

Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)

Revisiting Neural Networks for Continual Learning: An Architectural Perspective

Aojun Lu¹, Tao Feng², Hangjie Yuan³, Xiaotian Song¹ and Yanan Sun^{1*}

¹Sichuan University

²Tsinghua University

³Zhejiang University

aojunlu@stu.scu.edu.cn, fengtao.hi@gmail.com, hj.yuan@zju.edu.cn
songxt@stu.scu.edu.cn, ysun@scu.edu.cn

<https://arxiv.org/pdf/2404.14829>

SIN5006 - INTELIGÊNCIA COMPUTACIONAL
ENTREGA FINAL

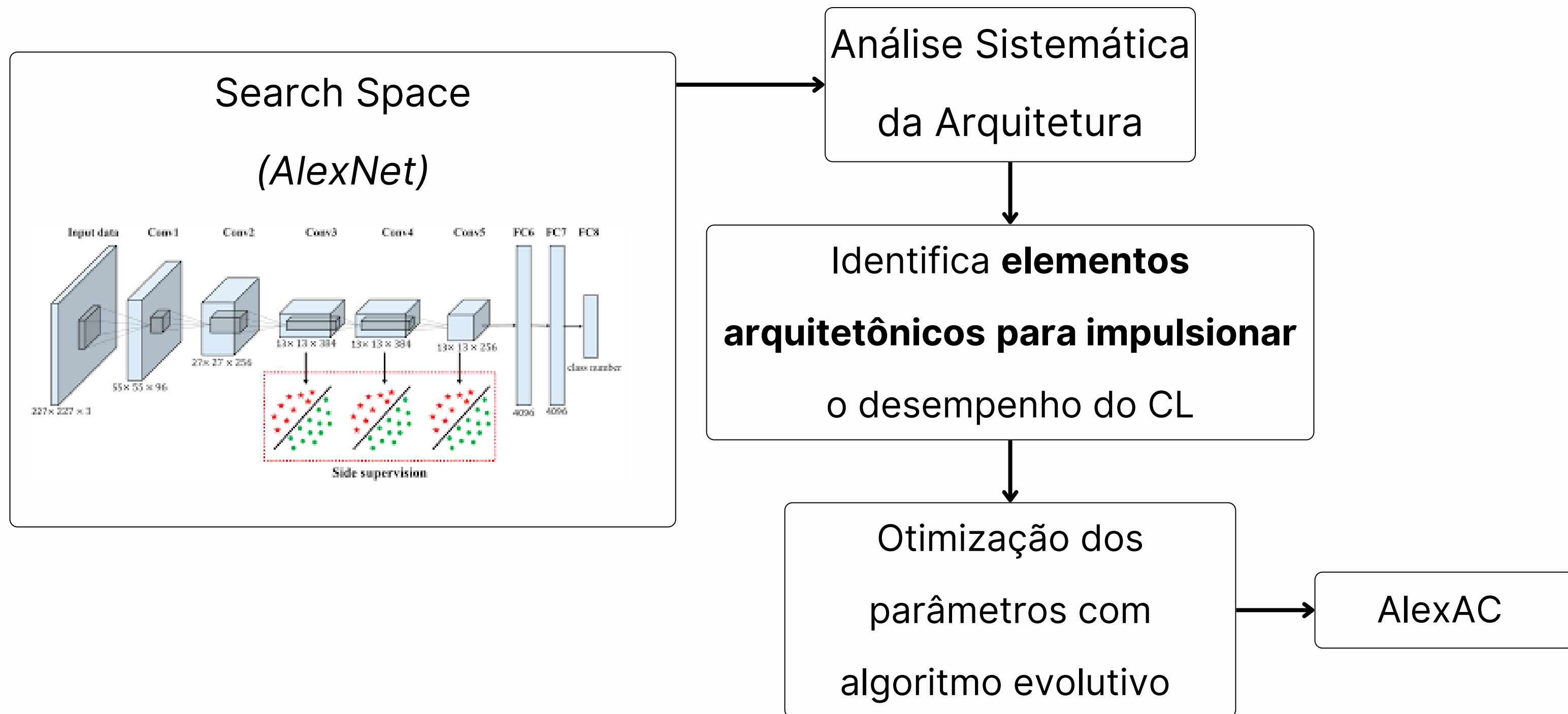
NILTON TADASHI ENTA - 12730911
VITOR CAMARGO PINHEIRO - 12693156

ArchCraft | Sumário Executivo

ArchCraft é um novo método de Neural Architecture Search (NAS).
Focado em **Aprendizado Contínuo (CL)** para preencher a lacuna entre o design da arquitetura de rede e o CL.

1. Revisitando a Metodologia
2. Base de Dados e Métodos Aplicados
3. Tarefa 1: Influência da largura e profundidade no contexto de CL
4. Tarefa 2: Impacto do down-sampling e onde isso ocorre

ArchCraft | Revisitando a Metodologia



ArchCraft | Base de Dados e Métodos Aplicados

AlexNet:

- ↓ Conv1 (64 filtros, kernel=4) + ReLU + Dropout(0.2) + MaxPool
- ↓ Conv2 (128 filtros, kernel=3) + ReLU + Dropout(0.2) + MaxPool
- ↓ Conv3 (256 filtros, kernel=2) + ReLU + Dropout(0.5) + MaxPool
- ↓ Flatten
- ↓ FullyConnected 1 (2048) + ReLU + Dropout(0.5)
- ↓ FullyConnected 2 (2048) + ReLU + Dropout(0.5)

ArchCraft | Base de dados

 Images

CIFAR-100

 Edit

Introduced by Krizhevsky et al. in [Learning multiple layers of features from tiny images](#)

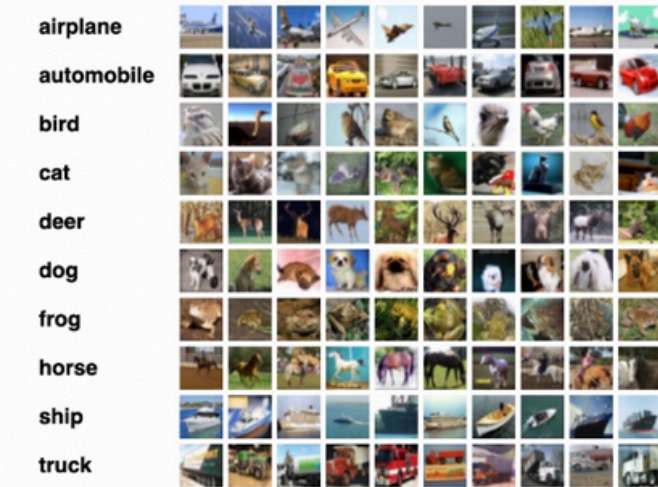
The **CIFAR-100** dataset (Canadian Institute for Advanced Research, 100 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. There are 600 images per class. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs). There are 500 training images and 100 testing images per class.

