Understanding Catastrophic Forgetting: A Three-Axis Structural Analysis in Class-Incremental Learning

Gabriel F. Sarmento[1], Liedson D. O. Leite[1], Thyago T. M. de Oliveira[1], Rodolfo C. Cavalcante[1], Leandro L. Minku[2], Gustavo H. F. de M. Oliveira[1]

[1]Sistemas de Informação – Universidade Federal de Alagoas (UFAL)
Unidade Educacional de Penedo, Penedo, AL – Brasil

[2]School of Computer Science – University of Birmingham
Edgbaston, Birmingham, B15 2TT – Reino Unido

{gabriel.sarmento, liedson.leite, rodolfo.cavalcante}@arapiraca.ufal.br

{gustavo.oliveira, thyago.oliveira}@penedo.ufal.br, l.l.minku@bham.ac.uk

Abstract. *Continual learning aims to integrate new knowledge without forgetting the past. However, in Class-Incremental Learning, models are evaluated by aggregate metrics that obscure the structure of catastrophic forgetting. Therefore, an analysis protocol for the temporal, decisional, and representational axes is proposed to investigate when, how, and why it occurs. Five strategies (Naive, EWC, LwF, ER, DER++) were evaluated in four benchmarks. The results show that models with similar overall performance may present different failure mechanisms, caused by the non-uniformity of the manifestation of forgetting. It can be concluded that this multidimensional evaluation is indispensable for accurate diagnoses.*

Resumo. *O aprendizado contínuo visa integrar novos conhecimentos sem esquecer o passado. Contudo, na Aprendizagem Incremental de Classe, os modelos são avaliados por métricas agregadas que ocultam a estrutura do esquecimento catastrófico. Por isso, propõe-se um protocolo de análise para os eixos temporal, decisório e representacional para investigar quando, como e por que ele ocorre. Foram avaliadas cinco estratégias (Naive, EWC, LwF, ER, DER++) em quatro benchmarks. Os resultados mostram que modelos com desempenho global semelhante podem apresentar diferentes mecanismos de falha, causados pela não uniformidade da manifestação do esquecimento. Conclui-se que esta avaliação multidimensional é indispensável para diagnósticos precisos.*

1. Introdução

No contexto de Aprendizado Contínuo (*Continual Learning*), modelos são expostos a dados de forma sequencial, sendo atualizados continuamente à medida que novas tarefas são introduzidas. Nesse contexto, uma *tarefa* corresponde, em geral, a um subconjunto de classes. Um dos principais desafios desse paradigma é a indisponibilidade do conjunto completo de dados anteriores durante o treinamento, restringindo o modelo a acessar apenas pequenas amostras ou nenhuma, o que dificulta a preservação do conhecimento já adquirido. Nesse cenário, redes neurais artificiais tendem a se degradar rapidamente nas tarefas previamente aprendidas, fenômeno conhecido como esquecimento catastrófico [French 1999, McCloskey and Cohen 1989].

No cenário de *Class-Incremental Learning* (CIL), esse desafio se torna ainda mais relevante. Nesse caso, o modelo precisa predizer o dado de entrada na classe correta considerando todas as outras que já aprendeu até o momento, sem receber qualquer indicação de qual grupo ou etapa de treinamento o dado pertence [van de Ven et al. 2022]. Isso exige, ao mesmo tempo, estabilidade para preservar o conhecimento anterior e plasticidade para incorporar as novas classes, dilema conhecido como estabilidade-plasticidade [Mermillod et al. 2013].

Apesar de diversas abordagens terem surgido na literatura para resolver esse desafio, seus processos de avaliação comparativa para esse contexto apoiam-se predominantemente em métricas agregadas, como acurácia média (A_k) e *forgetting* (F_k), sendo k o número de tarefas aprendidas pelo modelo [Lopez-Paz and Ranzato 2017, Chaudhry et al. 2018]. Embora convenientes, essas métricas calculam a média das tarefas individuais para resumir o desempenho final em um único valor global, ocultando variações específicas ao longo do treinamento e entre classes. Como consequência, modelos distintos podem apresentar valores semelhantes, mas exibir comportamentos qualitativamente diferentes ao longo do treinamento. Essa limitação no processo avaliativo dificulta a interpretação do esquecimento catastrófico, pois não permite identificar quando, como e por que ele ocorre. Nesse contexto, torna-se necessária uma análise mais estruturada do fenômeno, que vá além da avaliação puramente numérica das métricas tradicionais.

De forma complementar, existem algumas análises na literatura que podem auxiliar a entender melhor a granularidade do problema, como os mapas de esquecimento (*forgetting maps*), matrizes de confusão [Rebuffi et al. 2017] e projeções t-SNE [Chen et al. 2025]. No entanto, elas são tipicamente empregadas de forma isolada nos trabalhos e não constituem um quadro sistemático para avaliação do esquecimento catastrófico. Diante desse cenário, este trabalho busca compreender esse fenômeno, respondendo à seguinte pergunta de pesquisa: como, quando e por que o esquecimento catastrófico se manifesta ao longo do treinamento? Partimos da hipótese de que o esquecimento em CIL não é apenas quantitativo, mas também estrutural, manifestando-se em três dimensões complementares. No eixo temporal, o esquecimento está relacionado ao momento e à intensidade com que o desempenho se degrada ao longo da sequência de tarefas. No eixo decisório, refere-se à forma como os erros de classificação se reorganizam entre classes antigas e novas. Já no eixo representacional, o fenômeno está associado ao formato das representações internas aprendidas pela rede neural, como sobreposição, colapso ou rigidez entre clusters de classes.

