Guidelines for Writing a Scientific Paper (Machine Learning bias)

Artur Jordão
arturjordao[at]usp.br

Introdução

Guidelines for Writing a Scientific Paper

- Este guia ilustra como escrever um artigo científico a partir de exemplos extraídos de artigos publicados em importantes veículos de publicação
  - International Conference on Learning Representations (ICLR)
  - Neural Information Processing Systems (NeurIPS)
  - International Conference on Machine Learning (ICML)
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)
  - Computer Vision and Pattern Recognition (CVPR)
  - International Conference on Computer Vision (ICCV)

- A maioria dos exemplos contempla apenas pequenos trechos do trabalho original
  - Desta forma, para os interessados, é recomendado que leiam os artigos na íntegra

Delineando uma Introdução

Características e Finalidade

Introdução

- Parte mais importante (e complicada) do texto
- Apresentar a importância da linha de pesquisa do trabalho
- Apresentar onde esforços estão sendo concentrados
  - Apontar as soluções existentes e suas potenciais limitações (caso existam)
  - Indicar se existem lacunas na literatura e a importância de preencher tais lacunas
- Indicar a importância do trabalho atual
- Apresentar as contribuições do trabalho (*key insights*)
- No geral a introdução faz promessas para o leitor

Questões que devem ser pensadas antes de iniciar a introdução

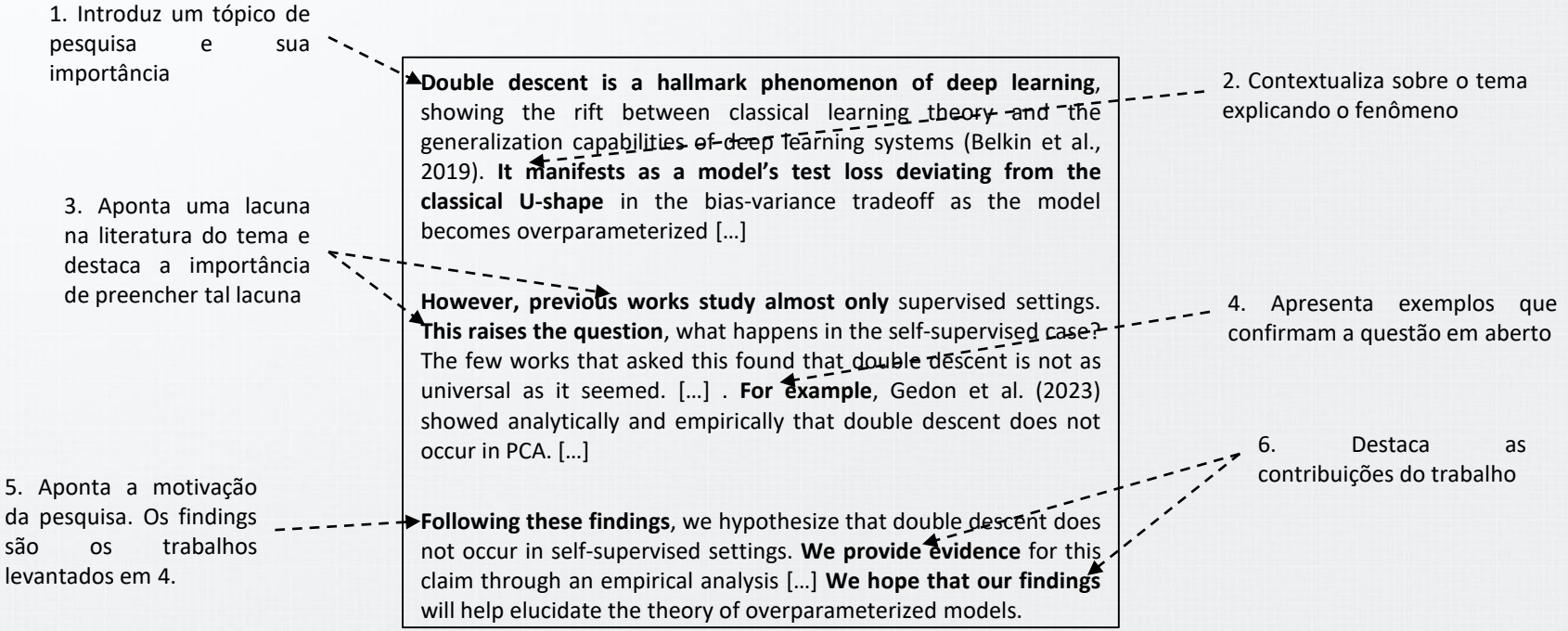
Introdução

- As decisões do trabalho atual estão alinhadas com as evidências da literatura?
- O trabalho apresenta maior ênfase prática ou teórica?
- O trabalho está propondo um método, um *dataset,* uma melhoria em soluções existentes ou uma inovação para uma tarefa específica?
  - **Evidências da literatura apoiam a proposta**?
  - Quais são os problemas/desafios associados ao estado da arte atual?
  - A proposta preenche lacunas da literatura? Quais são as lacunas?
- O trabalho é uma revisão da literatura?
  - Em caso afirmativo, o que acrescenta em relação às revisões existentes?

Exemplo

Introdução

- Fonte: Jayalath et al. *No Double Descent in Self-supervised Learning*. ICLR, 2023



Exemplo

Introdução

- Fonte: Evci et al. *Head2Toe: Utilizing Intermediate Representations for Better Transfer Learning.* ICML, 2022

1. Introduz um tópico e sua importância

Transfer learning is a widely used method for obtaining strong performance in a variety of tasks where training data is scarce (e.g., Zhu et al., 2020; Alyafeai et al., 2020; Zhuang et al., 2020). **A well-known recipe for transfer learning involves** the supervised or unsupervised pretraining of a model on a source task with a large training dataset (also referred to as upstream training) [...]

2. Contextualiza sobre o tema

FINETUNING's superior generalization in the low-data regime is counterintuitive […]. **How does FINETUNING learn from few examples successfully? We conjecture that** FINETUNING better leverages existing internal representations rather than discovering entirely new representations; FINETUNING exposes existing features buried deep in the net for use by the classifier. **Under this hypothesis**, features needed for transfer are already present in the pretrained network […]

3. Destaca alguns embasamentos e questionamentos que motivam o trabalho. Em outras palavras, esses são os pontos que motivam o trabalho

4. Apresenta o objeto de pesquisa

In this work, we propose and explore methods for selecting useful features from all layers of a pretrained net, […]

Our key contributions are as follows:

1. **We observe** a strong correlation between the degree to which a target […]
2. **We introduce** HEAD2TOE, an efficient transfer learning method for […]
3. On the VTAB collection of data sets, **we show that** […]
4. Critically, **HEAD2TOE outperforms** FINETUNING on OOD target domains […]

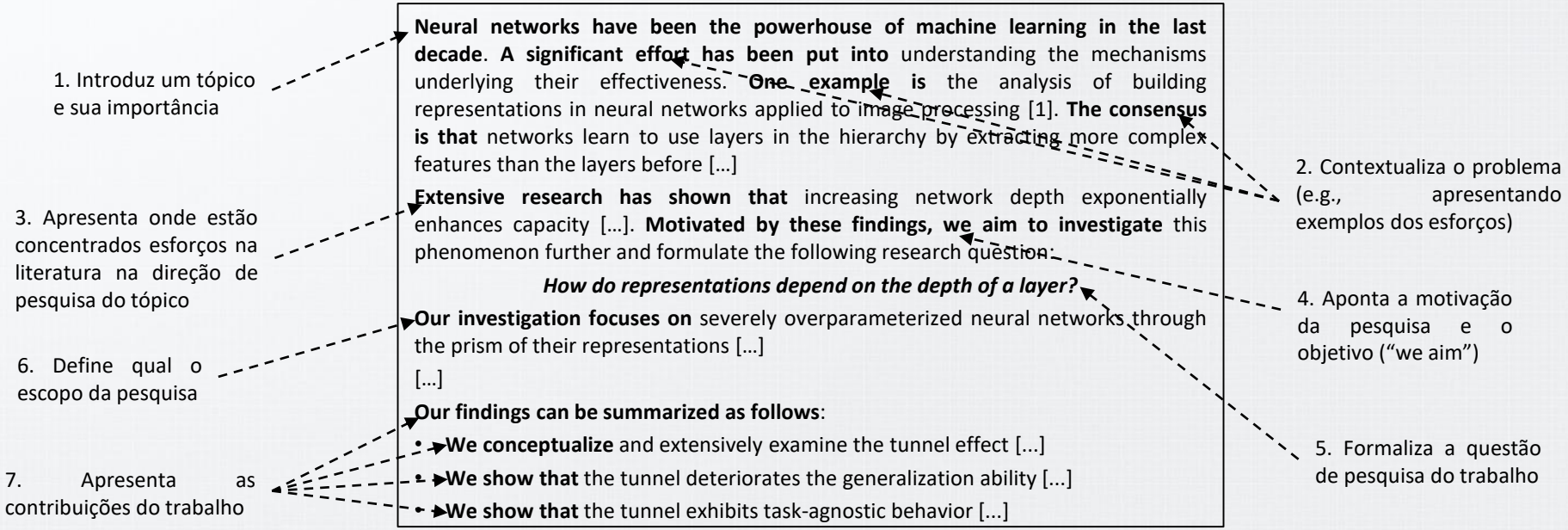
5. Destaca as contribuições do trabalho

Importante: Rodar experimentos em um dataset **não é contribuição**. Por exemplo "We conduct experiments on the VTAB collection..." porém, notem que os autores complementam, com "[...] we show that"

Exemplo

Introdução

- Fonte: Masarczyk et al. *The Tunnel Effect: Building Data Representations in Deep Neural Networks.* NeurIPS, 2023



Exemplo

Introdução

- Fonte: Duong et al. *Representational Dissimilarity Metric Spaces for Stochastic Neural Networks*. ICLR, 2023

1. Introduz a importância do tópico

2. Descreve as linhas de pesquisa existentes sobre o tópico

4. Explica sobre o tópico de atuação mencionando exemplos (omitido aqui por questões de clareza)

3. Aponta lacunas para melhorias nas pesquisas mencionadas em 2. Observe como a combinação das palavras "however" e "only" passam uma mensagem de "room for improvement"

5. Destaca a importância do tópico apontado em 4.

7. Destaca qual a contribuição. Devido às notas 1., 5. e 6, a contribuição é, de fato, importante

6. Identifica a lacuna da literatura sobre o tópico. Observem o "to our knowledge"; isso protege o autor de críticas como "acabou de sair um trabalho na mesma direção"

Comparing high-dimensional neural responses — [...] — is a **fundamental problem** in neuroscience and machine learning (Dwivedi & Roig, 2019; Chung & Abbott, 2021). **There are now many methods** for quantifying representational dissimilarity including canonical correlations analysis [...] **However, these methods only** compare deterministic representations [...]

