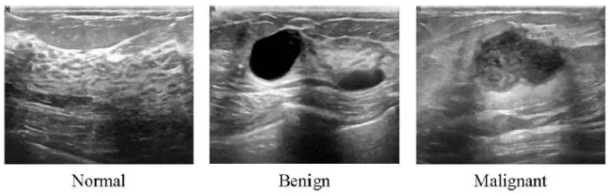

Figura 3: Exemplos de imagens de ultrassom de mamas do conjunto de dados BUS (Tsang [2022])

Para a avaliação das 25 métricas de complexidade, Vakanski and Xian [2021a] arquiteta 3 experimentos separados incluindo duas tarefas diferentes - *single-task* (classificação) e *multi-task* (classificação e segmentação), com a ideia de construir modelos e tarefas mais próximos dos utilizados no contexto de imagens médicas. A arquitetura

usada em ambas é inspirada nas redes *VGG* e *U-net*, sendo a Figura 4 uma representação gráfica disponibilizada por Vakanski and Xian [2021a].

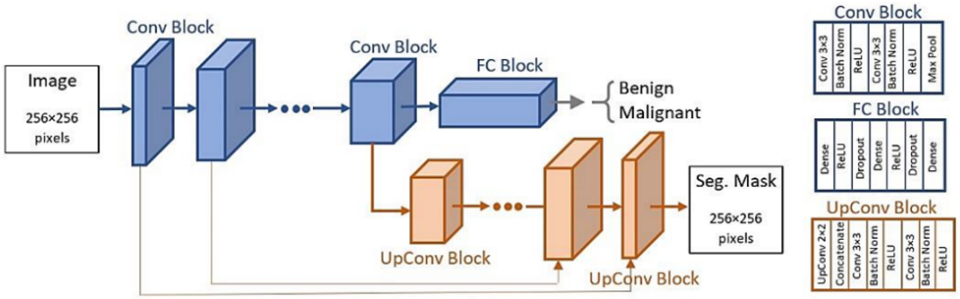

Figura 4: Representação gráfica da arquitetura de rede utilizada; em azul o ramo de classificação (inspirada na *VGG*; em laranja o ramo de segmentação (inspirada na *U-net*). Imagem retirada do repositório no github (Vakanski and Xian [2021b])

Cada experimento foi caracterizado pela variação de *Conv Blocks*, ou blocos de convolução, entre 2 (67 mil parâmetros) e 8 (152 milhões de parâmetros); cada preditor de mesma arquitetura foi treinado 20 vezes com diferentes *subsets* de treino e teste do conjunto de dados, e diferentes *random seeds* para a inicialização dos parâmetros. Mais formalmente, para cada experimento ε, são geradas 7 famílias de 20 preditores cada, tendo, portanto, 140 modelos representados por $\{h_1, h_2, ..., h_{20}\} \in \mathcal{H}_i, i \in \{1, ..., 7\}$, todos com *Adam Optimizer* como algoritmo de otimização. Os experimentos ε foram estruturados das seguintes formas:

1. Experimento ε_1: Apenas classificação (*Single-task*), usando o conjunto de dados *BUS-combined* (maior quantidade de dados) dividido em 80% treino e 20% teste; a *Learning Rate* a 10^{-5} e *mini-batch* de 2 são fixos. Os modelos, assim como no artigo anterior, também foram treinados até o critério de convergência de *cross-entropy loss* igual a $0,01$.

2. Experimento ε_2: Apenas classificação (*Single-task*), usando o conjunto de dados BUSIS (menor quantidade de dados) e *k-folds* (k=5) dividido em 60% treino, 20% validação e 20% teste; a *Learning Rate* também a 10^{-5} fixa. Diferente do ε_1, a intenção é aplicar uma abordagem próxima do real, e foi inserida regularização explícita com *early stopping* (critério de parada se o risco na validação não diminuir por 30 épocas), além de *Batch normalization* e *Dropout Layers*.

3. Experimento ε_3: Envolve classificação e máscara de segmentação (*Multi-task*), com o conjunto BUSIS e mesma estrutura de classificação que o ε_2. Além disso, apesar da ramificação de segmentação, as métricas de complexidade baseadas na rede foram calculadas apenas na de classificação.

2.5 Discussão de Resultados

Para a discussão, foram recolhidos alguns resultados sobre métricas avaliadas em comum de ambos os estudos, a fim da comparação do cenário geral e médico. Algumas métricas mais interessantes foram selecionadas, e os dados nas tabelas desta seção foram retirados dos próprios artigos.