Para investigar essa hipótese, conduzimos um estudo de caso comparando cinco estratégias representativas da literatura: Naive, EWC [Kirkpatrick et al. 2017], LwF [Li and Hoiem 2017], Experience Replay (ER) [Rolnick et al. 2019] e DER++ [Buzzega et al. 2020], em quatro benchmarks: MNIST, FMNIST, CIFAR-10 e CIFAR-100. Logo, as principais contribuições deste artigo são: (i) a proposição de um protocolo unificado de análise estrutural do esquecimento; (ii) evidências empíricas de que métricas globais podem induzir interpretações equivocadas; e (iii) a demonstração de que diferentes estratégias de aprendizado levam a organizações distintas do espaço de representações, revelando direções importantes de pesquisa futura na área.

Este trabalho está organizado em seis seções. A Seção 2 apresentará a

fundamentação teórica de CIL, bem como algumas abordagens, as métricas atuais e suas limitações. A Seção 3 apresentará o protocolo proposto para a análise nos três eixos, enquanto a Seção 4 descreve o protocolo experimental do estudo de caso. A Seção 5 apresenta e discute os resultados e, por fim, a Seção 6 encerra com conclusões e direções futuras.

2. Class-Incremental Learning e Limitações da Avaliação Tradicional

Esta seção apresenta os fundamentos de Class-Incremental Learning (CIL), abordando as principais estratégias de mitigação do esquecimento catastrófico na Subseção 2.1 e as limitações das métricas tradicionais de avaliação na Subseção 2.2.

2.1. Class-Incremental Learning (CIL) e suas Principais Abordagens

No Aprendizado Contínuo, o modelo é treinado sobre uma sequência de tarefas $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$, com acesso restrito ou nulo ao conjunto completo de dados anteriores, buscando aprender novos conhecimentos sem degradar os já adquiridos. Cada tarefa $\mathcal{T}_k$ é associada a um conjunto de dados $\mathcal{D}_k = \{(x_i, y_i)\}$ e, no cenário *class-incremental* (CIL), isso introduz novas classes, de modo que o conjunto de rótulos cresce ao longo do tempo: $\mathcal{Y}_1 \subset \mathcal{Y}_2 \subset \cdots \subset \mathcal{Y}_K$ [van de Ven et al. 2024].

Para mitigar esse desafio, as abordagens da literatura podem ser organizadas em três famílias principais: métodos de regularização, métodos baseados em replay e métodos arquiteturais. Os **métodos de regularização** atuam no espaço de pesos da rede neural, adicionando termos à função de perda que penalizam alterações em tarefas anteriores, podendo usar como método a matriz de Fisher [Kirkpatrick et al. 2017]. Essas restrições tendem a se acumular à medida que novas tarefas são incorporadas, limitando a capacidade de adaptação do modelo [van de Ven et al. 2022]. Em contraste, **métodos baseados em replay** reintroduzem conhecimento por meio de amostras de tarefas passadas durante o novo treinamento [Rolnick et al. 2019]. A terceira família, a dos **métodos arquiteturais**, modifica explicitamente a estrutura da rede neural para acomodar novas tarefas, seja por expansão dinâmica de parâmetros, seja por isolamento parcial de representações. Exemplos incluem Progressive Networks e Dynamic Expandable Networks [Yoon et al. 2017]. Apesar de relevantes, os métodos da terceira família não foram considerados neste estudo por frequentemente introduzirem crescimento estrutural do modelo, dificultando comparações diretas com abordagens baseadas em replay e regularização.

2.2. Métricas de Avaliação em CIL e Principais Limitações

No cenário CIL, o desempenho dos modelos é usualmente avaliado por métricas agregadas como a acurácia média e o esquecimento médio [Lopez-Paz and Ranzato 2017]. Para defini-las, considere $a_{k,j} \in [0, 1]$ como a acurácia no conjunto de teste da tarefa j após o treinamento até a tarefa k. A acurácia média após aprender a tarefa k é dada por $A_k = \frac{1}{k}\sum_{j=1}^{k} a_{k,j}$, enquanto o esquecimento médio é definido por $F_k = \frac{1}{k-1}\sum_{j=1}^{k-1}(\max_{l<k} a_{l,j} - a_{k,j})$ [Chaudhry et al. 2018]. Embora amplamente utilizadas, essas métricas condensam o desempenho em valores únicos, ocultando dinâmicas estruturais que podem ser analisadas sob diferentes perspectivas complementares.