The criteria for deciding whether an image belongs to a class were as follows:

- The class name should be high on the list of likely answers to the question “What is in this picture?”
- The image should be photo-realistic. Labelers were instructed to reject line drawings.
- The image should contain only one prominent instance of the object to which the class refers.
- The object may be partially occluded or seen from an unusual viewpoint as long as its identity is still clear to the labeler.

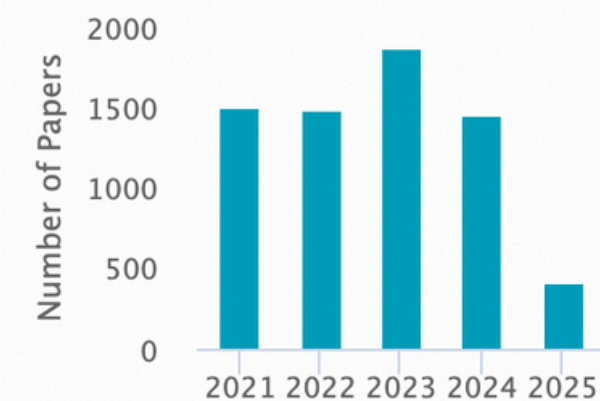
Source:  <https://www.cs.toronto.edu/~kriz/cifar.html>

Homepage



Source: [https://www.cs.toronto.edu/~kriz/...](https://www.cs.toronto.edu/~kriz/)

Usage



ArchCraft | Base de dados

The CIFAR-100 dataset

This dataset is just like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs). Here is the list of classes in the CIFAR-100:

Superclass	Classes
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food containers	bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

Yes, I know mushrooms aren't really fruit or vegetables, and bears aren't really carnivores.

<https://www.cs.toronto.edu/~kriz/cifar.html>

ArchCraft | Base de Dados e Métodos Aplicados

Base de Dados:

- CIFAR100 (*100 classes*)

Representação do CL:

- Incremento em lotes de mesmo tamanho
- As novas classes são treinadas com $\frac{1}{3}$ das épocas do treino inicial

Search Space:

- AlexNet

Algoritmo Evolutivo:

- Mutações estocásticas nos parâmetros selecionados da arquitetura
- Mantém a complexidade em até 1 milhão de parâmetros ou até 100 iterações

Medidas de Desempenho:

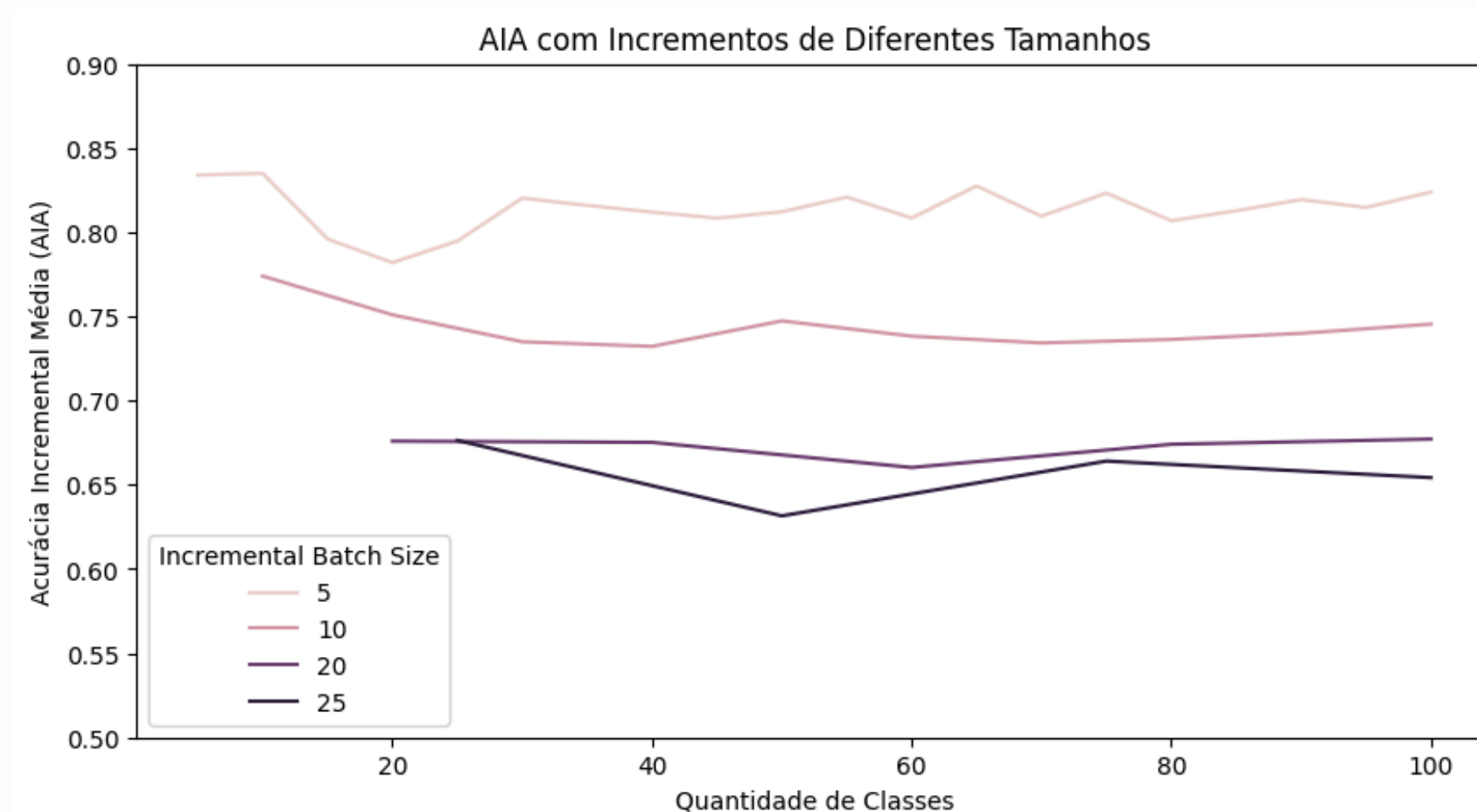
- Acurácia Incremental Média (AIA)
- Catastrophic Forgetting
- Estabilidade e Plasticidade

TAREFA 1: Qual a influência da largura e profundidade no contexto de CL e como ele afeta o Catastrophic Forgetting?