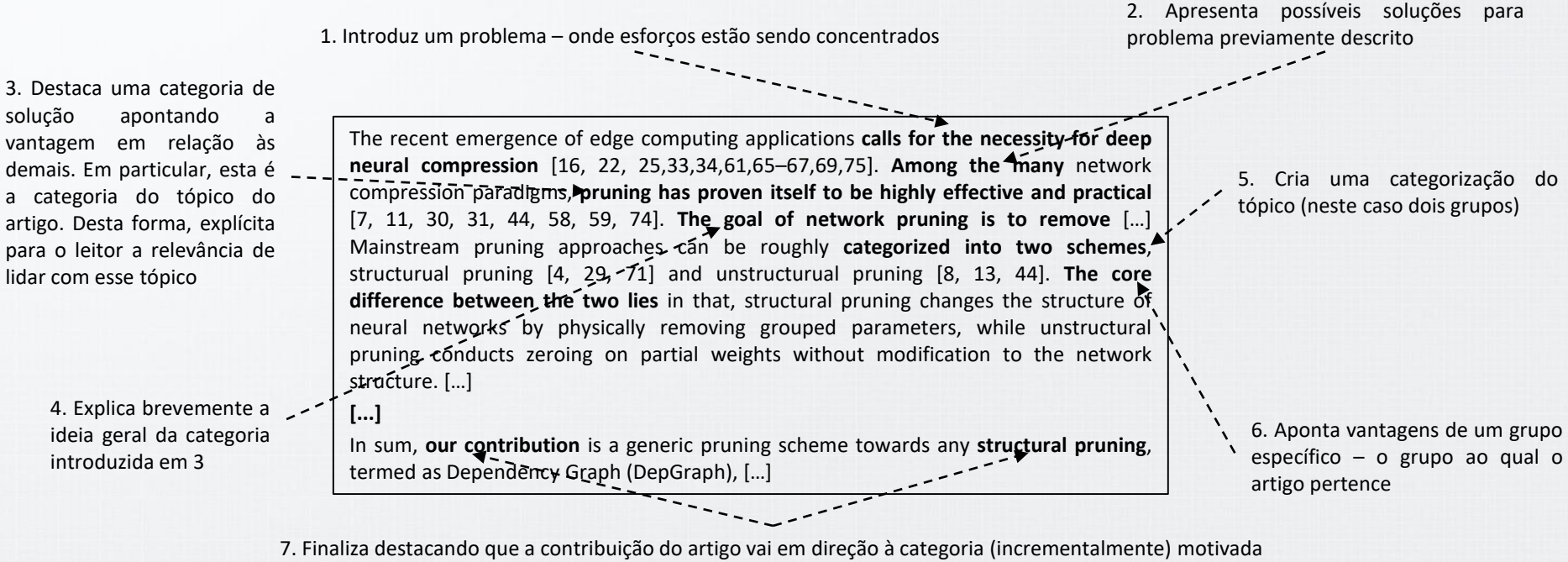
Stochastic networks may be conceptualized as functions mapping [...] **A rich literature in neuroscience has built on top of this insight** (Shadlen et al., 1996; Abbott & Dayan, 1999; Rumyantsev et al., 2020). **However, to our knowledge, no studies** have compared [...]

To address these gaps, we formulate a novel class of metric spaces over stochastic neural representations [...]

Exemplo

Introdução

- Fonte: Fang et al. *DepGraph: Towards Any Structural Pruning*. CVPR, 2023



Exemplo

Introdução

- Fonte: Mirzasoleiman et al. *Coresets for Data-efficient Training of Machine Learning Models.* ICML 2020

Mathematical optimization lies at the core of training large scale machine learning systems, and **is now widely used over** massive data sets with great practical success, assuming sufficient data resources are available. **Achieving this success, however, also requires large amounts of (often GPU) computing**, as well as concomitant financial expenditures and energy usage (Strubell et al., 2019). **Significantly decreasing these costs without decreasing** the learnt system's resulting accuracy is **one of the grand challenges** of machine learning and artificial intelligence today (Asi & Duchi, 2019).

[...]

[...] Incremental Gradient (IG) methods, such as Stochastic Gradient Descent (SGD) and its accelerated variants, [...] **The majority of the work** speeding up IG methods has thus primarily **focused on reducing** the variance of the gradient estimate [...]

However, **the direction that remains largely unexplored** is how to carefully select a small subset $S \subseteq V$ of the full training data V, [...] **If such a subset S can be quickly found, then this would directly lead to a speedup** [...]

Here we develop Coresets for Accelerating Incremental Gradient descent (CRAIG), for selecting a subset of training data points **to speed up training of large machine learning** models [...] **Our key idea is to** [...]

1. Introduz um tópico de pesquisa e sua relevância

2. Apresenta problemas no escopo do tópico

3. Contextualiza a importância e os desafios para mitigar os problemas apontados em 2.

4. Explora o que a literatura tem investigado para lidar com o problema anteriormente descrito.

5. Aponta uma lacuna na literatura e destaca a importância de preencher tal lacuna

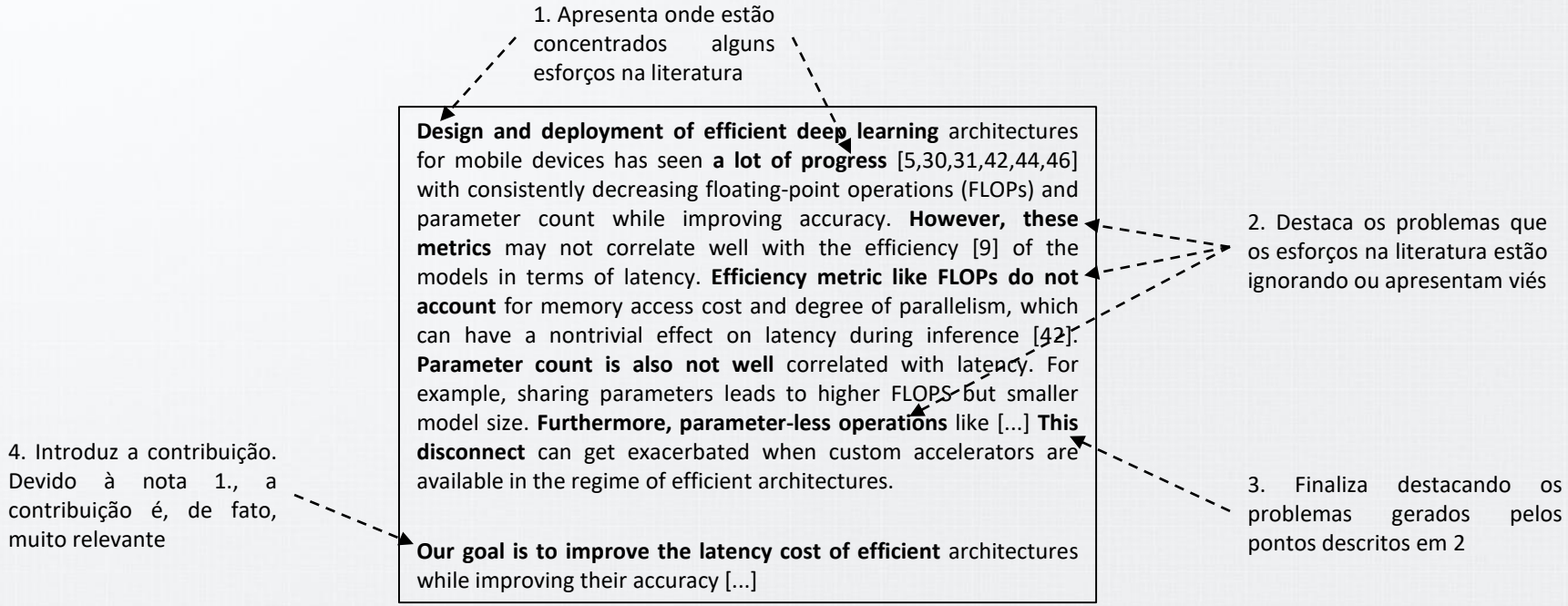
6. Descreve a solução que artigo está propondo. Observe que a solução proposta mitiga todas as questões levantadas anteriormente (2, 5 e 7)

7. Explica a ideia geral da solução proposta

Exemplo

Introdução

- Fonte: Vasu et al. *MobileOne: An Improved One millisecond Mobile Backbone*. CVPR, 2023



Trabalhos Relacionados e o Estado da Arte

Características e Finalidade

Trabalhos Relacionados

- A ideia não é apenas descrever o que já existe em relação ao tópico do artigo
  - É preciso elaborar sobre a contribuição do artigo em relação ao conhecimento existente

- Qual é o estado da arte e onde/como a fronteira do conhecimento está sendo avançada
  - **Descrever como as soluções abordam o problema em questão e como seu artigo inova nesse aspecto**

- Corroborar ou refutar *insights* existentes na literatura

- Suportar decisões e escolhas metodológicas/experimentais
  - *Benchmarks* e *baselines* (modelos, arquiteturas de redes neurais) utilizados
  - Configuração experimental

Exemplos

Trabalhos Relacionados

- Fonte: He et al. *Deep Residual Learning for Image Recognition*. CVPR, 2016
  - Best paper

1. Explicitamente mostra que existe um trabalho “concorrente”

2. Explica uma particularidade do trabalho citado

> **Concurrent with our work**, “highway networks” [41, 42] present shortcut connections with gating functions [15]. **These gates are data-dependent** and have parameters, **in contrast to our identity** shortcuts that are parameter-free. **When a gated shortcut is** “closed” (approaching zero), the layers in highway networks represent non-residual functions. **On the contrary, our formulation always** learns residual functions; our identity shortcuts are never closed, and all information is always passed through, with additional residual functions to be learned. **In addition, highway networks have not demonstrated accuracy gains** with extremely increased depth (e.g., over 100 layers).

3. Menciona uma distinção do trabalho atual com o artigo citado considerando a explicação em 2.

4. Detalha outra particularidade do mesmo trabalho

5. Novamente, destaca uma distinção do que acabou de ser explicado em 4.

6. Conclui a comparação com o trabalho apontando vantagens adicionais (além das apontadas em 3 e 5)

Exemplos

Trabalhos Relacionados

- Fonte: Davari et al. *Reliability of CKA as a Similarity Measure in Deep Learning*. ICLR, 2023

1. Explicita que existe um trabalho alinhado com o foco do trabalho atual

Most closely related to our work, Ding et al. (2021) demonstrated that CKA lacks sensitivity […]. Also, Nguyen et al. (2022) found that the previously observed high CKA similarity between representations of later layers in large capacity models […] **We distinguish ourselves from these papers by providing** theoretical justifications to CKA sensitivity to outliers and to directions of high variance which were only empirically observed in Ding et al. (2021); Nguyen et al. (2021). **Secondly, we do not only** present situations in which CKA gives unexpected results **but** we also show how CKA values can be manipulated to take on arbitrary values.

2. Descreve a diferença do trabalho em relação ao que foi descrito nas sentenças anteriores

O ponto 2 é muito importante: não faz sentido mencionar que existem trabalhos similares e não explicitar para o leitor qual a(s) diferença(s) com o trabalho atual

Exemplos

Trabalhos Relacionados

- Fonte: Touvron et al. *ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training*. PAMI, 2023

1. Aponta um trabalho que complementa alguma das contribuições feitas no artigo

2. Em seguida, destaca qual a novidade do artigo atual, isto é, contrasta com a literatura

Concurrent Works. The MLP-Mixer model **by Tolstikhin et al. [63] brings complementary insights to ours**: they achieve interesting performance with larger MLP models [...] **In contrast, we focus on** faster models trained on ImageNet1k, **and provide more** architectural ablations. **Another concurrent related work** includes the work of [...] **In comparison to our approach**, MLP-Mixer uses a MLP [...]

3. Introduz e explica (brevemente – ideia geral) outro trabalho na mesma direção

4. Novamente, destaca a novidade do artigo atual com o que acabou de ser explicado

Os itens 2 e 4 são muito importantes. Se o que o artigo atual aborda perguntas que já foram respondidas, qual a contribuição? Isto é, por que o artigo deveria ser aceito?

Exemplos

Trabalhos Relacionados

- Podemos “agrupar” uma série de trabalhos
  - Isso evita descrever cada trabalho de maneira muito específica

- Fonte: Mirzasoleiman et al. *Coresets for Data-efficient Training of Machine Learning Models.* ICML, 2020

1. Inicia descrevendo uma categoria de estudos em relação ao tópico de pesquisa do artigo

Convergence of IG methods has been long studied under various conditions (Zhi-Quan & Paul, 1994; Mangasariany & Solodovy, 1994; Bertsekas, 1996; Solodov, 1998; Tseng, 1998), however IG's convergence rate has been characterized only more recently (see (Bertsekas, 2015b) for a survey). **In particular**, (Nedic & Bertsekas, 2001) provides a […] **While these works provide** convergence on the full dataset, **our analysis provides** the same convergence rates on subsets obtained by CRAIG.

2. Especifica somente um trabalho da categoria apresentada em 1. para fornecer uma ideia geral ao leitor. Se está sumarizando não faz sentido dar muitos exemplos

3. Contrasta o trabalho atual com a categoria descrita e não somente com um artigo

Exemplos

Trabalhos Relacionados

- Podemos “agrupar” uma série de trabalhos
  - Isso evita descrever cada trabalho de maneira muito específica

- Fonte: Mirzasoleiman et al. *Coresets for Data-efficient Training of Machine Learning Models.* ICML, 2020

1. Descreve onde esforços em determinado tema estão sendo concentrados, abordando a categoria (*variance reduction*) à qual os trabalhos citados pertencem

Techniques for speeding up SGD, are mostly focused on variance reduction techniques (Roux et al., 2012; ShalevShwartz & Zhang, 2013; Johnson & Zhang, 2013; Hofmann et al., 2015; Allen-Zhu et al., 2016), and accelerated gradient methods when the regularization parameter is small (Frostig et al., 2015; Lin et al., 2015; Xiao & Zhang, 2014). […] **Our CRAIG method and analysis are complementary** to **variance reduction** and accelerated methods. CRAIG can be **applied to all these methods** as well to speed them up.