Ao tentarmos avaliar os modelos, por exemplo, no eixo **temporal**, as métricas agregadas não indicam quando o esquecimento ocorre ao longo da sequência de tarefas. Isto é, elas não distinguem se a degradação de uma tarefa anterior ocorre de forma abrupta, caindo drasticamente logo após o aprendizado da tarefa imediatamente subsequente, ou de maneira gradual, em que a interferência se acumula lentamente à medida que múltiplas novas tarefas são adicionadas. No eixo de **decisão**, a agregação por tarefa mascara o comportamento de classes individuais. Por exemplo, uma tarefa pode apresentar 80% de acurácia mesmo que algumas classes antigas estejam sendo sistematicamente confundidas com novas, enquanto as demais permanecem corretas. Assim, o erro fica concentrado em classes específicas, mas não é evidenciado por A_k e F_k. Por fim, no eixo **representacional**, as métricas não informam como o espaço de representações evolui ao longo do treinamento. Dois modelos com desempenho semelhante podem apresentar estruturas internas distintas, como separação preservada ou sobreposição entre classes antigas [Ramasesh et al. 2020], indicando mecanismos de esquecimento diferentes que não são capturados por A_k e F_k.

Essas limitações mostram que a avaliação padrão em CIL não caracteriza adequadamente o fenômeno do esquecimento, motivando a necessidade de protocolos que integrem explicitamente as dimensões temporal, decisória e representacional.

3. Protocolo para Análise Estrutural do Esquecimento

Esta seção apresenta o protocolo de análise estrutural do esquecimento, organizado em três eixos complementares: temporal, decisório e representacional, conforme ilustrado na Figura 1 e discutido nas subseções a seguir, respectivamente.

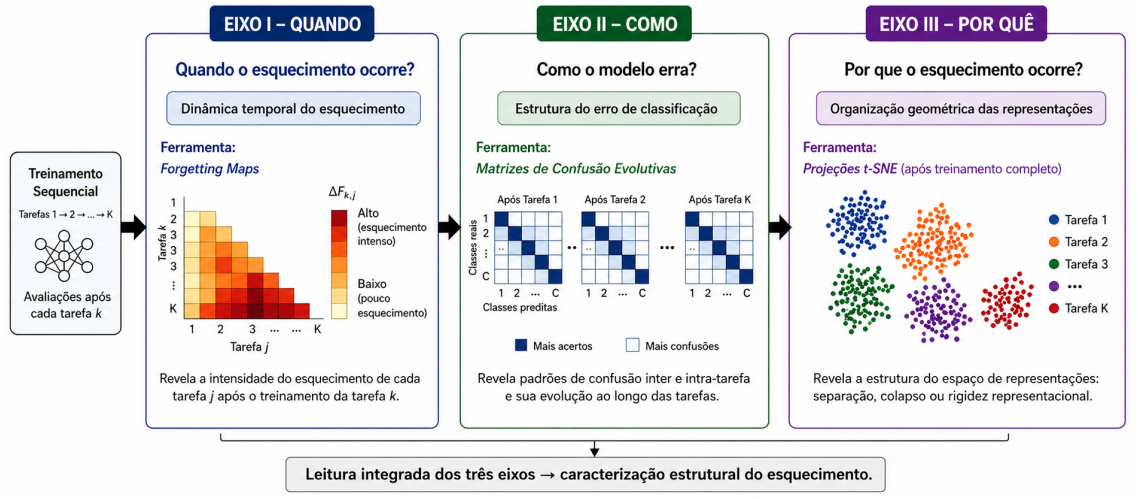


Figura 1. Fluxo do protocolo para análise estrutural do esquecimento.

3.1. Eixo I: Dinâmica Temporal do Esquecimento

O primeiro eixo investiga ***QUANDO*** e com que intensidade o esquecimento ocorre ao longo da sequência de tarefas. Para capturar essa dinâmica, utiliza-se a sistematização dos ***forgetting maps***, uma representação visual bidimensional derivada da decomposição da métrica de *forgetting* (F_k) proposta por [Chaudhry et al. 2018]. Cada célula (k, j) registra a queda de acurácia da tarefa j imediatamente após o treinamento da tarefa k,

definida como $\Delta F_{k,j} = \max_{l \leq k-1} a_{l,j} - a_{k,j}$, para $j < k$, onde $a_{l,j}$ representa a acurácia da tarefa j após o treinamento da tarefa l.

A visualização como um mapa de calor triangular permite descompactar a métrica global e revelar a topografia do esquecimento ao longo do tempo, possibilitando identificar tanto as tarefas mais vulneráveis (análise por coluna) quanto aquelas que atuam como catalisadoras de maior interferência nas representações anteriores (análise por linha).

3.2. Eixo II: Estrutura do Erro de Classificação

O segundo eixo investiga ***COMO*** o modelo erra após o treinamento sequencial. Para capturar essa estrutura, propõe-se o uso de **matrizes de confusão evolutivas** [Rebuffi et al. 2017], construídas ao final do treinamento de cada tarefa, para permitir acompanhar como os padrões de erro se transformam à medida que novas tarefas são introduzidas. Em cada matriz, as linhas representam as classes reais e as colunas as classes preditas; a diagonal principal indica acertos, enquanto as entradas fora da diagonal evidenciam os principais padrões de confusão. A análise evolutiva permite identificar quando determinados vieses surgem e como se intensificam ou se dissipam com o treinamento.