A ← Artigo

- A[1] Como diferentes componentes arquitetônicos (como largura, profundidade, normalização em lote e camadas de *pooling*), afetam a capacidade de aprendizado e retenção de modelos.
- A[2] Abordagem bayesiana para adaptar dinamicamente a profundidade e largura da rede ao longo de tarefas sequenciais, visando melhor desempenho e eficiência.
- A[3] Como o tamanho do modelo (profundidade e largura) impacta o aprendizado contínuo online.

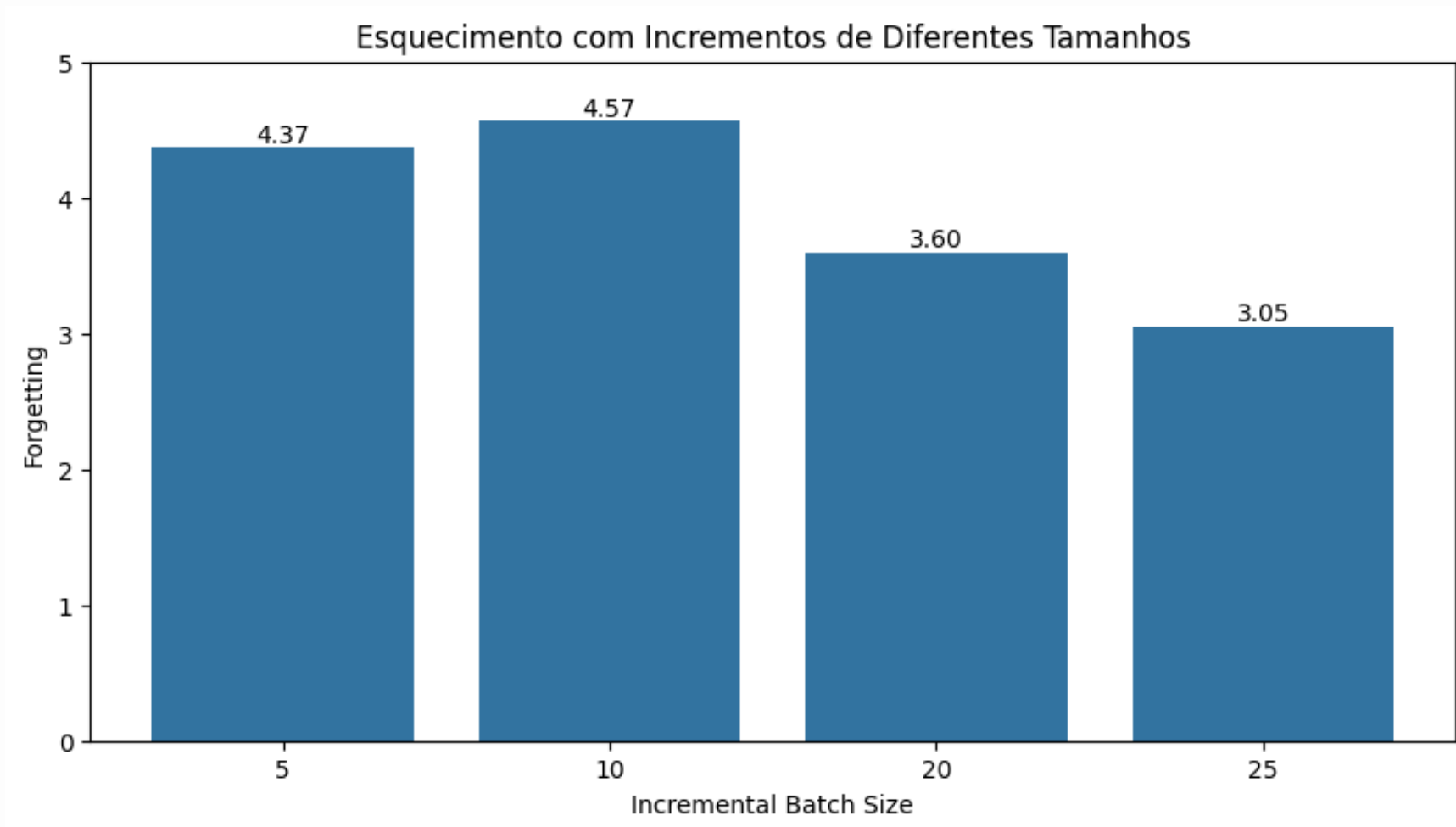
ArchCraft | Resultados



Estabilidade vs. Plasticidade:

- Pior desempenho de modelos maiores, possuem mais parâmetros.
 - **PONTO DE SATURAÇÃO:** Pode dificultar a manutenção da estabilidade das características aprendidas à medida que novos dados são introduzidos.
- Os torna mais propensos ao sobreajuste (*overfitting*) e ao esquecimento de informações previamente aprendidas.