2. Introduz o complemento do trabalho em relação aos esforços existentes, utilizando a categoria mencionada

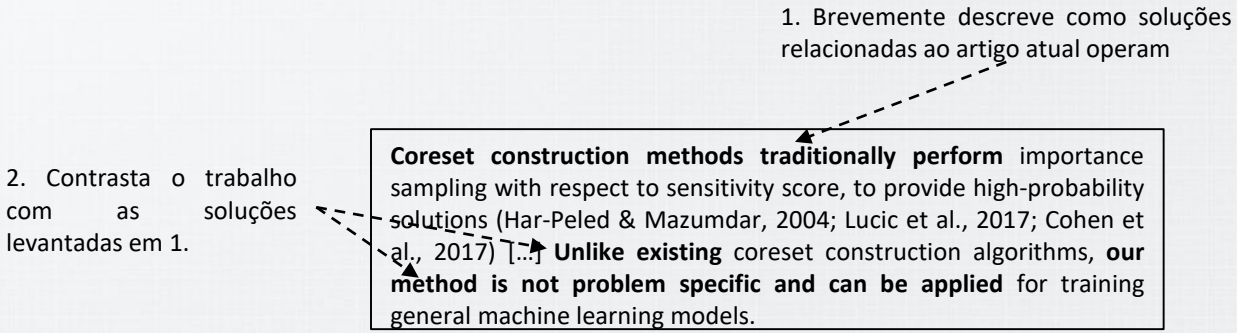
3. Aponta que o trabalho não só complementa o que existe mas pode também ser utilizado em conjunto

Exemplos

Trabalhos Relacionados

- Podemos “agrupar” uma série de trabalhos
  - Isso evita descrever cada trabalho de maneira muito específica

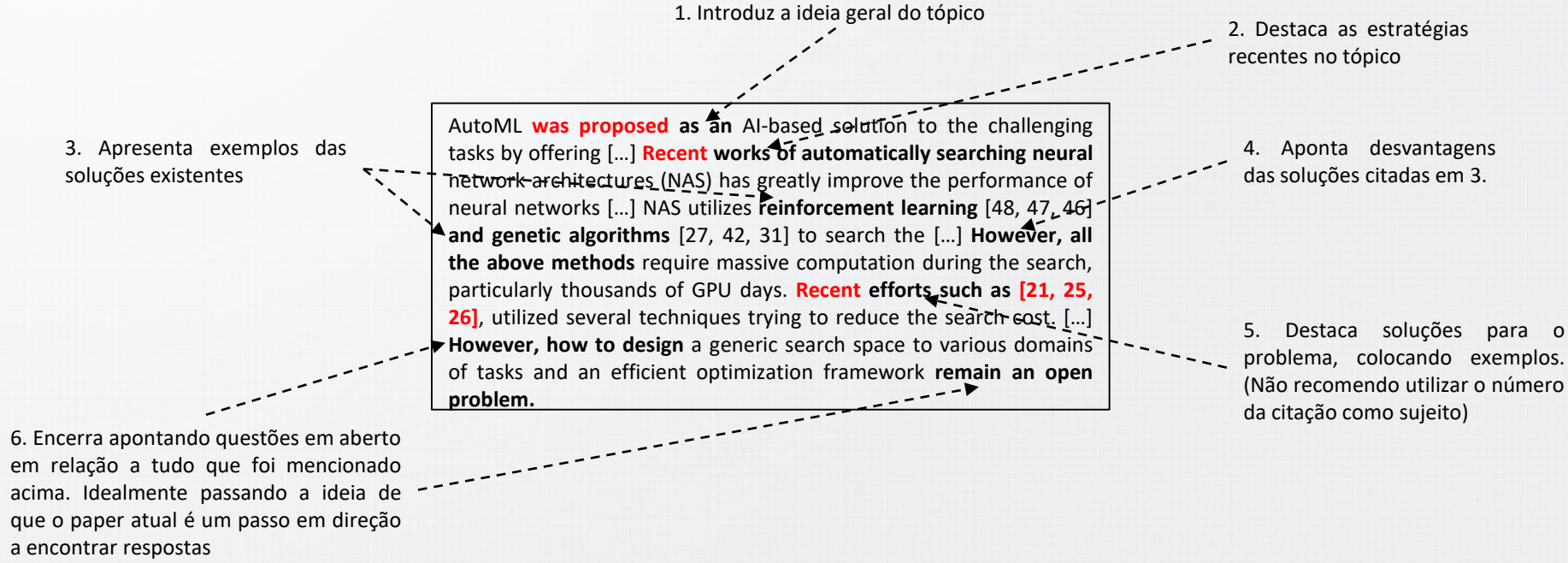
- Fonte: Mirzasoleiman et al. *Coresets for Data-efficient Training of Machine Learning Models.* ICML, 2020



Exemplos

Trabalhos Relacionados

- Fonte: Li et al. *AM-LFS: AutoML for Loss Function Search*. ICCV, 2019



Exemplos

Trabalhos Relacionados

- Fonte: Jordão et al. *When Layers Play the Lottery, all Tickets Win at Initialization*. ICCV 2023

1. Aponta onde alguns esforços da literatura estão sendo dedicados

2. Descreve que apesar desses esforços ainda existem questões a serem exploradas – lacuna na literatura

[…]. Hence, **many studies have focused on** removing layers instead of other structures [38, 8, 44, 42]. **Unfortunately, none of these efforts** have been done in the direction of LTH and pruning at initialization. **To bridge this gap, we take a step towards** understanding the behavior of LTH when the pruning process removes layers.

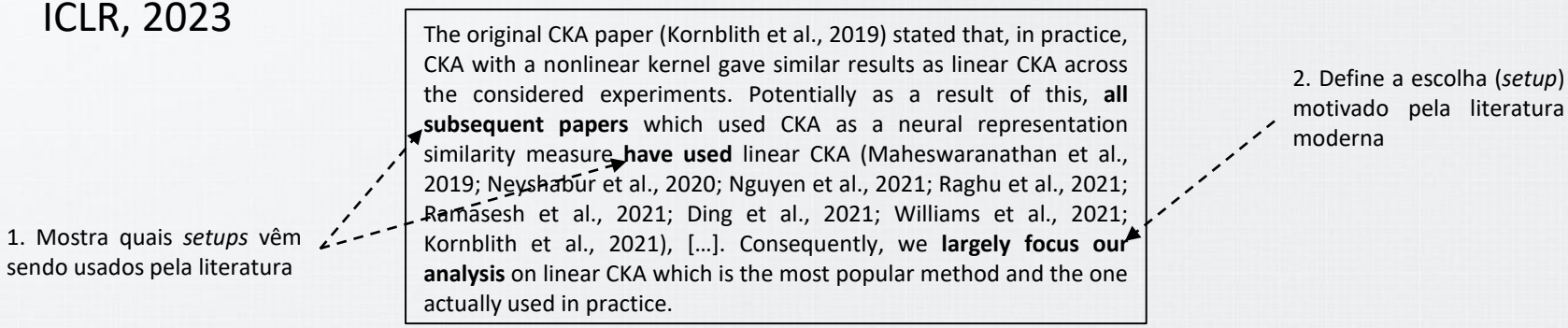
3. Finaliza explicitamente apontando qual lacuna o trabalho preenche. Mais especificamente, o que foi levantado em 2.

Exemplos

Trabalhos Relacionados

- Justificando escolhas
  - Podemos usar o *related work* para fazer um "*warm up*" das escolhas metodológicas e experimentais do artigo. Nessa direção cuidado para não especificar muito (não é o papel principal do *related work*)

- Fonte: Davari et al. *Reliability of CKA as a Similarity Measure in Deep Learning*, ICLR, 2023



O ponto 2 é importante porque sempre existirá n opções para qualquer *setup* metodológico/experimental. Usar a literatura como motivação para definir *setups* deixa a mensagem "Sabemos que existem mais opções, mas estamos seguindo as escolhas do estado da arte"

Exemplos

Trabalhos Relacionados

- Destacar contribuição em relação ao conhecimento existente

- Fonte: Lei et al. *A Comprehensive Survey of Dataset Distillation*. PAMI, 2023

1. Introduz um trabalho que se correlaciona com o artigo atual

2. Descreve a diferença para o trabalho mencionado

We note that there is a concurrent survey on dataset distillation [37]. **The most notable difference between [37] and our survey is** the taxonomy of DD. **It plainly classifies** DD algorithms into four different categories of meta-model matching, gradient matching, trajectory matching, and distribution matching, while our hierarchical taxonomy first introduces […] **Therefore, our taxonomy provides a more** systematic […]

3. Explica brevemente a singularidade que constitui a diferença no trabalho

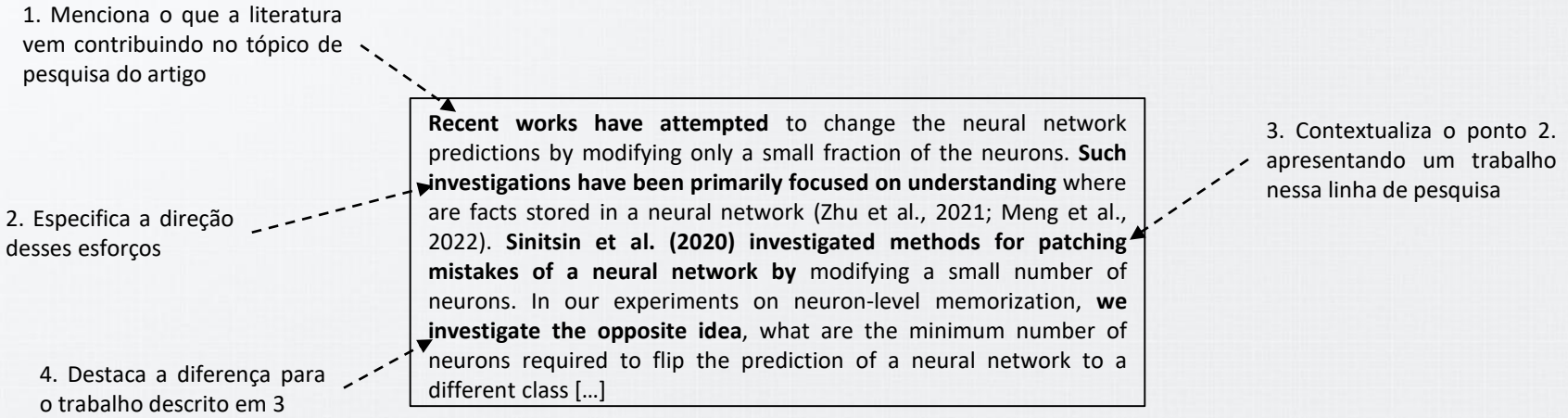
4. Conclui a sentença explicitando qual a diferença para o trabalho mencionado anteriormente, enfatizando o ponto levantado em 3. Observe que devido ao ponto 3. o artigo atual traz contribuições

Exemplos

Trabalhos Relacionados

- Destacar contribuição em relação ao conhecimento existente

- Fonte: Maini et al. *Can Neural Network Memorization Be Localized?*. ICML, 2023



Definição do Problema e Preliminares

Características e Finalidade

Definição do Problema e Preliminares

- Uma parte essencial do artigo é a descrição **formal** do problema
  - Formalismo matemático elimina ambiguidade
  - Evita que o leitor não entenda o que está sendo feito

- Descreve as variáveis do problema e o objetivo do trabalho a partir de uma descrição formal

- Dependendo da complexidade do formalismo essa seção pode aparecer
  - Na própria introdução (como subseção)
  - Antes do *related wo*rk
  - Após o *related work*

Exemplos

Definição do Problema e Preliminares

- Fonte: Lei et al. *A Comprehensive Survey of Dataset Distillation*. PAMI, 2024

1. Introduz como visualizar a descrição de um tema a partir de um formalismo matemático

2.1 Formalizing Dataset Distillation

Given a target dataset (source training dataset) $\mathbf{T} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{m}}$, the **objective of dataset distillation** is to extract the knowledge of T into a small synthetic dataset $\boldsymbol{S} = \{(\boldsymbol{s_j}, \boldsymbol{y_j})\}_{\boldsymbol{j=1}}^{\boldsymbol{n}}$, where $n \ll m$ [...]