3.3. Eixo III: Organização do Espaço de Representações

O terceiro eixo investiga ***POR QUE*** o esquecimento ocorre. Para tornar essa estrutura visível, propõe-se utilizar as **projeções t-SNE** [van der Maaten and Hinton 2008] das representações extraídas da penúltima camada da rede ao final do treinamento. O t-SNE é um método de redução de dimensionalidade não-linear que preserva relações de vizinhança, mapeando essas representações para um espaço bidimensional, de modo que pontos próximos no espaço original permaneçam próximos na visualização. Cada ponto representa uma amostra do conjunto de teste, colorida de acordo com a tarefa de origem.

A análise dessas projeções permite identificar padrões estruturais das representações associados ao esquecimento. Quando tarefas distintas formam clusters bem separados, o modelo preserva estrutura discriminativa; quando os pontos se concentram em uma única região, ocorre *colapso representacional*; e quando um único cluster domina (tipicamente associado às primeiras tarefas), observa-se *rigidez representacional*, em que novas tarefas não estabelecem regiões próprias no espaço representacional.

4. Estudo de caso: Validação Experimental

Esta Seção apresenta o estudo de caso e a configuração experimental adotada. A Figura 2 resume os datasets, arquiteturas, estratégias e o cenário de Class-Incremental utilizados, cujos detalhes são discutidos nas subseções a seguir.

4.1. Datasets

A configuração experimental utiliza quatro benchmarks consolidados na literatura: MNIST, FMNIST, CIFAR-10 e CIFAR-100 que podem ser obtidos através da biblioteca Avalanche[1]. Esses datasets diferem principalmente em termos de complexidade visual, número de classes e tipo de representação, permitindo uma análise abrangente do comportamento dos modelos em diferentes níveis de dificuldade.

[1] `https://pypi.org/project/avalanche-lib/`

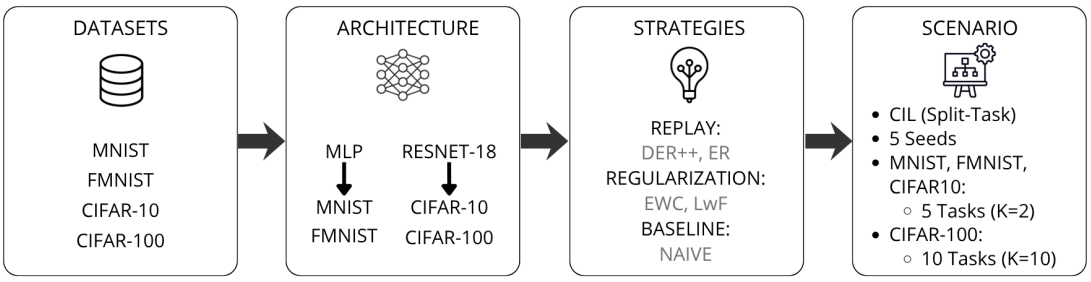


Figura 2. Fluxo da configuração experimental: datasets, arquiteturas, estratégias e cenário class-incremental.

O MNIST e o FMNIST são compostos por 10 classes e imagens em escala de cinza com dimensão 28×28. O MNIST apresenta dígitos manuscritos, enquanto o FMNIST introduz itens de vestuário, como camisetas, sapatos e bolsas, tornando a tarefa de classificação ligeiramente mais desafiadora.

O CIFAR-10 também possui 10 classes, porém utiliza imagens coloridas de 32×32, com maior variabilidade e maior complexidade visual. As classes incluem objetos como aviões, automóveis e animais, exigindo maior capacidade de generalização do modelo.

Por fim, o CIFAR-100 amplia significativamente a dificuldade do problema ao conter 100 classes distintas, também em imagens coloridas de 32×32. Esse aumento no número de categorias impõe maior demanda de discriminação fina entre classes semanticamente próximas.

4.2. Arquiteturas das Redes Neurais Experimentadas

As arquiteturas foram definidas e fixadas de modo a isolar o seu desempenho para enfatizar apenas o impacto das estratégias de aprendizado contínuo. Para os datasets MNIST e FMNIST, definiu-se um *Multilayer Perceptron (MLP)* com duas camadas ocultas de 256 neurônios cada, ativação ReLU e regularização via Dropout de 0.2. Essa escolha é suficiente para capturar os padrões estruturais desses dados sem introduzir complexidade desnecessária.

Para CIFAR-10 e CIFAR-100, definiu-se uma ResNet-18 *slim*, treinada do zero (sem pré-treinamento), de modo a evitar qualquer influência de conhecimento prévio. A arquitetura é adaptada para entradas de 32×32 pixels com ajuste no *stride* inicial, mantendo compatibilidade com os dados.

Em todos os casos, o treinamento ocorre de forma sequencial usando lotes (*batches*) de tamanho 64 e ajuste de pesos usando 20 épocas por tarefa. A otimização é realizada com o algoritmo Stochastic Gradient Descent (SGD), utilizando taxa de aprendizado no conjunto $\{0.001, 0.005, 0.01\}$, *momentum* 0.9 e *weight decay* de 5×10^{-4}.

4.3. Estratégias de Aprendizado Contínuo e Hiperparametrização

Para o estudo de caso, foram avaliadas cinco estratégias representativas de aprendizado contínuo, oriundas das famílias de **regularização** e **replay**. Como baseline, utiliza-se o método **Naive**, no qual o modelo é treinado sequencialmente sem mecanismos de mitigação do esquecimento.