ArchCraft | Resultados



Observações Sobre o Aprendizado:

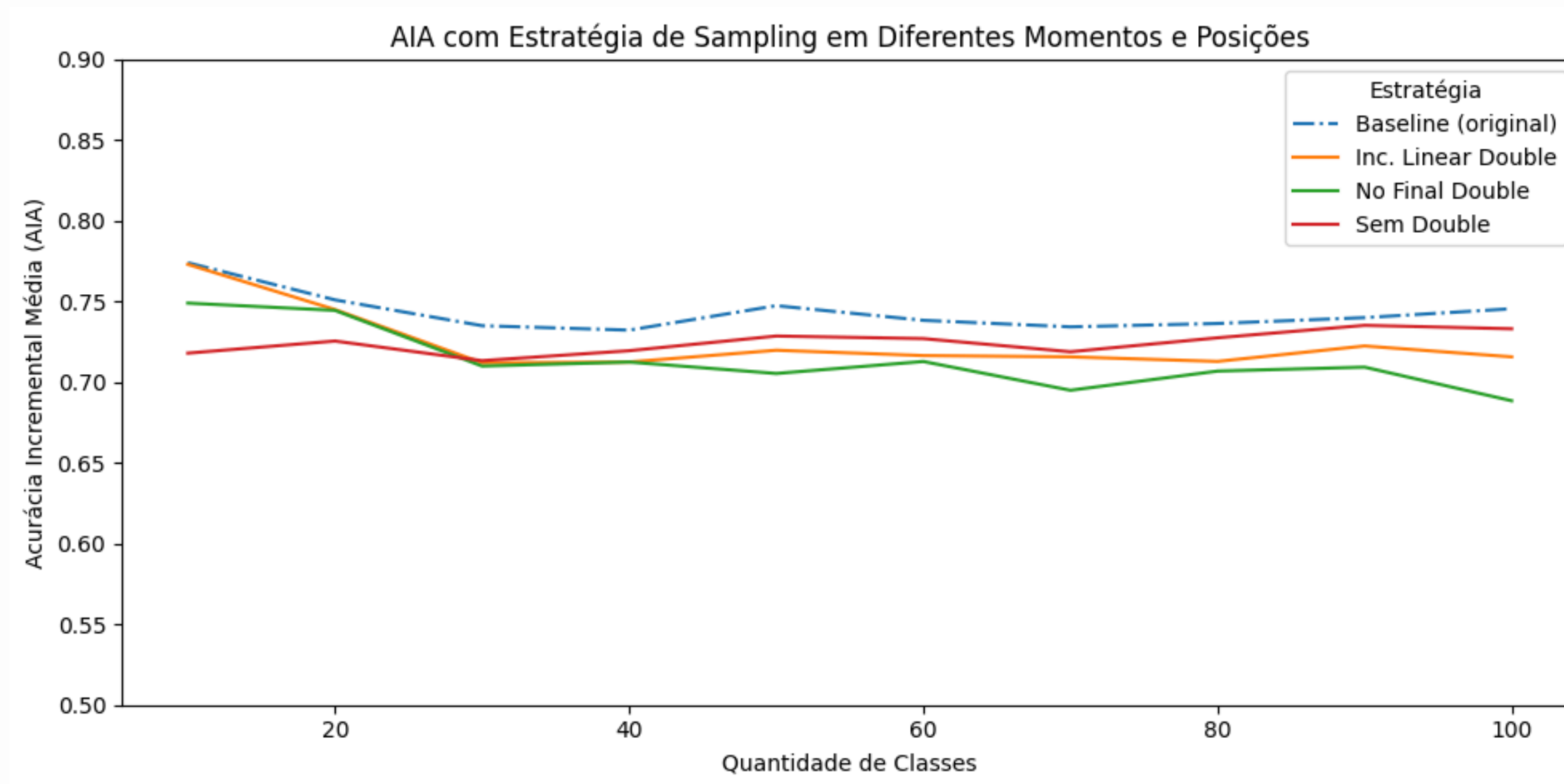
- Impacto da Largura (Over-parametrização)
- Modelos maiores podem ser piores na generalização na tarefa de CL.
 - Versão "slim" do ResNet-18 (menor largura) demonstrou a maior precisão e tendência A[2]
- Em contraste com a largura, aumentar a profundidade (camadas) nem sempre melhora o desempenho de CL

ArchCraft | Resultados

TAREFA 2: Qual o impacto do *down-sampling* e onde ele isso ocorre?

- *Down-sampling*: Refere-se à redução do tamanho do mapa de características, muitas vezes pela metade, em locais escolhidos dentro da arquitetura.
- *Double-sampling*: refere-se a uma estratégia de design arquitetural, onde o número de canais de saída é duplicado em pontos específicos da rede neural.
- **O impacto pode variar dependendo do tipo de pooling utilizado.**
- Capacidade de extração de características: As posições onde o número de canais é aumentado são consideradas críticas para o desempenho do Continual Learning (CL), pois influenciam significativamente a capacidade da rede de extrair características