2. A partir do formalismo em 1. descreve o objetivo do tema do artigo

Exemplos

Definição do Problema e Preliminares

- Fonte: Davari et al. *Reliability of CKA as a Similarity Measure in Deep Learning*. ICLR, 2023

1. Define variáveis e descreve o que elas representam em determinado problema

Let $\mathbf{X} \in \mathbb{R}^{n \times d_1}$ **denote** a set of ANN internal representations, i.e., the neural activations of a specific layer with d_1 **neurons in a network**, in response to $n \in N$ **input examples**. **Let** $\mathbf{Y} \in \mathbb{R}^{n \times d_2}$ **be another set** of such representations […] **We are interested in** representation similarity measures, which try to capture a certain notion of **similarity between X and Y**.

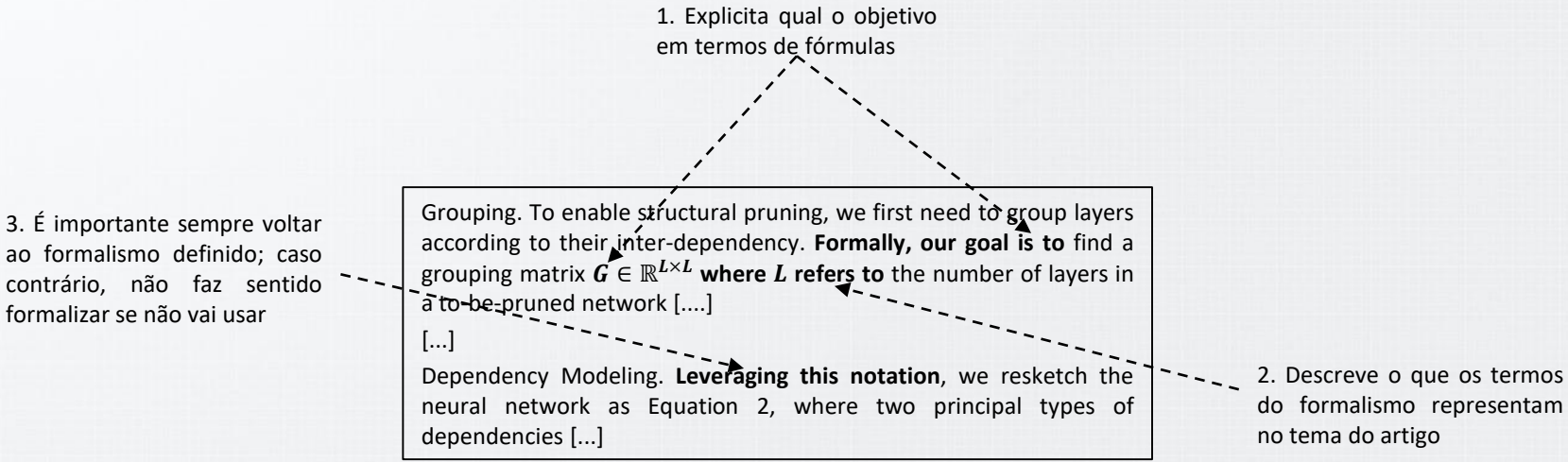
3. A partir da descrição formal, menciona o objetivo do estudo em relação às variáveis descritas em 1.

Note que os autores explicitamente destacam no que estão interessados; isso elimina qualquer brecha para "por que também não fizerem isso?"

Exemplos

Definição do Problema e Preliminares

- Fonte: Fang et al. *DepGraph: Towards Any Structural Pruning*. CVPR, 2023



Exemplos

Definição do Problema e Preliminares

- Fonte: Kolossov et al. *Towards a Statistical Theory of Data Selection Under Weak Supervision*. ICLR, 2024
  - Oral

1. Introduz algumas definições formais envolvendo o tópico

Given a sample of size N, it is often useful to select a subsample of smaller size $n < N$ to be used for statistical estimation or learning. Such a data selection step is useful to reduce the requirements of data labeling and the computational complexity of learning. **We assume to be given N unlabeled samples $\{x_i\}_{i \leq N}$, and to be given access to a 'surrogate model' that can predict labels y_i** better than random guessing. **Our goal is to select a subset of the samples, to be denoted by $\{x_i\}_{i \in G}$, of size $|G| = n < N$.** We then acquire labels for this set and we use them to train a model via regularized empirical risk minimization.

2. Descreve as restrições impostas durante a exploração do problema, delimitando o escopo do trabalho

3. Após apresentar as notações e *constraints*, explica o foco do trabalho

Exemplos

Definição do Problema e Preliminares

- Fonte: Gorishniy et al. *TabR: Tabular Deep Learning Meets Nearest Neighbors in 2023*. ICLR, 2024

1. Introduz o formalismo do problema

3. A partir do formalismos apresentados, explica o trabalho e suas particularidades

3.1 PRELIMINARIES

Notation. For a given supervised learning problem on tabular data, **we denote the dataset as** $\{(x_i, y_i)\}_{i=1}^{n}$ **where** $x_i \in X$ **represents the** i**-th object's features and** $y_i \in \mathbb{Y}$ **represents the** i**-th object's label**. Depending on the context, the i index can be omitted. **We consider three types of tasks: binary classification** $\mathbb{Y} = \{0, 1\}$**, multiclass classification** $\mathbb{Y} = \{1, \ldots, C\}$ **and regression** $\mathbb{Y} = \mathbb{R}$. For simplicity, [...] The dataset is split into three disjoint parts: $1, n = I_{train} \cup I_{val} \cup I_{test}$, where the "train" part [...]

Let's consider a generic feed-forward retrieval-free network $f(x) = P(E(x))$ informally partitioned into two parts: encoder $E : \mathbb{X} \to \mathbb{R}^d$ **and predictor** $\mathbb{P}: \mathbb{R}^d \to \hat{\mathbb{Y}}$**.** To incrementally make it retrieval-based, [...]

2. Aponta o foco do trabalho de acordo com o formalismo

Exemplos

Definição do Problema e Preliminares

- Fonte: Duong et al. *Representational Dissimilarity Metric Spaces for Stochastic Neural Networks*. ICLR, 2023

1. Introduz as definições e variáveis necessárias

> [...] Intuitively, we desire a notion of distance such that $\boldsymbol{d}(\boldsymbol{X_i}, \boldsymbol{X_j}) = \boldsymbol{d}(\boldsymbol{X_i}, \boldsymbol{X_j}\boldsymbol{\Pi})$ for any permutation matrix, $\boldsymbol{\Pi} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ **[...] Let** $\boldsymbol{\phi_k} : \mathbb{R}^{\boldsymbol{M}\times\boldsymbol{nk}} \mapsto \mathbb{R}^{\boldsymbol{M}\times\boldsymbol{n}}$ be a fixed, "preprocessing function" for each network and let $\boldsymbol{G}$ be a set of nuisance transformations on $\mathbb{R}^{\boldsymbol{n}}$. Williams et al. (2021) showed that:
>
> $$d(X_i, X_j) = \min_{T\in G} \left|\left|\phi_i(X_i) - \phi_j(X_j)T\right|\right|_F \qquad (2)$$
>
> [...] **How can Equation** (2) **be generalized to measure representational distances in this case?**

2. Após introduzir as definições, levanta questões de pesquisa usando o formalismo descrito

Notem que não há definições que não são utilizadas (G é usado posteriormente)

Definições que devem ser evitadas

Definição do Problema e Preliminares

- Evite formalismos que nunca serão utilizados

- Nunca descreva “fórmulas” óbvias
  - Por exemplo, média, desvio padrão, normalização max-min, Z-score. Exceção: modificações ou variações para o problema
  - Métricas bem conhecidas na literatura (e.g., acurácia). Nesse caso, opte por citar um artigo que descreve a métrica, preferencialmente *surveys* ou *benchmarks*

Óbvio e prolixo

Let μ be the average across all samples calculated in terms of $\mu \leftarrow \frac{1}{n-1}\sum_{i=1}^{n} x_i$. Let σ denote the standard deviation estimated by $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu) * (x_i - \mu)}$. After estimating these metrics, we normalize the data using this average and the standard deviation in terms of $x \leftarrow \frac{x-\mu}{\sigma}$ before starting analyzing them.

Conciso e refinado

It is standard practice to center the data column-wise (z-score) before analyzing them.

Experimentos

Características e Finalidade

Experimentos

- Confirmar hipóteses e questões de pesquisas levantadas (idealmente na introdução)
  - Mostra o que foi prometido na introdução

- Corroborar ou refutar *insights* da literatura

- Questões que devem ser pensadas **após** concluir os experimentos
  - Todas as questões de pesquisas levantadas possuem, no mínimo, um experimento associado?
  - Uma hipótese levantada demonstrou-se verdadeira? Por qual experimento?
  - **Toda afirmação (sem referência) no texto possui um experimento para apoiá-la?**
  - Algum experimento deixou questões em aberto (*room for improvement*). Em caso afirmativo, porque não responder no artigo atual?

Exemplos

Experimentos

- Refutar/Corroborar evidências da literatura

- Fonte: Davari et al. *Reliability of CKA as a Similarity Measure in Deep Learning*. ICLR 2023.

Fig. 5 shows the CKA map of f_{θ^*} along with the CKA map of three scenarios we investigated […]. This is **surprising and contradictory to the previous findings** (Kornblith et al., 2019; Raghu et al., 2021) as it suggests that it is possible to achieve a strong […]. **Similarly we observe that** […]

1. Contrasta a descoberta obtida no experimento com observações de trabalhos anteriores

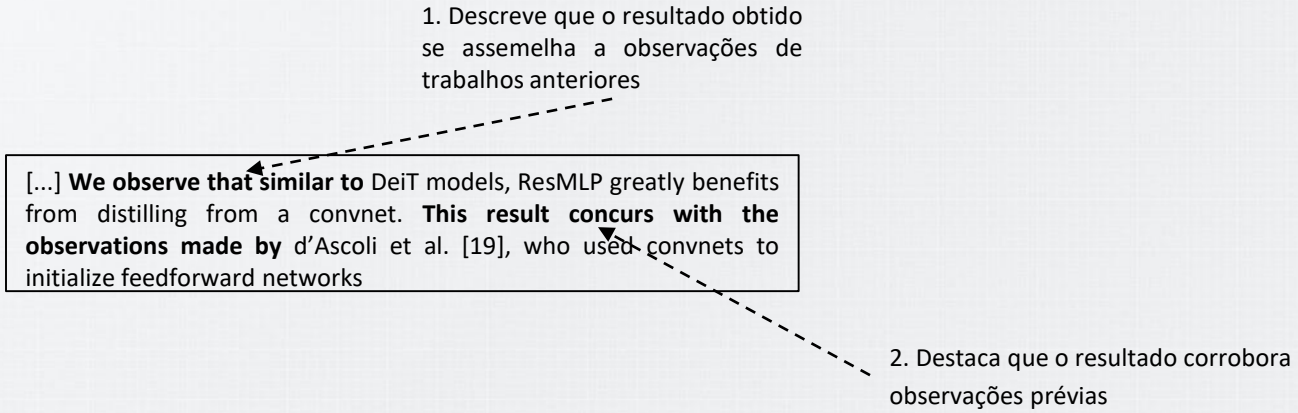
2. Corrobora que alguns fatores são similares a descobertas anteriores

Exemplos

Experimentos

- Refutar/Corroborar evidências da literatura

- Fonte: Touvron et al. *ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training*. PAMI, 2023



Exemplos

Experimentos

- Revisitar questões de pesquisas

- Fonte: Pham et al. *Towards Data-Agnostic Pruning At Initialization: What Makes a Good Sparse Mask?.* NeurIPS 2023

1. Após abordar os aspectos principais do experimento, finaliza descrevendo qual questão de pesquisa a figura ajuda responder (mesmo que parcialmente)

Figure 5 shows the experimental results for different […] Results of all settings in Figure 5 **provide evidence to support our** Node-Path Balancing principle and the existence of a specific balancing region between nodes and paths at given […]

Exemplos

Experimentos

- Justificar escolhas

- Fonte: Fang et al. *DepGraph: Towards Any Structural Pruning.* CVPR, 2023

1. Relembra o leitor qual o foco do trabalho. Principalmente se o objetivo não é bater algum método do estado da arte

[...] **The target of this work is not** to provide state-of-the-art results for various models, **thus we only use** the most basic importance criterion in this work.