Todas as estratégias foram implementadas com a biblioteca Avalanche para garantir a consistência de implementação. Os hiperparâmetros foram definidos por busca aleatória nas três primeiras tarefas de cada dataset, utilizando 10% dos dados de treino como validação. Os hiperparâmetros finais correspondem aos valores que apresentaram melhor estabilidade entre plasticidade e preservação de conhecimento durante a validação preliminar conduzida neste estudo.

Na família de **regularização**, são considerados os métodos **EWC** e **LwF**. O **EWC** foi avaliado com $\lambda \in \{10, 100, 400, 1000, 5000\}$, sendo selecionado $\lambda = 1000$ (ou $\lambda = 100$ no CIFAR-100), com taxa de aprendizado de 0.01, batch size de 64 e 20 épocas. O **LwF** foi avaliado com $\alpha \in \{0.5, 1, 2, 5\}$ e $T \in \{1, 2, 4\}$, sendo selecionados $\alpha = 1.0$ e $T = 1.0$, com taxa de aprendizado de 0.01, batch size de 32 e 20 épocas.

Na família de **replay**, são avaliados os métodos **ER** e **DER++**. O **ER** foi avaliado com buffers de tamanho $\{200, 500, 1000, 2000\}$, sendo selecionado 2000, com taxa de aprendizado de 0.005, batch size de 32 e 10 épocas. O **DER++** foi avaliado com $\alpha, \beta \in [0.1, 1.0]$ e buffers entre 200 e 2000 amostras, sendo selecionados $\alpha = 0.5$, $\beta = 0.5$ e buffer de 2000, com taxa de aprendizado de 0.001, batch size de 64 e 20 épocas.

4.4. Cenário Experimental de Class-Incremental

O cenário adotado utiliza o protocolo de avaliação *split-task*. Nesse contexto, o conjunto de classes de cada dataset é particionado em tarefas disjuntas, que são apresentadas sequencialmente ao modelo ao longo do treinamento.

Para os datasets com 10 classes (MNIST, FMNIST e CIFAR-10), são definidas 5 tarefas com 2 classes cada (por exemplo, classes 0–1 na primeira tarefa, 2–3 na segunda, e assim sucessivamente). Já no CIFAR-100, o particionamento é feito em 10 tarefas com 10 classes cada. Essa configuração controla explicitamente o número de tarefas e o grau de interferência entre elas.

Dessa forma, o modelo de rede neural é treinado de forma sequencial, tendo acesso, em cada etapa, aos dados da tarefa corrente e, no caso das estratégias baseadas em *replay*, a um subconjunto restrito de dados anteriores mantido em memória. Após o treinamento na tarefa k, o modelo é avaliado em todos os conjuntos de teste das tarefas já observadas, sem acesso ao identificador da tarefa durante a inferência (*single-head*). Por exemplo, ao final da terceira tarefa no MNIST, o modelo deve classificar corretamente entre todas as seis classes já vistas, mesmo sem saber a qual tarefa pertence cada amostra.

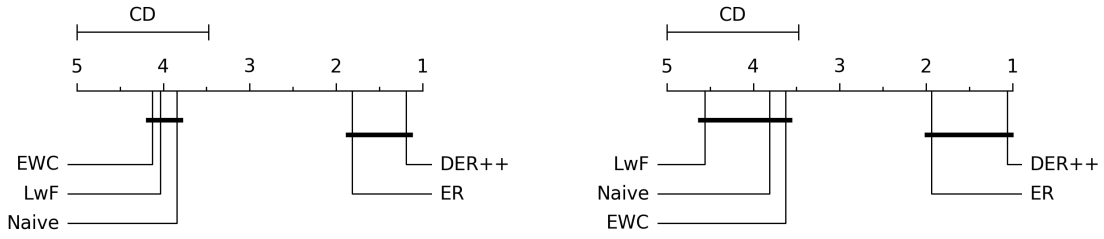
Toda a execução segue um protocolo determinístico, com controle de sementes e padronização das divisões de dados, de modo a garantir a reprodutibilidade dos experimentos. Nesse contexto, cada configuração experimental é repetida com cinco sementes fixas (0, 42, 1337, 2024, 999), permitindo capturar a variabilidade decorrente da inicialização dos modelos. Com isso, obtém-se uma avaliação mais robusta para fornecer uma base consistente para a análise estatística dos resultados.

5. Resultados

Esta seção analisa o esquecimento de forma progressiva. Primeiro, avalia-se o desempenho global das estratégias (Subseção 5.1). Em seguida, investiga-se quando, como e por que o esquecimento ocorre ao longo do treinamento (Subseção 5.2).