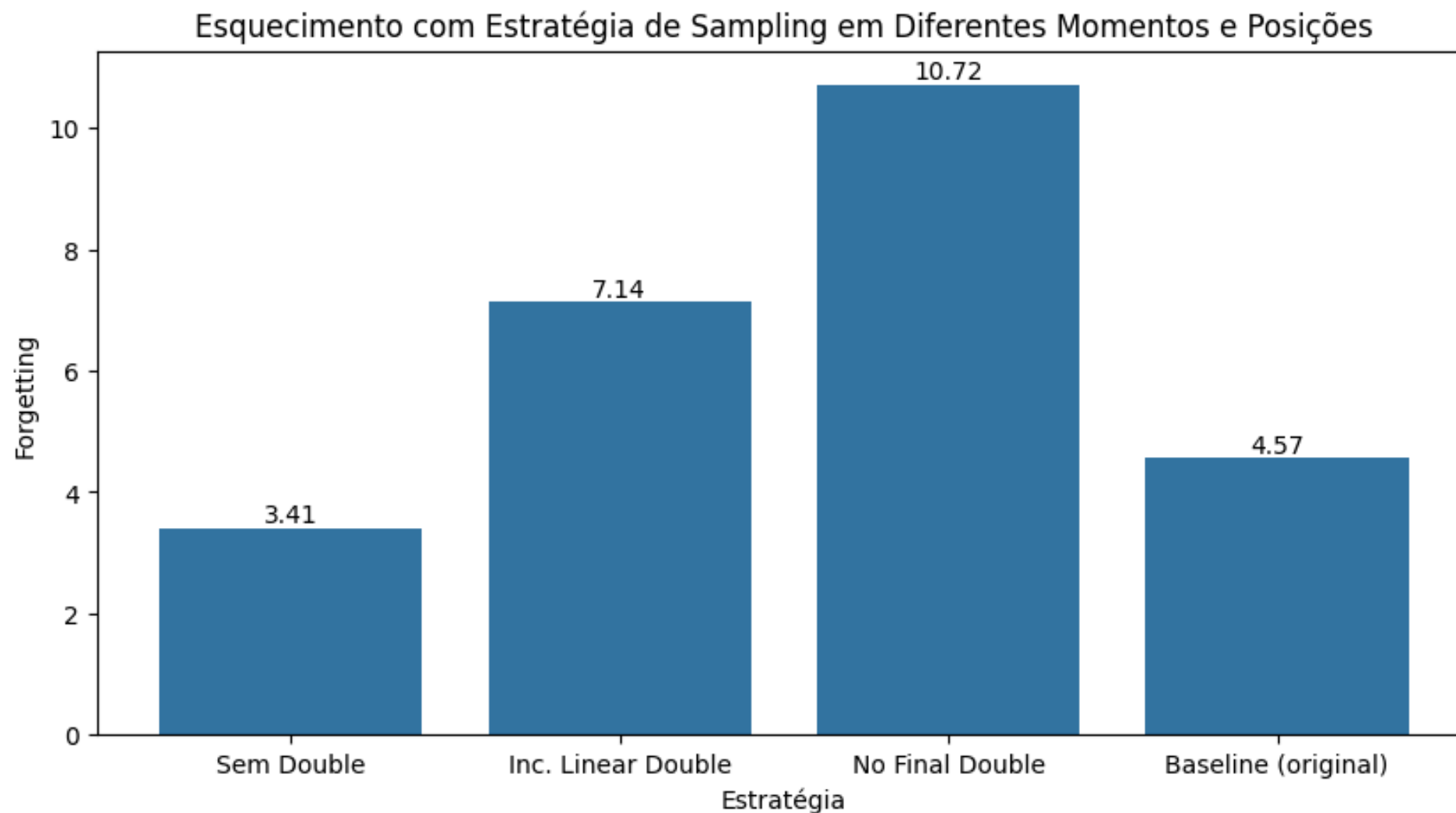
ArchCraft | Resultados



Estabilidade vs. Plasticidade:

- Ponto de Equilíbrio na terceira execução, 30 classes
- Diferentes estratégias com double incrementado apresentaram queda no desempenho

ArchCraft | Resultados



Observações Sobre o Aprendizado:

- A estratégia sem double supera o original
- Perda de desempenho com o aumento dos parâmetros de double

Code:

[0, 0, 0, 1, 2] ← Baseline (original)

[0, 0, 0, 0, 0] ← Sem Double

[0, 1, 2, 4, 8] ← Inc. Linear Double

[0, 0, 2, 2, 2] ← No Final Double

ArchCraft | Referências

LU, Aojun; FENG, Tao; YUAN, Hangjie; SONG, Xiaotian; SUN, Yanan. **Revisiting neural networks for continual learning: an architectural perspective**. arXiv:2404.14829, 2024. Disponível em: <https://arxiv.org/abs/2404.14829>. Acesso em: 30 abr. 2025.

Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Timothy Nguyen, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. **Architecture Matters in Continual Learning**. arXiv:2202.00275v1, 2022

Eunhae Lee. **THE IMPACT OF MODEL SIZE ON CATASTROPHIC FORGETTING IN ONLINE CONTINUAL LEARNING**. arXiv:2407.00176v1, 2024. Massachusetts Institute of Technology

Jeevan Thapa, Rui Li. **Bayesian Adaptation of Network Depth and Width for Continual Learning**. Proceedings of the 41 International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

ANEXOS

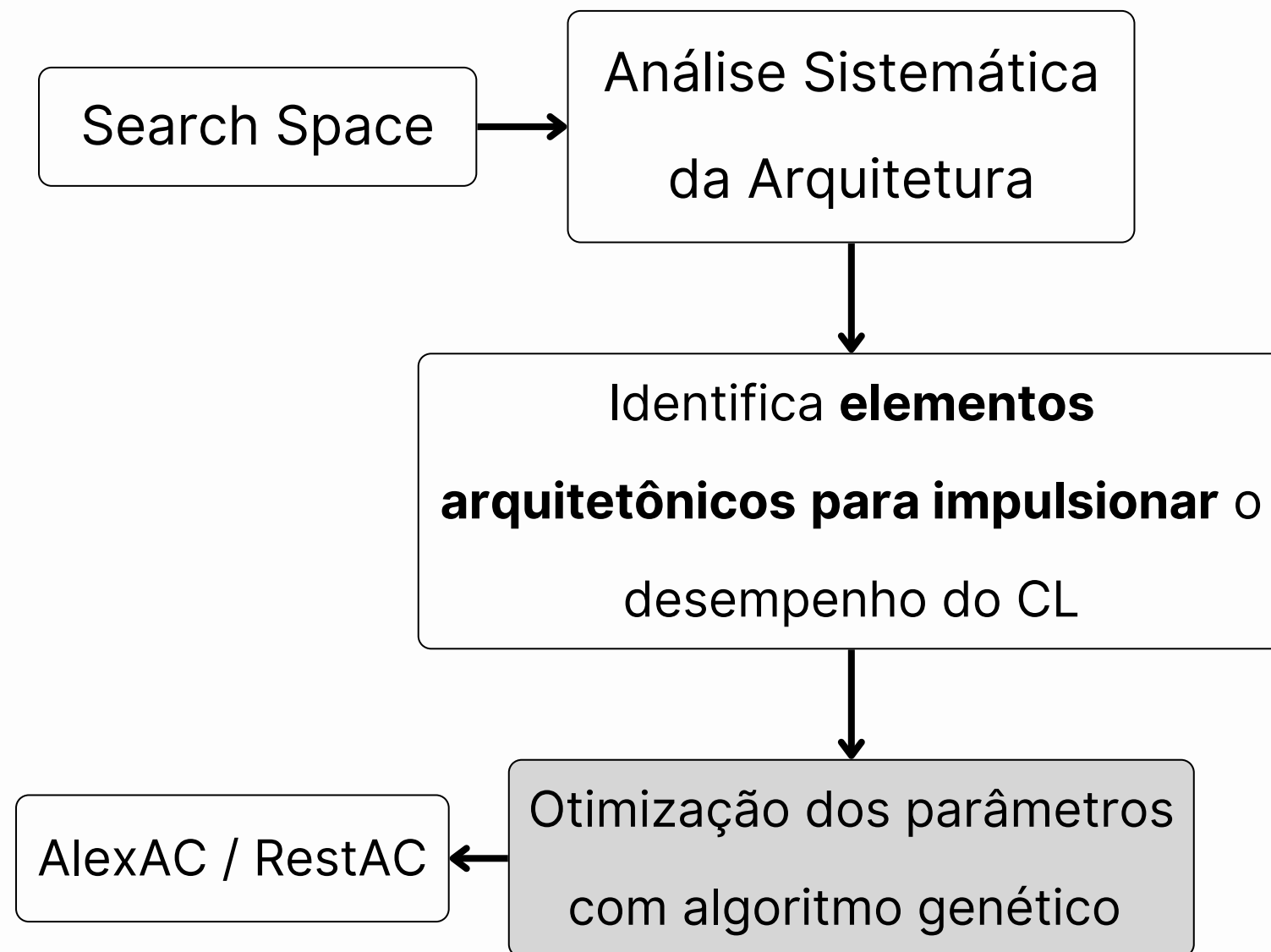
ArchCraft | Medidas de Acurácia para o CL