É importante mencionar que os estudos não são completamente compatíveis, tendo metodologias de experimentação que se diferenciam além de somente na natureza dos dados. Dentre os três experimentos realizados por Vakanski and Xian [2021a], o que é mais similar ao de Jiang et al. [2019] é o experimento ε_1 - 140 modelos gerados apenas com a variação do número de *Conv Blocks*, e treinados até a *loss* atingir $0,01$. Além disso, os experimentos ε_2 e ε_3 sofreram *Underfitting* nos modelos com menor número de *Conv Blocks* (i.e. o modelo não foi complexo suficiente para aprender o problema), e ambos os erros de treino e teste foram muito altos, ocasionando um *Gap* pequeno que cresce com o aumento de *Conv Blocks*, sugerindo conclusões equivocadas. A solução dos autores foi calcular o coeficiente de Kendall também para o erro de teste (τL), ao invés de apenas para o *Gap*.

Com essas informações, a comparação principal será entre os resultados dos Coeficientes de Kendall e Kendall Granulado do experimento na base CIFAR-10 e o experimento ε_1, presentes nas colunas 3-6 da Tabela 2.

		Artigo Geral		Artigo Médico					
		CIFAR-10		Exp. 1 (ε1)		Exp. 2 (ε1)		Exp. 3 (ε1)	
Métricas	**Ref**	**τ (ΔL)**	**Ψ**	**τ (ΔL)**	**Ψ**	**τ (ΔL)**	**τ (L)**	**τ (ΔL)**	**τ (L)**
vc-dimension	1	-0,251	-0,154	-0,280	-0,299	0,331	-0,485	0,430	-0,394
inverse margin	2	-0,124	-0,121	0,290	0,325	-0,320	0,281	-0,375	0,349
frob. distance	3	-0,263	-0,341	0,165	0,139	-0,114	-0,094	-0,122	0,052
parameter-norm	4	0,073	0,052	-0,211	-0,288	0,366	-0,422	-0,214	0,120
path-norm	5	0,373	0,311	0,234	0,288	-0,023	0,164	-0,357	0,339
path-norm/margin	6	0,374	0,305	0,281	0,331	-0,092	0,141	-0,436	0,401
pacbayes-flatness	7	0,303	0,346	0,327	0,373	-0,211	0,144	0,417	-0,306
pacbayes-mag-flatness	8	0,365	0,315	0,491	0,533	-0,372	0,381	-0,342	0,211
pacbayes-mag-init	9	0,175	0,052	-0,276	-0,297	0,32	-0,469	0,389	-0,387
pacbayes-mag-orig	10	0,41	0,284	-0,285	-0,299	0,32	-0,469	0,387	-0,385

Tabela 1: Tabela com as métricas selecionadas, número de referência e coeficientes de correlação de Kendall τ e Kendall Granulado ψ para o estudo geral na base CIFAR-10, e experimentos 1, 2 e 3 do estudo médico ($\tau\Delta L$ para o *Gap* e τL para o erro de teste). A graduação de cor vai de laranja para os valores mais negativos, passa por branco e vai até o azul para os valores mais positivos. Todos os valores foram retirados dos artigos analisados.

Dentre esses dois casos (geral e médico exp. ε_1), várias métricas apresentaram resultados semelhantes, assim como resultados opostos (e.g. correlações positivas em um e negativas em outro). Em especial, as métricas *Path-norm, Path-norm/margin, Pacbayes-flatness e Pacbayes-mag-flatnesss* mostraram correlação positiva em ambos os coeficientes e em ambos os casos. Em contrapartida, a margem inversa, *frob. distance*, *pacbayes-mag-orig* e *-init* foram para lados opostos, invertendo o sinal dos coeficientes entre os casos.

Entretanto, é possível perceber algo intrigante quando olhamos as métricas mencionadas no parágrafo anterior nos experimentos ε_2 e ε_3. Ao observar os coeficientes das métricas que concordaram com o Caso Geral no experimento ε_1 (1, 5, 6 e 8), isso se mantém nos outros experimentos apenas para os coeficientes com base no erro de teste (τL), e não no *Gap* ($\tau\Delta L$). Já para as métricas que discordaram (2, 9 e 10), é o inverso - as métricas passam a concordar com o Caso geral em ε_2 e ε_3 nos coeficientes com base no *Gap*. Em outras palavras, no segundo e terceiro experimento, apenas o τL mantém o comportamento observado em ε_1, seja ele concordante ou discordante com o Caso Geral. Isso acaba evidenciando mais o problema de *Underfitting* em ε_2 e ε_3, sugerindo que o coeficiente baseado no erro de teste tenha sido uma boa solução para resultados mais coerentes entre si. Na Tabela 2 abaixo, é possível enxergar esse padrão um pouco melhor; nela, foram incluídas as métricas mencionadas acima, e pintadas de azul quando são

concordantes com o Caso geral (ou seja, quando as correlações têm o mesmo sinal).