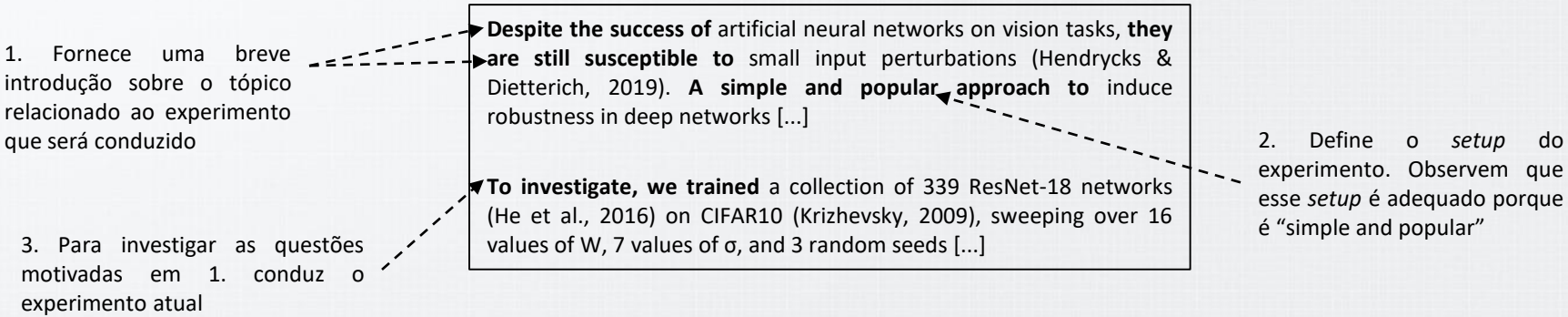
2. Com base na descrição usada em 1. descreve que a escolha experimental, dentre as possíveis, é adequada para alcançar o objetivo do trabalho

Exemplos

Experimentos

- Iniciar o experimento introduzindo sua motivação

- Fonte: Duong et al. *Representational Dissimilarity Metric Spaces for Stochastic Neural Networks.* ICLR, 2023



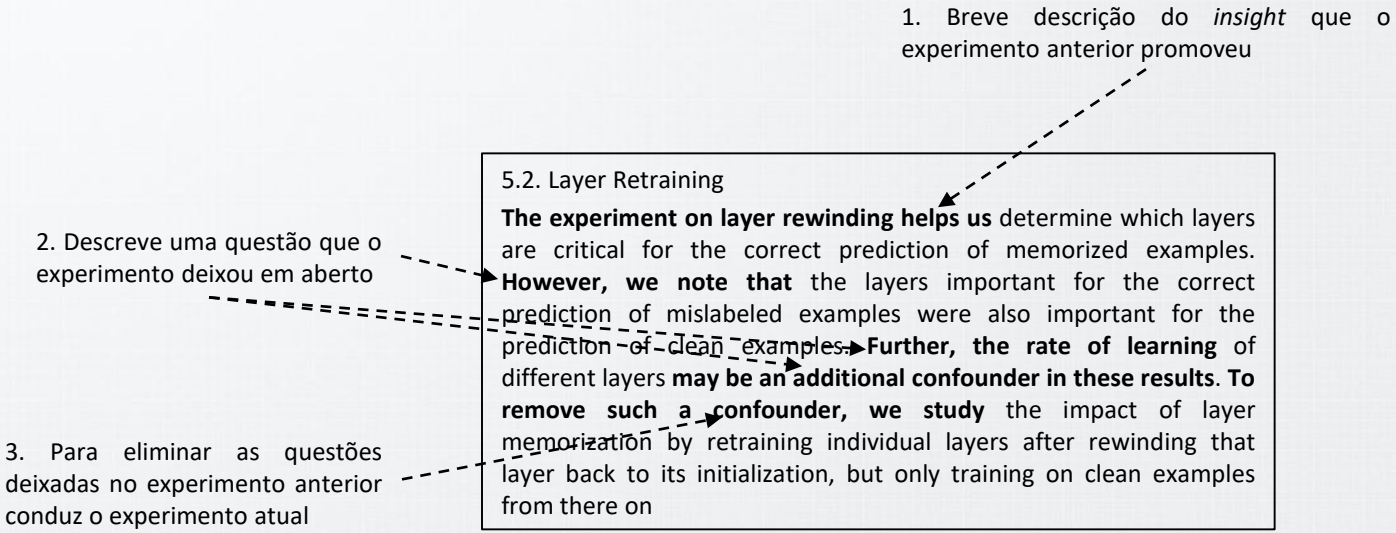
Notem que devido à motivação levantada no ponto 1 o experimento torna-se importante

Exemplos

Experimentos

- Experimentos que motivam experimentos

- Fonte: Maini et al. *Can Neural Network Memorization Be Localized?*. ICML, 2023

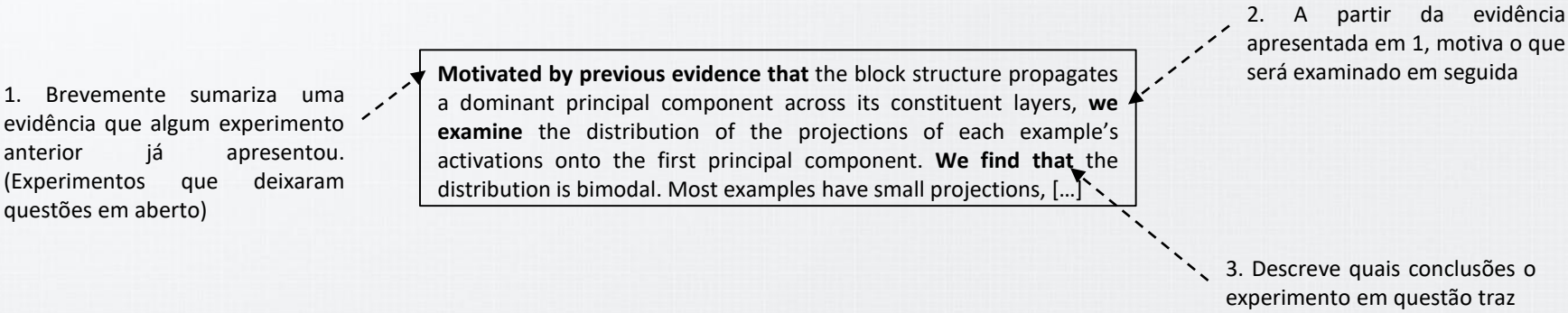


Exemplos

Experimentos

- Experimentos que motivam experimentos

- Fonte: Nguyen et al. *On the Origins of the Block Structure Phenomenon in Neural Network Representations.* Transactions on Machine Learning Research, 2022



Exemplos

Experimentos

- Experimentos que motivam experimentos

- Fonte: Nguyen et al. *On the Origins of the Block Structure Phenomenon in Neural Network Representations.* Transactions on Machine Learning Research, 2022

1. Brevemente sumariza o que os experimentos anteriores demonstraram

2. A partir do que foi demonstrado, motiva o que será examinado neste experimento (ou conjunto de experimentos)

In the previous section, we characterize the signals the block structure propagates across its layers, and explore their implications on other aspects of the network internals. **Informed by these findings, we next explore what happens to** the block structure and the dominant images over time, from initialization until the model converges, and how this process varies across different training runs.

[...]

Overall these findings suggest that the uniqueness of the block structure representations in large-capacity models can be attributed to both initialization parameters and the image minibatches the models receive throughout training.

3. Conclui quais descobertas os resultados obtidos neste experimento trouxeram

Exemplos

Experimentos

- Experimentos que corroboram experimentos

- Fonte: Masarczyk et al. *The Tunnel Effect: Building Data Representations in Deep Neural Networks.* NeurIPS, 2023

1. Apresenta as descobertas do experimento (em particular, da figura)

[...]

Figure 3 reveals that for VGG-19 the inter-class representations variation decreases throughout the tunnel, meaning that representations [...] **This view aligns with the observation from Figure 2**, where the rank of the representations drops [...]

2. Reforça que esse insight está alinhado com observações apresentadas em experimentos anteriores

Exemplos

Experimentos

- Experimentos que corroboram experimentos

- Fonte: Masarczyk et al. *The Tunnel Effect: Building Data Representations in Deep Neural Networks.* NeurIPS, 2023

1. Apresenta as descobertas do experimento

[...]
In all tested scenarios, **we observe a consistent** relationship between the start of the tunnel and the drop in OOD performance. [...] **This observation aligns with our earlier findings suggesting** that the tunnel is a prevalent characteristic of the model rather than [...]

2. Reforça que esse *insight* está alinhado com observações apontadas em experimentos anteriores

Exemplos

Experimentos

- Concluindo experimentos

- Fonte: *Maini et al. Can Neural Network Memorization Be Localized?*. ICML, 2023

1. Introduz o objetivo do experimento

2. Menciona onde os resultados obtidos podem ser encontrados

To assess the impact of the choice of values of p_{gen} and p_{mem} on the ability of example-tied dropout to localize memorization in neurons, we run a grid search on 12 different parameter combinations. **Results for the** change in the efficacy of the method with a change in these values **are presented in Table 2. We find that** the method is robust to a large range of values (and combinations) of p_{gen} and p_{mem}. In general, **we observe the trend that** as we increase the capacity of the generalization neurons [...]

3. Conclui o experimento

O ponto 3 é muito importante: **nunca concluam um experimento com "Tabela 2 mostra os resultados obtidos"**

Exemplos

Experimentos

- Concluindo experimentos

- Fonte: Gorishniy et al. *TabR: Tabular Deep Learning Meets Nearest Neighbors In 2023.* ICLR, 2024

1. Discute os resultados obtidos com o experimento

2. Conclui o experimento enfatizando a mensagem principal deixada por ele

The obtained results highlight the retrieval technique and embeddings for numerical features (Gorishniy et al., 2022) (used in MLP-PLR and TabR) as two powerful architectural elements that improve the optimization properties of tabular DL models. **Interestingly, the two techniques** are not fully orthogonal, but none of them can recover the full power of the other,[…].

The main takeaway. TabR becomes a new strong deep learning solution for tabular data problems and demonstrates a good potential of the retrieval-based approach. TabR demonstrates strong average performance and achieves the new state-of-the-art on several datasets.

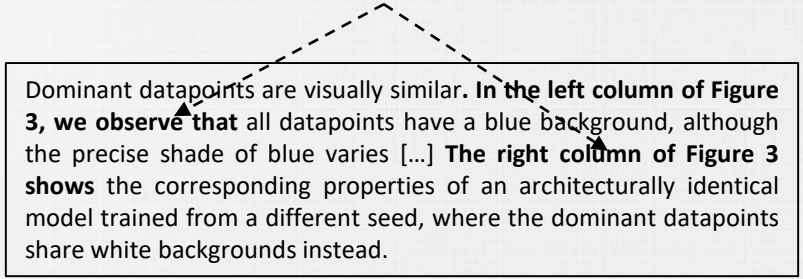
Exemplos

Experimentos

- Explicando figuras

- Fonte: Maini et al. *Can Neural Network Memorization Be Localized?*. ICML, 2023

1. Explicita o que deve ser notado na figura (Atenção para não ter redundância com o *caption*)



Exemplos

Experimentos

- Explicando figuras

- Fonte: Chen Et Al. *Which Layer is Learning Faster? A Systematic Exploration of Layer-Wise Convergence Rate for Deep Neural Networks*. ICLR, 2023

1. Aponta os principais pontos que devem ser notados na figura

In the following analysis, the first fully-connected layer (1-32) […]. The gradient values and the convergence processes for these layers are shown in Fig. 1 (a). **Two observations can be obtained from the plots**: 1) The gradient of Hidden layer 1 is nearly always smaller than the gradient of Hidden layer 2. 2) Although shallower layers have smaller gradients, they seem to converge faster. […]

Visão Geral

Experimentos

- Uma boa prática é iniciar os experimentos com o setup experimental
  - Definição dos datasets/*benchmarks*
  - *Setup* da máquina
  - *Setup* do treinamento e calibragem de parâmetros
  - Arquiteturas utilizadas

> We conduct experiments on three standard datasets: CIFAR-10, […]. Following [42], we use ResNet-20 for CIFAR-10, VGG-19 for CIFAR-100 […] We run five seeds with experiments on CIFAR-10[…] More details on our experimental setting are in Appendix A

Fonte: Pham et al. *Towards Data-Agnostic Pruning At Initialization: What Makes a Good Sparse Mask?* NeurIPS 2023

> We evaluate HEAD2TOE on the VTAB-1k benchmark using two popular vision architectures, ResNet-50 (Wu et al., 2018) and ViT-B/16 (Dosovitskiy et al., 2021), both pretrained on ImageNet-2012.
>
> [...] We perform five-fold cross validation for each task and method in order to pick the best hyperparameters. All methods search over the same learning rates and training steps (two values of each). [...] We repeat each evaluation using 3 different seeds and report median values and share standard deviations in Appendix D. More details on hyperparameter selection and values used are shared in Appendix A.