5.1. Desempenho Global das Estratégias

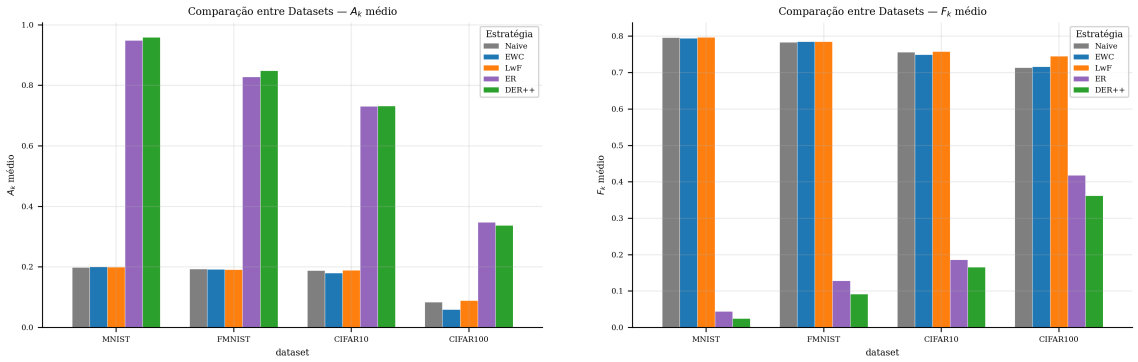
A Figura 3 apresenta os rankings de Friedman para A_k e F_k. A Figura ordena os algoritmos com melhor desempenho da direita para a esquerda. Os resultados evidenciam a separação estatística clara entre métodos de replay (DER++, ER) e os demais, visto que eles não são interligados pela linha horizontal da distância crítica de Nemenyi.



(a) Accuracy (A_k) **(b) Forgetting (F_k)**

Figura 3. Diagramas de diferença crítica (CD) com teste de Nemenyi ($\alpha = 0.05$).

Essa distinção é reforçada na Figura 4: enquanto DER++ e ER mantêm alto A_k em todos os cenários, Naive, EWC e LwF permanecem em patamares significativamente inferiores (em torno de ~ 0.20 nos cenários mais simples). Em F_k, observa-se o padrão complementar: métodos sem replay apresentam maior esquecimento nos cenários mais simples, enquanto métodos baseados em replay controlam a degradação. No entanto, no CIFAR-100, observa-se um comportamento contraintuitivo: métodos sem replay apresentam valores relativamente menores de F_k, apesar de exibirem desempenho final significativamente inferior, sugerindo que o menor esquecimento está associado a aprendizado limitado.



(a) A_k médio **(b) F_k médio**

Figura 4. Comparação global entre estratégias e datasets.

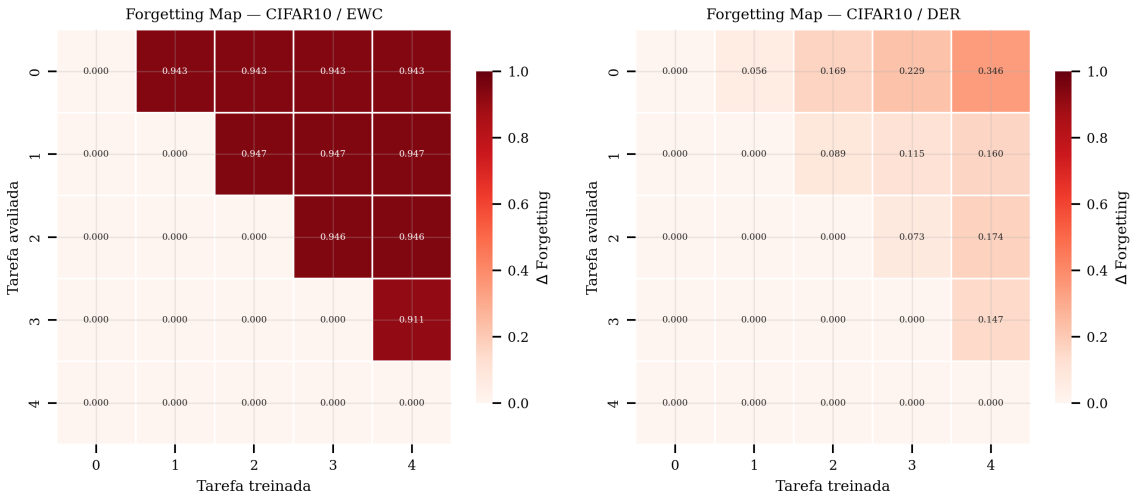
Esses resultados evidenciam uma limitação central das métricas agregadas: valores semelhantes de F_k podem corresponder a comportamentos estruturalmente distintos. Em particular, o contraste entre a redução de F_k em métodos sem replay e o aumento desse valor em métodos baseados em replay no CIFAR-100 indica que a métrica não captura adequadamente as diferenças na dinâmica de aprendizado entre abordagens. Além disso, diferenças substanciais de A_k entre estratégias não são proporcionalmente explicadas por variações em F_k, reforçando que a análise global é insuficiente para caracterizar o fenômeno. Assim, a análise não responde adequadamente às perguntas de pesquisa,

pois não revela como o esquecimento se manifesta internamente, motivando diretamente a necessidade da análise estrutural proposta neste trabalho.

5.2. Análise Estrutural do Esquecimento Catastrófico

Critério de seleção para análise estrutural: Embora a avaliação global considere todas as estratégias e *datasets*, as análises estruturais concentram-se em duas estratégias representativas e contrastantes: EWC e DER++. Essa escolha baseia-se na avaliação global (Subseção 5.1), na qual os testes de Friedman e Nemenyi evidenciaram a separação entre métodos com e sem *replay*, além da similaridade entre LwF e EWC, e ER e DER++. Além disso, como os diferentes *datasets* apresentaram padrões estruturais semelhantes, apenas o CIFAR-10 foi utilizado nas análises dos eixos temporal, decisório e representacional. Essa escolha deve-se ao seu nível intermediário de complexidade visual, suficientemente desafiador para evidenciar os efeitos do esquecimento catastrófico.