Parâmetros de Treino:

SGD: Stochastic Gradient Descent

SGD (learning rate: 0.01,
momentum: 0.9,
weight decay: 0.0002,
maximize: False,
differentiable: False)

ArchCraft | Metodologia



Novos **indivíduos (descendentes)** são criados em um algoritmo evolutivo.

Seleção dos pais:

- Dois indivíduos (soluções) são escolhidos aleatoriamente da população atual.
- O que tiver melhor desempenho (fitness) é escolhido para ser o "pai".

Mutação:

- O pai é copiado para criar um descendente.

Em seguida, uma parte do código desse descendente é alterada aleatoriamente (mutação).

ArchCraft | Metodologia

Algoritmo Genético:

old_code \leftarrow [largura, profundidade, pooling[2, 4, 6, 6, 6], double[0, 0, 6, 6, 6]]

WHILE parâmetros < 1 milhão **AND** tentativas < 101:

old_code \leftarrow mutação $\frac{1}{6}$ de probabilidade

largura \leftarrow largura * (ou "/") random == mod(4) e mínimo 8

profundidade \leftarrow profundidade * (ou "/") random

pooling, double \leftarrow ajuste proporcional

Executa 20 iterações de 5 mutações consecutivas

ArchCraft | Métodos (Visão Geral)

ArchCraft é um novo método de Neural Architecture Search (NAS).

Focado em Aprendizado Contínuo (CL) para preencher a lacuna entre o **design da arquitetura de rede e o CL**.

I. Search Space: Espaço de busca personalizado obtido a partir de **busca empírica** [1].

II. Friendly Search Strategy: Uma estratégia baseada em algoritmos genéticos para explorar o Search Space [2].

[1] O ArchCraft ele é projetado com base em **observações empíricas sobre seus componentes arquitetônicos**.

O experimento consistiu com uma ResNet. Partindo de configurações consideradas ótimas para essas tarefas (Incremental Learning (IL)): Max Pooling, Skip connection e Global Avg. Pooling

[2] **Inicialização:** Uma população de arquiteturas é gerada aleatoriamente

Avaliação da performance: durante as iterações evolutivas os indivíduos são avaliados por Acurácia Incremental Média (AIA).

Geração de novos indivíduos: novas arquiteturas são geradas a partir de dois indivíduos, escolhidos aleatoriamente, chamados de filho. O filho é resultado da mutação dessas duas arquiteturas.

Nova população: É gerada a partir dos filhos com melhor aptidão.

PSEUDOCODIGO

DEFINE VARIÁVEIS LOCAIS

incremento <- 50

indivíduo_id <- 0

code <- [profundidade, largura, pool[1, ..., 5], double[1, ..., 5]]

chosen_network <- 'arch_craft'

Algoritmo TrainModel:

grad_clip <- 10

epoch <- 2

learning_rate <- 0.01

Carrega os dados CIFAR-100 divididos em tarefas incrementais

Se rede == 'arch_craft': importa e instancia arquitetura Net personalizada

Se rede == 'alexnet': importa e instancia versão AlexNet

Mostra o número total de parâmetros do modelo

Inicializa o treinamento com otimizador SGD (Gradiente Descendente Estocástico)

Imprime a função de perda e o otimizador

Inicializa matrizes para armazenar acurácias, perdas, e medidas por tarefa

for tarefa **in** classes:

Treina o modelo

Salva os resultados

for conjunto **in** range(tarefas já processadas):

Testa a conjunto

Salva os resultados

Calcula a acurácia média após a tarefa

Calcula forgetting <- $\text{sum}(\text{acurácia máxima} - t[x]) / \text{len}(t)$