Fonte: Evci et al. *Head2Toe: Utilizing Intermediate Representations for Better Transfer Learning*. ICML, 2022

Visão Geral

Experimentos

- Idealmente a seção/subseção de um experimento deveria mencionar a ideia central por trás do experimento
  - Por exemplo, (i) qual questão de pesquisa o experimento explora/responde? (ii) Quais "promessas" feitas na introdução o experimento "cumpre"?

> In our experimental evaluation we wish to **address the following questions**: (1) How do loss and accuracy of IG applied to the subsets returned by CRAIG compare to loss and accuracy of IG applied to the entire data? (2) How small is the size of the subsets that we can select with CRAIG and still get a comparable performance to IG applied to the entire data? And (3) How well does CRAIG scale to large data sets, and extends to non-convex problems?

Fonte: Mirzasoleiman et al. *Coresets for Data-efficient Training of Machine Learning Models.* ICML, 2020

> 3.3. Outline of experimental results
>
> The experiments presented in Sections 4 and 5 **aim to answer a fundamental question: what happens when neural networks lose plasticity?** […]
>
> **Section 5 asks what properties cause plasticity loss?** […]
>
> **Section 6.2 addresses the question: how can we mitigate plasticity loss?** […]

Fonte: Lyle et al. *Understanding Plasticity in Neural Networks*. ICML, 2023

Teaser

Características e Finalidade

Teaser

- *Teaser* é uma figura que vem logo no início do artigo
  - Tipicamente ao lado do *abstract* ou na primeira página
- Apresentar resultados (quantitativos e qualitativos)
- Mostrar a ideia central do trabalho
- **Convencer visualmente o leitor que o trabalho é interessante e relevante**
- O *caption* do *teaser* precisa deixar a figura autocontida
  - O leitor precisa olhar e compreender a ideia da figura

Teaser

- Apresentar resultados

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

Abstract

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we

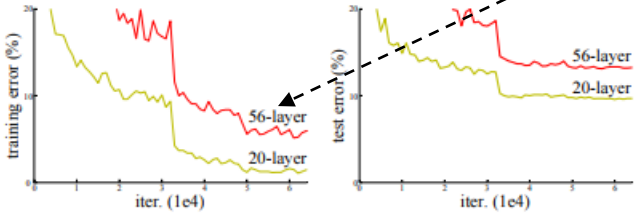


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

1. (Para entendedores) Esse comportamento chama muito a atenção. Após ver essa figura torna-se obrigatório ler o artigo

2. O *caption* reforça o que a figura quer ilustrar. Além disso, ele complementa mencionando que o comportamento ilustrado ocorre em outros cenários (não mostrados na ilustração)

Teaser

- Apresentar resultados

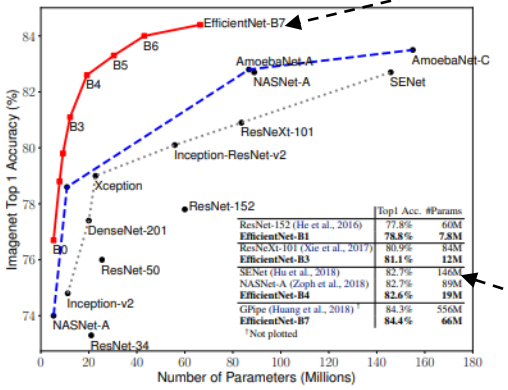
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Mingxing Tan [1] Quoc V. Le [1]

Abstract

Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more resources are available. In this paper, we systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance. Based on this observation, we propose a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective *compound coefficient*. We demonstrate the effectiveness of this method on scaling up MobileNets and ResNet.

To go even further, we use neural architecture search to design a new baseline network and scale it up to obtain a family of models, called *EfficientNets*, which achieve much better accuracy and efficiency than previous ConvNets. In particular, our EfficientNet-B7 achieves state-of-the-art 84.4%



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**78.8%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.1%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.6%**	**19M**
GPipe (Huang et al., 2018) [1]	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

[1]Not plotted

Figure 1. **Model Size vs. ImageNet Accuracy.** All numbers are for single-crop, single-model. Our EfficientNets significantly outperform other ConvNets. In particular, EfficientNet-B7 achieves new state-of-the-art 84.4% top-1 accuracy but being 8.4x smaller and 6.1x faster than GPipe. EfficientNet-B1 is 7.6x smaller and 5.7x faster than ResNet-152. Details are in Table 2 and 4.

1. Destaca como o método proposto não é superado pela literatura em ambos os eixos simultaneamente (Pareto). Isso é, a solução proposta domina as demais da literatura

2. Apresenta resultados quantitativos para complementar a visualização. Isso facilita ao leitor observar, quantitativamente, os ganhos que o método obtém

Teaser

- Apresentar resultados

ALOFT: A Lightweight MLP-like Architecture with Dynamic Low-frequency Transform for Domain Generalization

Jintao Guo[1,2] Na Wang[1,2] Lei Qi[3*] Yinghuan Shi[1,2*]
[1] State Key Laboratory for Novel Software Technology, Nanjing University
[2] National Institute of Healthcare Data Science, Nanjing University
[3] School of Computer Science and Engineering, Southeast University
{guojintao, wangna}@smail.nju.edu.cn, qilei@seu.edu.cn, syh@nju.edu.cn

Abstract

Domain generalization (DG) aims to learn a model that generalizes well to unseen target domains utilizing multiple source domains without re-training. Most existing DG works are based on convolutional neural networks (CNNs). However, the local operation of the convolution kernel makes the model focus too much on local representations (e.g., texture), which inherently causes the model more prone to overfit to the source domains and hampers its generalization ability. Recently, several MLP-based methods have achieved promising results in supervised learning tasks by learning global interactions among different patches of the image. Inspired by this, in this paper, we first analyze the difference between CNN and MLP methods in DG and find that MLP methods exhibit a better generalization ability because they can better capture the global representations (e.g., structure) than CNN methods.

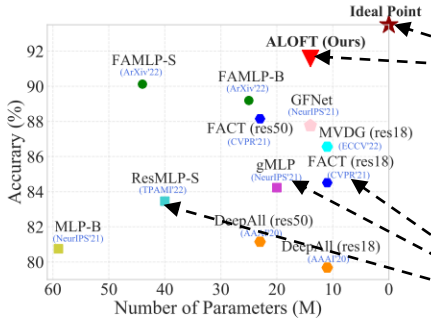


Figure 1. Comparison of the SOTA CNN-based methods, the latest MLP-like models, and our method on PACS. Among the SOTA CNN-based and MLP-based methods, our method can achieve the best performance with a relatively small-sized network.

1. Notem o *Ideal Point* em uma curva de Pareto e como o método proposto se aproxima desse ponto

2. Coloca a referência nos demais métodos para destacar que está superando artigos de importantes veículos de publicação (CVPR, PAMI, NeurIPS, etc.)

Teaser

- Apresentar resultados

PIXMIX: Dreamlike Pictures Comprehensively Improve Safety Measures

Dan Hendrycks* UC Berkeley
Andy Zou* UC Berkeley
Mantas Mazeika UIUC
Leonard Tang Harvard University
Bo Li UIUC
Dawn Song UC Berkeley
Jacob Steinhardt UC Berkeley

Abstract

In real-world applications of machine learning, reliable and safe systems must consider measures of performance beyond standard test set accuracy. These other goals include out-of-distribution (OOD) robustness, prediction consistency, resilience to adversaries, calibrated uncertainty estimates, and the ability to detect anomalous inputs. However, improving performance towards these goals is often a balancing act that today's methods cannot achieve without sacrificing performance on other safety axes. For instance, adversarial training improves adversarial robustness but sharply degrades other classifier performance metrics. Similarly, strong data augmentation and regularization techniques often improve OOD robustness but harm anomaly detection, raising the question of whether a Pareto improvement on all existing safety measures is possible. To meet this challenge, we design a new data augmentation strategy utilizing the natural structural complexity of pictures such as fractals, which outperforms numerous baselines, is near Pareto-optimal, and roundly improves safety measures.

1. Introduction

A central challenge in machine learning is building models that are reliable and safe in the real world. In addition to performing well on the training distribution, deployed models should be robust to distribution shifts, consistent in

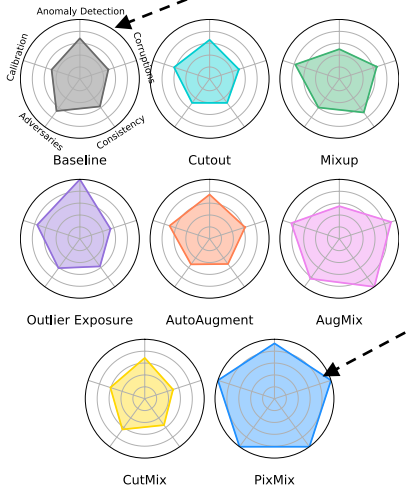


Figure 1. Normalized performance of different methods on five different model safety measures. PIXMIX is the only method that significantly outperforms the baseline in all five safety measures.

1. Define a legenda comum para todos os eixos do radar plot

2. De acordo com o radar plot, o método proposto se destaca em diferentes métricas de qualidade

Teaser

- Apresentar resultados

Apresenta o comportamento da visualização proposta em duas arquiteturas: uma conhecida por enfrentar colapso durante o treinamento (a) e outra que contorna esse problema (b). Desta forma, os autores reforçam a habilidade de visualização capturada pela abordagem proposta

Visualizing the Loss Landscape of Neural Nets

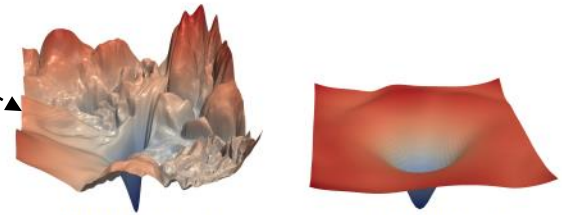
Hao Li[1], **Zheng Xu**[1], **Gavin Taylor**[2], **Christoph Studer**[3], **Tom Goldstein**[1]
[1]University of Maryland, College Park [2]United States Naval Academy [3]Cornell University
{haoli,xuzh,tomg}@cs.umd.edu, taylor@usna.edu, studer@cornell.edu

Abstract

Neural network training relies on our ability to find "good" minimizers of highly non-convex loss functions. It is well-known that certain network architecture designs (e.g., skip connections) produce loss functions that train easier, and well-chosen training parameters (batch size, learning rate, optimizer) produce minimizers that generalize better. However, the reasons for these differences, and their effect on the underlying loss landscape, are not well understood. In this paper, we explore the structure of neural loss functions, and the effect of loss landscapes on generalization, using a range of visualization methods. First, we introduce a simple "filter normalization" method that helps us visualize loss function curvature and make meaningful side-by-side comparisons between loss functions. Then, using a variety of visualizations, we explore how network architecture affects the loss landscape, and how training parameters affect the shape of minimizers.

1 Introduction

Training neural networks requires minimizing a high-dimensional non-convex loss function – a task that is hard in theory, but sometimes easy in practice. Despite the NP-hardness of training general neural loss functions [3], simple gradient methods often find global minimizers (parameter configurations with zero or near-zero training loss), even when data and labels are randomized before training [43]. However, this good behavior is not universal; the trainability of neural nets is highly dependent on network architecture design choices, the choice of optimizer, variable initialization, and a variety of other considerations. Unfortunately, the effect of each of these choices on the structure of the underlying loss surface is unclear. Because of the prohibitive cost of loss function evaluations (which requires looping over all the data points in the training set), studies in this field have remained predominantly theoretical.