Análise Temporal: A Figura 5 apresenta os *forgetting maps* focados nos representantes contrastantes EWC e DER++. No EWC, o esquecimento ocorre imediatamente após cada nova tarefa, caracterizando colapso abrupto. No DER++, a degradação cresce com a distância temporal, indicando interferência acumulativa. Esses comportamentos revelam que o esquecimento não ocorre de forma uniforme, mas depende da estratégia adotada, evidenciando mecanismos distintos e dinâmicas temporais específicas de esquecimento. Embora a análise temporal caracterize *quando* o esquecimento ocorre, ela não revela como essa degradação se distribui ao nível das classes.

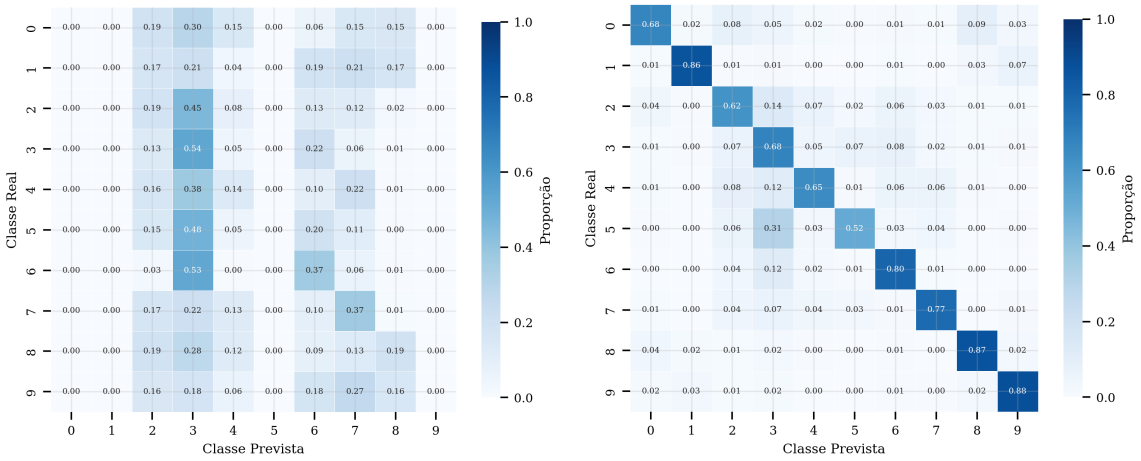


(a) EWC **(b) DER++**

Figura 5. *Forgetting maps* no CIFAR-10 para EWC e DER++.

Análise de Decisão: As matrizes de confusão da Figura 6 evidenciam como os erros se distribuem ao nível de classes. Por exemplo, no EWC, há forte viés de recência, com predições concentradas fortemente nas classes da última tarefa. Já o DER++ mantém erros semanticamente consistentes, preservando a diagonal dominante. Esses padrões distinguem exemplos de falha que não podem ser inferidos a partir de métricas globais.

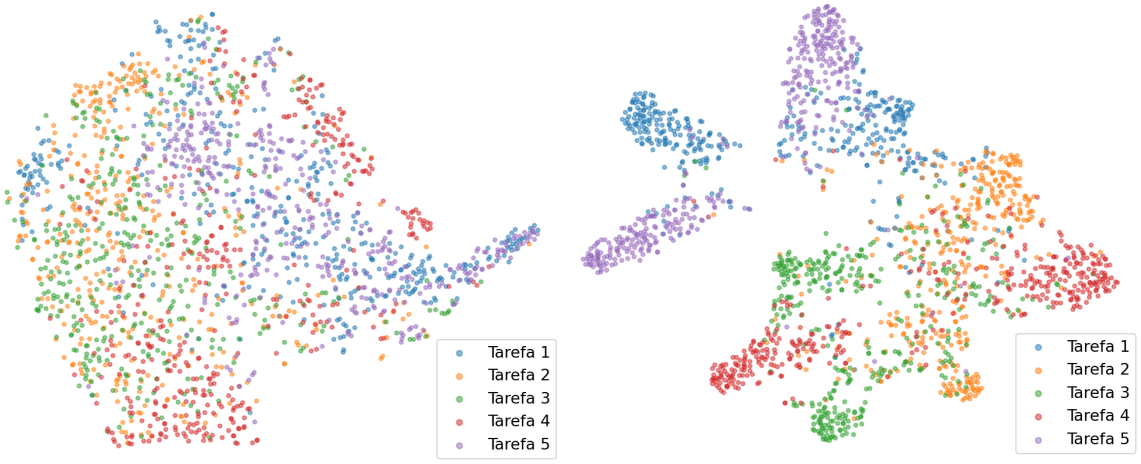
Embora essas análises revelem como o erro se estrutura entre classes, elas ainda não explicam a origem desses padrões. Para isso, é necessário examinar a organização do espaço de representações.