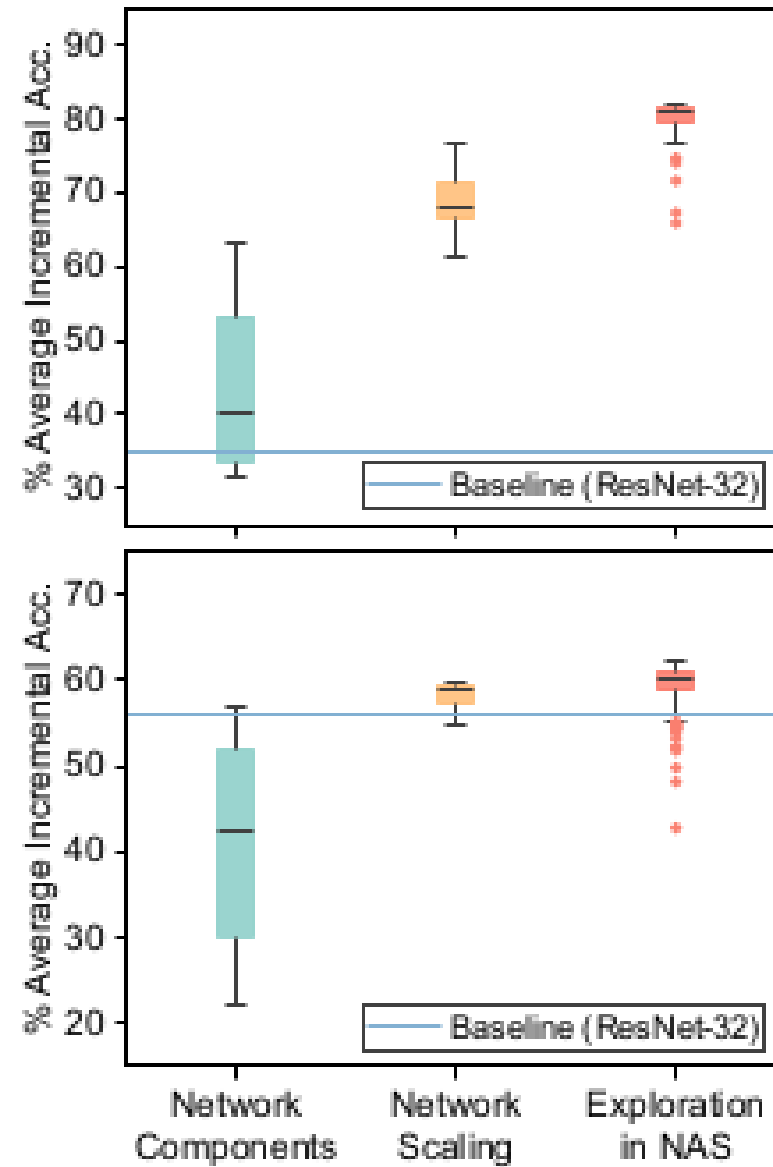
Imprime matriz de acurácia

Calcula e imprime métrica média ap (AIA)

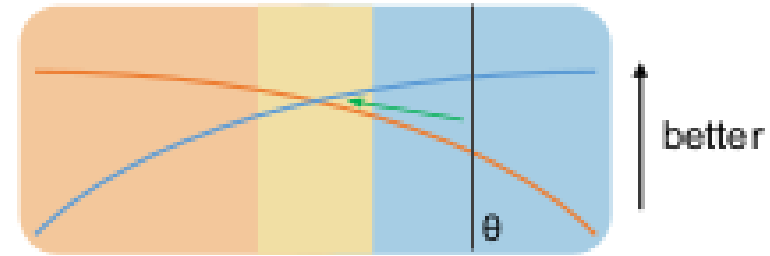
Calcula e imprime acurácia

Retorna a acurácia final

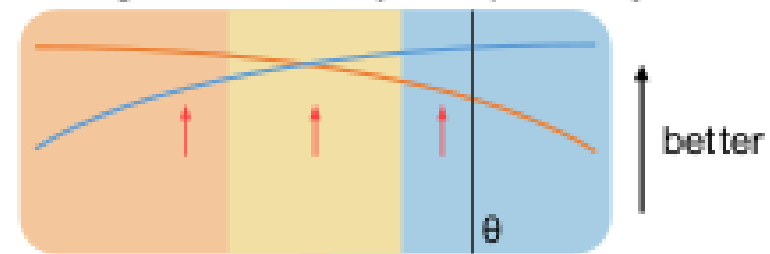
Anexos | Resultados



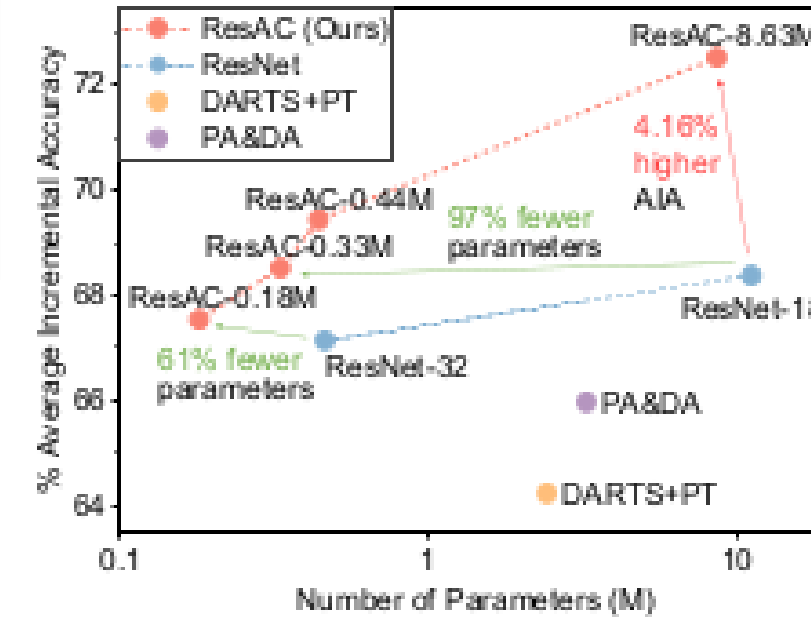
Existing solutions:
Designing novel learning algorithms
to balance the stability and plasticity



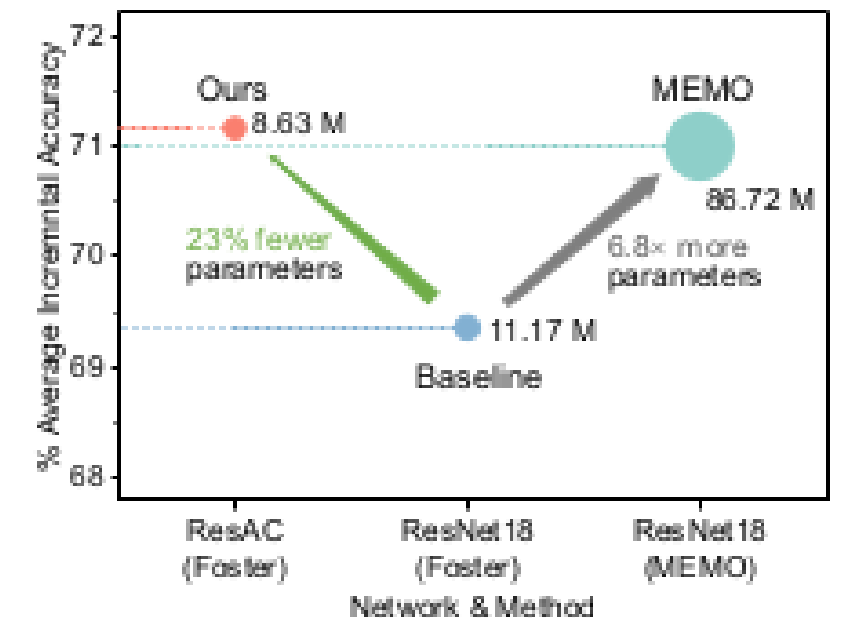
Our solution:
Designing novel basic architectures
with greater stability and plasticity