(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Teaser

- Mostrar a ideia central do trabalho

A figura não é trivial. Porém, observe como o *abstract* ajuda a explicá-la (*first gets worse and then gets better*)

DEEP DOUBLE DESCENT: WHERE BIGGER MODELS AND MORE DATA HURT

Preetum Nakkiran* Harvard University
Gal Kaplun† Harvard University
Yamini Bansal† Harvard University
Tristan Yang Harvard University
Boaz Barak Harvard University
Ilya Sutskever OpenAI

ABSTRACT

We show that a variety of modern deep learning tasks exhibit a "double-descent" phenomenon where, as we increase model size, performance first gets *worse* and then gets better. Moreover, we show that double descent occurs not just as a function of model size, but also as a function of the number of training epochs. We unify the above phenomena by defining a new complexity measure we call the *effective model complexity* and conjecture a generalized double descent with respect to this measure. Furthermore, our notion of model complexity allows us to identify certain regimes where increasing (even quadrupling) the number of train samples actually *hurts* test performance.

1 INTRODUCTION

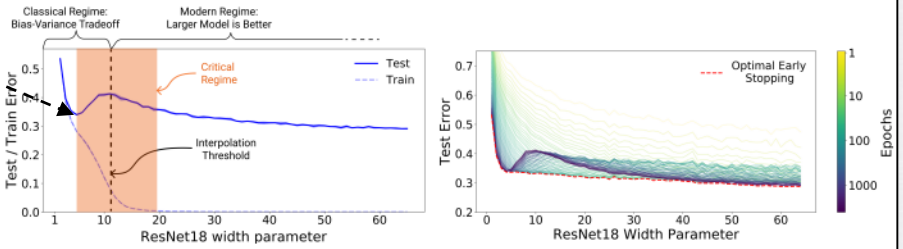


Figure 1: **Left:** Train and test error as a function of model size, for ResNet18s of varying width on CIFAR-10 with 15% label noise. **Right:** Test error, shown for varying train epochs. All models trained using Adam for 4K epochs. The largest model (width 64) corresponds to standard ResNet18.

Teaser

- Mostrar a ideia central do trabalho

A combinação do *abstract* com o *teaser* deve ser harmônica. Aqui o *abstract* define as abreviações usadas no *teaser*. Podemos entender o *teaser* como uma continuação visual do *abstract*

Searching for A Robust Neural Architecture in Four GPU Hours

Xuanyi Dong[1,2], Yi Yang[1]
[1]University of Technology Sydney [2]Baidu Research
xuanyi.dong@student.uts.edu.au, yi.yang@uts.edu.au

Abstract

Conventional neural architecture search (NAS) approaches are based on reinforcement learning or evolutionary strategy, which take more than 3000 GPU hours to find a good model on CIFAR-10. We propose an efficient NAS approach learning to search by gradient descent. Our approach represents the search space as a directed acyclic graph (DAG). This DAG contains billions of sub-graphs, each of which indicates a kind of neural architecture. To avoid traversing all the possibilities of the sub-graphs, we develop a differentiable sampler over the DAG. This sampler is learnable and optimized by the validation loss after training the sampled architecture. In this way, our approach can be trained in an end-to-end fashion by gradient descent, named Gradient-based search using Differentiable Architecture Sampler (GDAS). In experiments, we can finish one searching procedure in four GPU hours on CIFAR-10, and the discovered model obtains a test error of 2.82% with only 2.5M parameters, which is on par with the state-of-the-art.

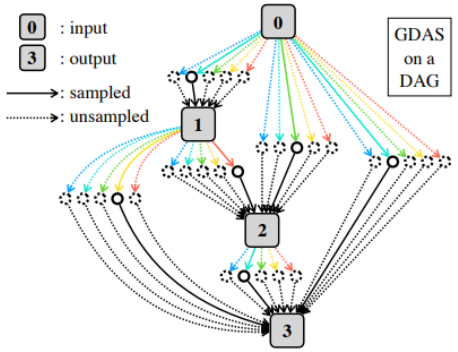


Figure 1. We utilize a DAG to represent the search space of a neural cell. Different operations (colored arrows) transform one node (square) to its intermediate features (little circles). Meanwhile, each node is the sum of the intermediate features transformed from the previous nodes. As indicated by the **solid** connections, the neural cell in the proposed GDAS is a sampled sub-graph of this DAG. Specifically, among the intermediate features between every two nodes, GDAS samples one feature in a differentiable way.

1. Introduction

Teaser

- Mostrar a ideia central do trabalho

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 46, NO. 1, JANUARY 2024

A Comprehensive Survey of Dataset Distillation

Shiye Lei and Dacheng Tao, *Fellow, IEEE*

(Survey paper)

***Abstract*—Deep learning technology has developed unprecedentedly in the last decade and has become the primary choice in many application domains. This progress is mainly attributed to a systematic collaboration in which rapidly growing computing resources encourage advanced algorithms to deal with massive data. However, it has gradually become challenging to handle the unlimited growth of data with limited computing power. To this end, diverse approaches are proposed to improve data processing efficiency. Dataset distillation, a dataset reduction method, addresses this problem by synthesizing a small typical dataset from substantial data and has attracted much attention from the deep learning community. Existing dataset distillation methods can be taxonomized into meta-learning and data matching frameworks according to whether they explicitly mimic the performance of target data. Although dataset distillation has shown surprising performance in compressing datasets, there are still several limitations such as distilling high-resolution data or data with complex label spaces. This paper provides a holistic understanding of dataset distillation from multiple aspects, including distillation frameworks and algorithms, factorized dataset distillation, performance comparison, and applications. Finally, we discuss challenges and promising directions to further promote future studies on dataset distillation.**

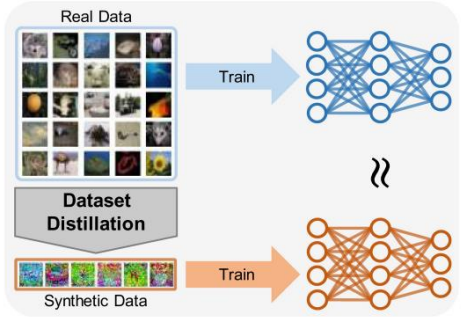


Fig. 1. Illustration of dataset distillation. The models trained on the large original dataset and small synthetic dataset demonstrate comparable performance on the test set.

Check List Geral

Check List

Check List Geral

- ❑ Toda afirmação deve ser seguida de uma referência, exceto as que o artigo demonstrará
  - Por exemplo, “Parameter count is also not well correlated with latency”. Quem confirma isso? O artigo em questão (ex. com um experimento) ou algum trabalho na literatura?
  - Afirmações sem referência ou demonstrações criam uma **lacuna** no artigo e o revisor pode facilmente usá-la para rejeitar o trabalho

- ❑ Evite escrever informação/sentença desnecessária
  - Imagine que o leitor tem uma capacidade limitada de informações que prestará atenção
  - Se o leitor começar a “pular” algumas sentenças porque **são desnecessárias**, ele pode acabar pulando sentenças que **são importantes** para o entendimento do trabalho
  - *“A **perfeição não** é alcançada quando **não** há mais nada a ser incluído. A **perfeição** é alcançada quando **não** há mais nada a ser retirado”* Antoine de Saint-Exupéry

Check List

Check List Geral

- ❑ Garanta que o texto não contém voz passiva
  - "[…] experiments were conducted" -> "We conduct experiments"
  - "can be solved in $O(N)$ complexity" -> "an algorithm solves in $O(N)$ complexity"
  - "values are presented in Table 2" -> "Table 2 presents the values"
  - Tipicamente, o uso da palavra "can" premedita o uso de voz passiva

- ❑ Verifique se o texto não possui excesso de that/which, this/these
  - Nunca na mesma sentença
  - Evitar fortemente em sentenças próximas
  - **De maneira geral, evite palavras repetidas próximas**

While these works **provide** convergence on the full dataset, our analysis **provides** the same convergence rates on subsets obtained by CRAIG.

While these works **provide** convergence on the full dataset, our analysis **promotes** the same convergence rates on subsets obtained by CRAIG.

While these works **provide** convergence on the full dataset, our analysis **leads to** the same convergence rates on subsets obtained by CRAIG.

Check List

Check List Geral

- ❑ Garanta que o fluxo das ideias entre as sentenças está coerente e suave
  - Dica: Pegue as sentenças $i-1, i, i+1$ e verifique se as ideias apresentam transição suave
  - Manter um fluxo bom é uma das tarefas cognitivas mais difíceis no processo de escrita (ChatGPT pode não ser uma alternativa viável)

Fluxo Grosseiro

Design and deployment of efficient deep learning architectures for mobile devices has seen a lot of progress with consistently decreasing floating-point operations (FLOPs) and parameter count while improving accuracy. FLOPs and parameters not correlate well with the efficiency of the models in terms of latency. Efficiency metric like FLOPs do not account for memory access cost and degree of parallelism, which can have a nontrivial effect on latency during inference. Parameter count is not well correlated with latency.

Fluxo Suave

Design and deployment of efficient deep learning architectures for mobile devices has seen a lot of progress with consistently decreasing floating-point operations (FLOPs) and parameter count while improving accuracy. **However, these** metrics **may** not correlate well with the efficiency of the models in terms of latency. Efficiency metric like FLOPs do not account for memory access cost and degree of parallelism, which can have a nontrivial effect on latency during inference. Parameter count is **also** not well correlated with latency.

Percebam como o uso de 4 palavras tornou o fluxo mais suave

Check List

Check List Geral

- ❑ Verifique (densamente) a gramática do texto

- ❑ Ferramentas úteis: Grammarly, Google Translator
  - ❑ Não coloque trechos longos, utilize para cada sentença
  - ❑ Atualize a página a cada 5-10 sentenças

- ❑ Foque em escrever/aprimorar sentenças pequenas
  - Em seguida agregue com as demais (regra do $i-1, i, i+1$) sentenças do parágrafo
  - Novamente, verifique se o fluxo das sentenças $i-1, i, i+1$ está coerente e suave

Check List

Check List Geral

- ❑ Use o ChatGPT para **refinar** sentenças. Refinar ≠ Escrever *from scratch*
  - Prompt “Assuma que você é autor e/ou editor das conferências ICLR, ICML e NeurIPS. Vou fornecer algumas sentenças e preciso que você ajude a aprimorá-las. Evite o uso de voz-passiva. Analise também se a sentença está clara e compreensiva.”

CVPR Review Policy

LLM policy: Remember that you can use an **LLM to refine your review text** if you think it is helpful. But you CAN'T show the paper to an LLM in any way, because doing so is a major violation of policy. We think we can detect people showing papers to LLMs, and we will prosecute people we catch. Don't do this.

What is the LLM Policy for referees in CVPR 2024?

(Detalhes: https://cvpr.thecvf.com/Conferences/2024/ReviewerGuidelines)

Referees may use any device, including an **LLM, to polish their review wording**, but must vouch for, and be responsible for, the accuracy of the review. It is a significant act of referee misconduct to allow an LLM to see a submission. PCs interpret showing a submission to an LLM as a deliberate referee violation of confidentiality. This would allow PCs to move to exclude the relevant referee from submitting to CVPR for up to two years. PCs intend to exercise this power.