(a) EWC (b) DER++

Figura 6. Matrizes de confusão após a tarefa 5 no CIFAR-10.

Análise Representacional: A Figura 7 ilustra essa organização, apresentando as projeções t-SNE. A partir das projeções, nota-se que no EWC ocorre colapso representacional, enquanto no DER++, os *clusters* mantêm-se bem definidos por tarefa. Esses padrões indicam que a capacidade de retenção está diretamente associada à organização do espaço de representações, conectando o comportamento observado à causa do esquecimento.



(a) EWC (b) DER++

Figura 7. Projeções t-SNE após a tarefa 5 no CIFAR-10.

Discussão: A análise conjunta dos três eixos fornece evidências consistentes para responder à pergunta de pesquisa de como, quando e por que o esquecimento catastrófico se manifesta no cenário CIL, e não apenas quantitativamente. No **eixo temporal**, ele ocorre como colapso imediato (estratégias de regularização) ou degradação progressiva (estratégias baseadas em *replay*). No **eixo decisório**, manifesta-se por padrões

distintos de erro, como a tendência a favorecer classes recentes ou a concentração excessiva de predições em poucas classes (característica da regularização), em contraste com a manutenção de acertos consistentes na diagonal principal (característica do *replay*). No **eixo representacional**, está diretamente relacionado à organização do espaço de representações, variando entre o colapso representacional (observado na regularização) e a preservação da separação de representações (alcançada pelo *replay*).

Assim, os resultados reforçam a necessidade de avaliações que considerem simultaneamente os três eixos propostos para permitir diagnósticos mais precisos e escolhas mais confiáveis de modelos em aplicações contínuas. Nesse contexto, observa-se que as estratégias baseadas em *replay*, especialmente DER++, apresentaram maior estabilidade temporal, decisória e representacional ao longo do treinamento. Contudo, a análise estrutural evidencia lacunas críticas nas abordagens existentes: os métodos de regularização carecem de mecanismos para evitar o colapso representacional imediato, enquanto os métodos de *replay* ainda precisam superar a interferência acumulativa que causa a degradação progressiva. Isso sugere que o aprimoramento de futuras estratégias deve transcender a simples preservação de parâmetros ou retenção de amostras, focando na manutenção ativa da integridade geométrica e semântica do espaço de características a longo prazo.

6. Conclusão e Trabalhos Futuros

Este trabalho investigou o fenômeno do esquecimento catastrófico no cenário de *Class-Incremental Learning* sob uma perspectiva estrutural, permitindo examinar o esquecimento além de métricas agregadas. Para isso, foi proposto um protocolo unificado de análise que integra três eixos complementares: temporal, decisório e representacional. O protocolo foi validado por meio de um estudo de caso envolvendo cinco estratégias representativas da literatura (Naive, EWC, LwF, ER e DER++) em quatro benchmarks com diferentes níveis de complexidade.

Os resultados obtidos demonstram que o esquecimento catastrófico em CIL se manifesta de forma estrutural e não apenas quantitativa, envolvendo dinâmicas temporais distintas, padrões específicos de erro e alterações na organização do espaço representacional. Em conjunto, os achados evidenciam que métricas agregadas são insuficientes para caracterizar adequadamente o comportamento dos modelos, enquanto o protocolo proposto permite identificar mecanismos distintos de esquecimento entre estratégias de aprendizado contínuo. Nesse contexto, estratégias baseadas em replay, especialmente DER++, apresentaram maior estabilidade ao longo do treinamento, preservando de forma mais consistente o desempenho e a estrutura das representações internas.

Apesar dos avanços, este trabalho apresenta algumas limitações. Os experimentos foram conduzidos em um protocolo controlado com fronteiras de tarefa explícitas e arquiteturas sem pré-treinamento, o que restringe a generalização dos resultados para cenários mais realistas. Como direções futuras, destacam-se: (i) a extensão da análise para cenários online e sem fronteiras de tarefa definidas; (ii) a aplicação do protocolo proposto em outros domínios além da visão computacional; e (iii) a investigação de estratégias adaptativas de gerenciamento de memória, como políticas dinâmicas de atualização de buffer, para mitigar limitações observadas em métodos de replay em cenários de maior complexidade.

Referências

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. (2020). Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15920–15930.

Chaudhry, A., Dokania, P. K., Ajanthan, T., and Torr, P. H. S. (2018). Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 556–572.

Chen, H., Wang, Y., Fan, Y., Jiang, G., and Hu, Q. (2025). Reducing class-wise confusion for incremental learning with disentangled manifolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10121–10130.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Li, Z. and Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

Lopez-Paz, D. and Ranzato, M. (2017). Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Mermillod, M., Bugaiska, A., and BONIN, P. (2013). The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, Volume 4 - 2013.

Ramasesh, V. V., Dyer, E., and Raghu, M. (2020). Anatomy of catastrophic forgetting: Hidden representations and task geometry. *arXiv preprint arXiv:2007.07400*.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. (2017). iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. (2019). Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

van de Ven, G. M., Soures, N., and Kudithipudi, D. (2024). Continual learning and catastrophic forgetting.

van de Ven, G. M., Tuytelaars, T., and Tolias, A. S. (2022). Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. (2017). Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*.