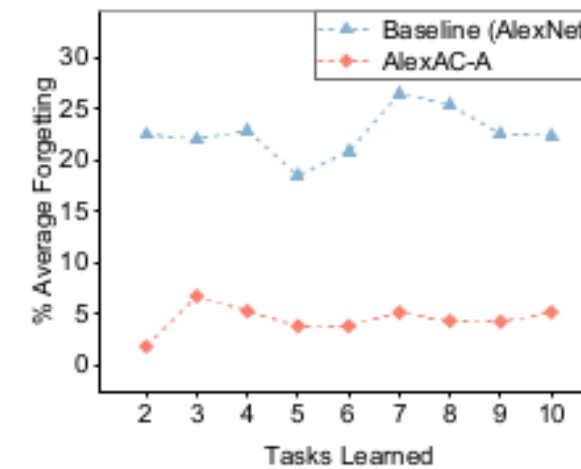
- Stability
- Excessive stability
- Well-balanced stability and plasticity
- Plasticity
- Excessive plasticity



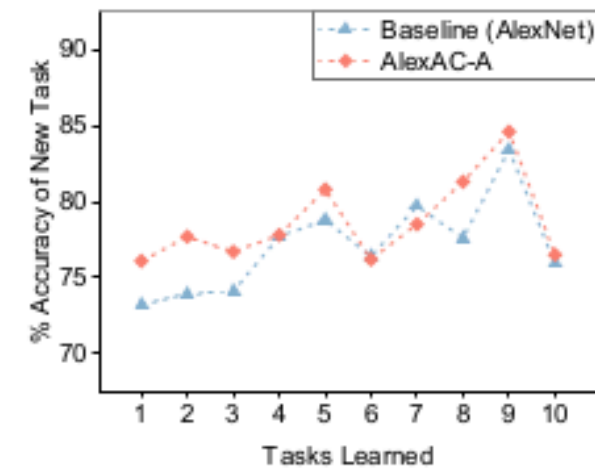
(a)



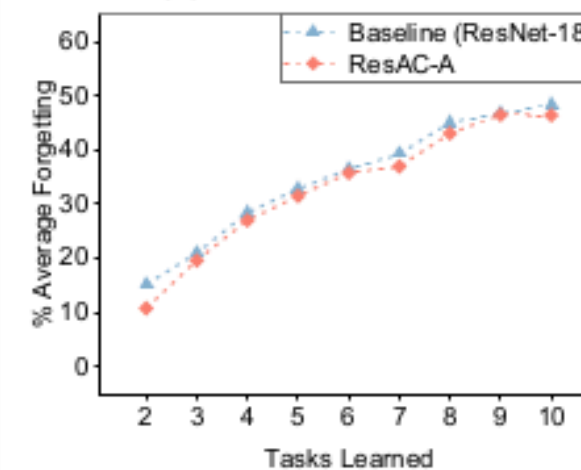
(b)



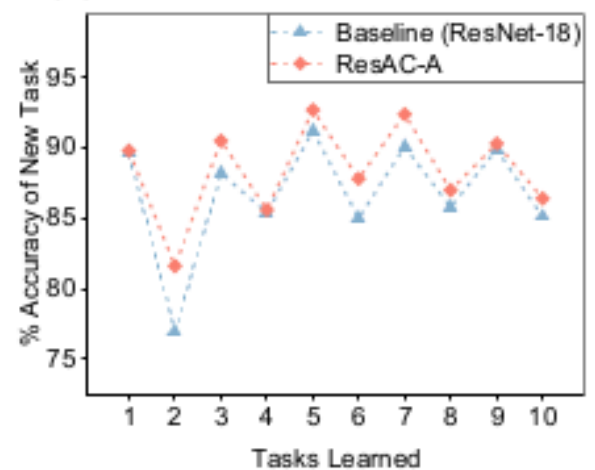
(a) AF in Task IL



(b) New Task Acc. in Task IL



(c) AF in Class IL



(d) New Task Acc. in Class IL

Anexos | Resultados

Method	Network	#P (M)	C100-inc5		C100-inc10		I100-inc5		I100-inc10		Max Improvement	
			LA	AIA	LA	AIA	LA	AIA	LA	AIA	LA	AIA
Replay	ResNet-32	0.46	39.10	58.17	40.02	58.21	-	-	-	-	-	-
	ResAC-B	0.44(↓ 4%)	40.45	59.67	42.79	59.99	-	-	-	-	+2.77	+1.78
	ResNet-18	11.17	40.04	58.80	43.23	60.42	36.30	57.30	41.00	59.21	-	-
	ResAC-A	8.63↓ 23%	42.99	62.52	46.62	63.36	36.78	57.40	42.44	60.07	+3.39	+3.72
iCaRL	ResNet-32	0.46	46.67	63.47	48.80	64.18	-	-	-	-	-	-
	ResAC-B	0.44↓ 4%	47.94	64.17	50.11	64.42	-	-	-	-	+1.31	+0.70
	ResNet-18	11.17	47.32	64.13	52.77	66.04	44.10	62.36	50.98	67.11	-	-
	ResAC-A	8.63↓ 23%	52.6	68.71	55.52	69.62	45.12	63.98	52.46	68.42	+5.28	+4.58
WA	ResNet-32	0.46	46.95	62.93	53.35	66.61	-	-	-	-	-	-
	ResAC-B	0.44↓ 4%	51.31	66.39	54.89	67.73	-	-	-	-	+4.36	+3.46
	ResNet-18	11.17	45.11	62.06	56.59	68.89	46.06	62.96	55.04	68.60	-	-
	ResAC-A	8.63↓ 23%	53.23	69.19	59.79	71.40	49.94	67.20	58.86	71.56	+8.12	+7.13
Foster	ResNet-32	0.46	47.78	62.36	54.36	67.14	-	-	-	-	-	-
	ResAC-B	0.44↓ 4%	53.50	67.34	58.17	69.44	-	-	-	-	+5.72	+4.98
	ResNet-18	11.17	49.03	61.97	55.98	68.38	53.26	65.20	60.58	69.36	-	-
	ResAC-A	8.63↓ 23%	57.22	69.99	61.44	72.54	54.32	66.41	61.94	71.16	+8.19	+8.02

Table 5: The CL performance of ArchCraft in *Class IL*. ‘#P’ represents the number of parameters of the network used.