Check List

Check List Geral

- ❑ Sempre utilizar figuras vetoriais (por exemplo, .pdf)
  - Ferramentas como *matplotlib* já possibilitam salvar figuras em formato vetorial (.pdf)
  - Um bom tamanho para figuras é ocupar uma coluna da página
  - \includegraphics[width=\columnwidth]
- ❑ Verifique se todas as figuras estão sendo citadas no texto
  - Todas as figuras devem ser discutidas – o papel das figuras é facilitar a compreensão de algum ponto discutido
- ❑ Garanta que os elementos da figura estão no mesmo idioma do artigo
  - Por exemplo, se o texto estiver em português, os eixos de um gráfico não devem estar em inglês (e vice-versa)

Check List

Check List Geral

- ❑ Nunca coloque equações e tabelas como figuras (png, jpg, etc.)
  - Em apresentações construam tabelas/equações usando o próprio ambiente (i.e., PowerPoint)

- ❑ Tabelas em LaTex:
  - https://www.tablesgenerator.com/
  - Use sempre \renewcommand{\arraystretch}{1.2}

$$w^* = (X^T X)^{-1} X^T Y$$

C1	C2	C3
1	(+) 0.08	(+) 0.40
2	(+) 0.00	(+) 0.72
3	(-) 0.50	(+) 0.34

$$w^* = (X^T X)^{-1} X^T Y$$

C1	C2	C3
1	(+) 0.08	(+) 0.40
2	(+) 0.00	(+) 0.72
3	(-) 0.50	(+) 0.34

Check List

Check List Geral

- ❑ Evite códigos
  - Não assuma que o leitor conhece determinada linguagem (Python, C/C++, etc)

- ❑ Exceções
  - Demonstrar a simplicidade da estratégia

We show versatility of einops by expressing common numpy (np) operations in Listing 1

```
1 np.transpose(x, [0, 3, 1, 2])          rearrange(x, 'b h w c -> b c h w')
2 np.reshape(x,
3   [x.shape[0]*x.shape[1], x.shape[2]])
4 np.squeeze(x, 0)                       rearrange(x, '() h w c -> h w c')
5 np.expand_dims(x, -1)                  rearrange(x, 'h w c -> h w c ()')
6 np.stack([r, g, b], axis=2)            rearrange([r, g, b], 'c h w -> h w c')
7 np.concatenate([r, g, b], axis=0)      rearrange([r, g, b], 'c h w -> (c h) w')
```

Fonte: Rogozhnikov. Einops: *Clear and Reliable Tensor Manipulations With Einstein-like Notation.* ICLR, 2022

Implementation. GN can be **easily implemented by a few lines of code** in PyTorch [50] and TensorFlow [51] where automatic differentiation is supported. Figure 3 shows the code based on TensorFlow.

```
def GroupNorm(x, gamma, beta, G, eps=1e−5):
    # x: input features with shape [N,C,H,W]
    # gamma, beta: learnable scale and offset, with shape [1,C,1,1]
    # G: number of groups for GN
    N, C, H, W = x.shape
    x = tf.reshape(x, [N, G, C // G, H, W])

    mean, var = tf.nn.moments(x, [2, 3, 4], keep dims=True)
    x = (x − mean) / tf.sqrt(var + eps)

    x = tf.reshape(x, [N, C, H, W])

    return x ∗ gamma + beta
```

Figure 3. Python code of Group Norm based on TensorFlow.

Fonte: Rogozhnikov. Group Normalization. ECCV, 2018

Check List

Check List Geral

- ❑ Evite colocar a citação para o arXiv quando o artigo está publicado em um veículo de publicação (*peer-review*)

- ❑ Referências padronizadas
  - https://dblp.uni-trier.de

Se são a mesma conferência, por que uma tem mais detalhes que a outra? Uma foi virtual e a outra não?

Uma publicação teve editor, editora e volume e a outra não? Por que uma tem abreviação e a outra não?

Não Padronizada

Kornblit et al. Why Do Better Loss Functions Lead to Less Transferable Features? In NeurIPS, 2021. 2021. Virtual-only

Mahabadi et al. Core-sets for Fair and Diverse Data Summarization. In Neural Information Processing Systems (NeurIPS), 2023.

Shangyun et al. Harder or different? a closer look at distribution shift in dataset reproduction. In International Conference on Machine Learning, 2020.

Dong e al. Attention is not all you need: pure attention loses rank doubly exponentially with depth. Editor Marina Meila and Tong Zhang, volume 139. ICML, 2021. Publisher PMLR.

Fan et al. Reducing transformer depth on demand with structured dropout. In International Conference on Learning Representations (ICLR), 2020.

Hollmann et al. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In International Conference on Learning Representations (ICLR).

Atemporal?

Padronizada

Kornblit et al. Why Do Better Loss Functions Lead to Less Transferable Features? In Neural Information Processing Systems (NeurIPS), 2021.

Mahabadi et al. Core-sets for Fair and Diverse Data Summarization. In Neural Information Processing Systems (NeurIPS), 2023.

Dong e al. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In International Conference on Machine Learning (ICML), 2021.

Shangyun et al. Harder or different? a closer look at distribution shift in dataset reproduction. In International Conference on Machine Learning (ICML), 2020.

Fan et al. Reducing transformer depth on demand with structured dropout. In International Conference on Learning Representations (ICLR), 2020.

Hollmann et al. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In International Conference on Learning Representations (ICLR), 2023.

Check List

Check List Geral

- ❑ Referências padronizadas
  - Mantenham somente as tags author, title, year, booktitle/journal (padronize com o nome por extenso seguido da sigla entre parênteses)

```
@inproceedings{Kolossov:2024,
author    = {Germain Kolossov and Andrea Montanari and Pulkit Tandon},
title = {Towards a statistical theory of data selection under weak supervision},
year = {2024},
booktitle = {International Conference on Learning Representations (ICLR)}
}
```

```
@inproceedings{Mirzasoleiman:2020,
author    = {Baharan Mirzasoleiman and Jeff A. Bilmes and Jure Leskovec},
title     = {Coresets for Data-efficient Training of Machine Learning Models},
booktitle   = {International Conference on Machine Learning (ICML)},
year      = {2020},
}
```

```
@inproceedings{Boull:2020,
author    = {Nicolas Boull{\'{e}} and Yuji Nakatsukasa and Alex Townsend},
title     = {Rational neural networks},
booktitle   = {Neural Information Processing Systems (NeurIPS)},
year      = {2020}
}
```

```
@inproceedings{He:2016,
author    = {Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun},
title     = {Deep Residual Learning for Image Recognition},
booktitle   = {Conference on Computer Vision and Pattern Recognition (CVPR)},
year      = {2016},
}
```

```
@article{Lei:2024,
author    = {Shiye Lei and Dacheng Tao},
title     = {A Comprehensive Survey of Dataset Distillation},
journal   = {IEEE Transactions on Pattern Analysis and Machine Intelligence},
year      = {2024},
}
```

Check List

Check List Geral

- ❑ Leia muito
  - Recomendação: 1 paper por dia (independentemente, se é IC, mestrado ou doutorado)
  - Evite leitura diagonal (isto é, leia observando minuciosamente os detalhes)
  - Como saber quais *insights* o artigo (ex. experimentos) corrobora ou refuta sem conhecer o que existe na literatura?
  - Como saber se as evidências encontradas são apoiadas pela literatura?

Check List

Check List Geral

- ❑ Recomendações de leitura (em ordem de preferência)
  - ICLR/NeruIPS
  - ICML
  - CVPR/PAMI
  - ICCV

- ❑ Sempre que disponível, verifique o OpenReview do paper (ICLR e NeurIPS)
  - Observe quais pontos fortes e fracos os revisores questionaram no artigo (utilize isso como motivação para as contribuições do seu trabalho)
  - Observe como os autores argumentaram com os revisores
  - Algumas discussões envolvendo detalhes técnicos e artigos correlatos ficam evidentes apenas no OpenReview do paper

> My main worry/question, since Geirhos paper has had a lot of attention (as well as mixup/cutmix), is why nobody has done this yet? The paper is very clear and there don't seem to be any hidden difficulty (except maybe using an auxiliary BN?)

Fonte: Li et al. *Shape-Texture Debiased Neural Network Training* (Openreview). ICLR, 2021

Check List

Check List Geral

- ❑ Antes de enviar o artigo para publicação (e.g., arXiv ou *camera-ready*) garanta que todos *acknowledgments* foram inseridos
  - Isso é **mandatório** quando os artigos fazem parte de projetos associados a órgãos de fomento (**FAPESP, CNPq**, etc.)

- ❑ Alguns veículos de publicação não autorizam a disponibilização gratuita do artigo
  - Atenção especial aos termos do *copyright*
  - Em caso de dúvidas não coloque o artigo no arXiv/*homepage*

Check List

Check List Geral

- ❑ Garanta que todos os *checklists* anteriores foram conferidos
  - Revisar um *paper* ruim pode melhorá-lo ligeiramente
  - O processo de revisão interna (entre os autores) é um processo custoso; portanto, **disponibilize a melhor versão possível do *paper***

- ❑ Um *paper* bem polido (com os *checklists* conferidos) torna o trabalho do revisor desafiador
  - O revisor precisa encontrar problemas graves que justifiquem a rejeição do artigo
  - Artigos com erros gramaticais, figuras ruins e difíceis de ler (fluxo ruim) são fáceis de rejeitar

- ❑ Antes de disponibilizar o artigo para revisão interna entre os autores, descanse do texto e revise novamente
  - Nesta etapa verifique mais uma vez se todos os *checklists* foram atendidos

Por que seguir esse Check-List?

Check List Geral

- Exemplos de comentários negativos
  - Fonte: Openreview ICLR 2024

Weaknesses:

The paper contains further inaccurate claims like:

[...]

which were not justified and references were not provided.

Questions:

1. Could you explain how one should read Figure 1? What is the x-axis? What is the y-axis?

https://openreview.net/forum?id=lnffMykYSj

Weaknesses:

Even though the reviewer appreciates many experiments for various tasks, this paper lacks descriptions and justifications of the baselines and how tuned the baseline methods are.

https://openreview.net/forum?id=QHVTxso1Is

Questions:

[...]

- Figure 1 is not referenced in the text

https://openreview.net/forum?id=B4nhr6OJWI

Weaknesses:

1. The writing needs improving, and at times, I find it hard to follow the text. For instance, the term "grokking tickets" is used 12 times before its formal definition presented, which presents a certain impediment to readers.

Questions:

1. The paper could be better organized. […] Excessive space is devoted to some trivial aspects, such as formulas for weight initialization.?

https://openreview.net/forum?id=WSsP7W8tqN

Weaknesses:

2. The writing is unclear in the sense that some notation seems not to be defined, such as p_θ, μ_z. This makes me sometimes a little bit confusing.

https://openreview.net/forum?id=LjygLD0AkT

Questions:

[...]

- Figures are of low resolution

https://openreview.net/forum?id=JWwvC7As4S

Weaknesses:

The paper is not very well written. E.g. the authors write “was wildly interpreted” – they probably mean widely. […] I would encourage them to formulate their thoughts as mathematical theorems with proofs. The algorithms are also unclear.

https://openreview.net/forum?id=0GZ1Bq4Tfr

Onde me Encontrar?

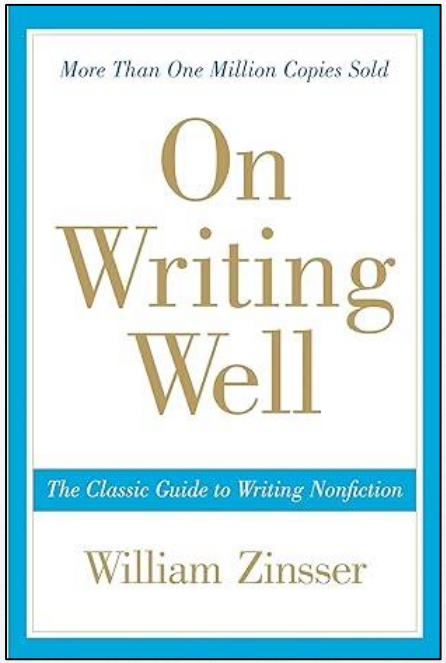
Check List Geral

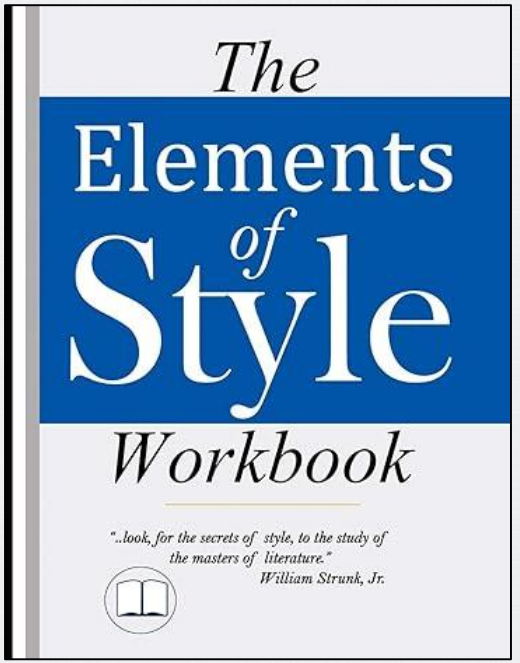
- Os slides desta apresentação estão disponíveis em:
  - https://github.com/arturjordao/PaperWriting

Bibliografia

Bibliografia

- Recomendações de leitura sobre escrita





Bibliografia

- Artigos recomendados
  - *Deep Residual Learning for Image Recognition*. CVPR, 2016
  - *Can Neural Network Memorization Be Localized?*. ICML, 2023
  - *Understanding Plasticity in Neural Networks*. ICML, 2023
  - *Towards a Statistical Theory of Data Selection Under Weak Supervision*. ICLR, 2024 (leitura até a seção 2)