		Artigo Geral		Artigo Médico					
		CIFAR-10		Exp. 1 (ε1)		Exp. 2 (ε1)		Exp. 3 (ε1)	
Métricas	Ref	τ (ΔL)	Ψ	τ (ΔL)	Ψ	τ (ΔL)	τ (L)	τ (ΔL)	τ (L)
vc-dimension	1	-0,251	-0,154	-0,280	-0,299	0,331	-0,485	0,430	-0,394
path-norm	5	0,373	0,311	0,234	0,288	-0,023	0,164	-0,357	0,339
path-norm/margin	6	0,374	0,305	0,281	0,331	-0,092	0,141	-0,436	0,401
pacbayes-mag-flatness	8	0,365	0,315	0,491	0,533	-0,372	0,381	-0,342	0,211
inverse margin	2	-0,124	-0,121	0,290	0,325	-0,320	0,281	-0,375	0,349
pacbayes-mag-init	9	0,175	0,052	-0,276	-0,297	0,32	-0,469	0,389	-0,387
pacbayes-mag-orig	10	0,41	0,284	-0,285	-0,299	0,32	-0,469	0,387	-0,385

Tabela 2: Recorte de algumas métricas da Tabela 1, com valores dos experimentos 1, 2 e 3 do caso médico, pintados de azul quando são concordantes (mesmo sinal) com o Caso Geral (3ª e 4º colunas) e sem coloração se tem sinal contrário.

A razão por trás do alinhamento ou discordância entre os estudos pode ser devido a alguma característica do aparato experimental, ou da própria métrica ao mudar para o contexto médico; Dado que dentre os dados observados não houve um padrão claro entre o tipo de métrica e os resultados baterem ou não, para avaliar a eficiência das métricas com relação à generalização entre casos genéricos e médicos com maior critério, seriam necessários maiores estudos, comparando experimentos mais próximos em método e controle.

3 Conclusão

Embora existam muitos estudos e artigos extensivos explorando o tema de generalização e métricas de complexidade, ainda existem casos particulares com espaço para investigação - sendo o caso médico um deles. Nos casos avaliados, o artigo feito por Jiang et al. [2019] contém experimentos extensivos e cuidadosos, que exigem grande poder computacional; o artigo de Vakanski and Xian [2021a] explora alguns experimentos, em busca de entender métricas de complexidade que podem explicar a generalização de modelos com imagens médicas. Dado que ambos os estudos destaquem o sucesso das métricas baseadas em PAC-bayes (em especial a *pacbayes-mag-flatness*), pode-se entender que são métricas promissoras no objetivo de compreender melhor a generalização de redes neurais.

Não obstante, na comparação entre ambos não fica claro se a diferença de resultados em algumas métricas, é devido as características do aparato experimental ou se realmente funcionam melhor ou pior no caso médico; Por esta razão, seria interessante uma análise de viabilidade da elaboração de um estudo, comparando algumas métricas de complexidade em *datasets* genéricos e médicos, de forma comparável metologicamente. De tal forma, seria possível compreender não só o desempenho dessas métricas em prever a generalização, mas também discriminar - se existirem - particularidades no cenário de imagens médicas

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.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. *CoRR*, abs/2010.11924, 2020. URL https://arxiv.org/abs/2010.11924.

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.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2014.

Lorenzo Rosasco and Tomaso Poggio. Generalization bounds and stability, 2009. URL https://www.mit.edu/ 9.520/spring09/Classes/class09$_s$*tability.pdf*.

Sik-Ho Tsang. Review — busi: Dataset of breast ultrasound images, 2022. URL https://sh-tsang.medium.com/review-busi-dataset-of-breast-ultrasound-images-174f3

Aleksandar Vakanski and Min Xian. Evaluation of complexity measures for deep learning generalization in medical image analysis. *CoRR*, abs/2103.03328, 2021a. URL https://arxiv.org/abs/2103.03328.

Aleksandar Vakanski and Min Xian. Repositório do github evaluation-of-complexity-measures-for-deep-learning-generalization-in-medical-image-analysis, 2021b. URL https://github.com/avakanski/Evaluation-of-Complexity-Measures-for-Deep-Learning-

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.

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*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017.