MCC: Multi-Centroid Curriculum for Pre-defined Centroid-based Classifiers

Nattan Gomes Silva Cortes[1], **Natália Azeredo Viana da Silva**[1], **Rogério Ferreira de Moraes**[1]

[1]Instituto Politécnico – Universidade Federal do Rio de Janeiro (UFRJ)
Rua Aloísio da Silva Gomes, 50 – Granja dos Cavaleiros
27930-560 – Macaé – RJ – Brasil

nattango132@gmail.com, natalia.azeredo.eng@gmail.com
rogeriofm@macae.ufrj.br

Abstract. *In classification problems, methods that fix class centers on the hypersphere, such as PEDCC and POD, assume that each class fits into a single center, but many encompass distinct visual subgroups. We propose the Multi-Centroid Curriculum (MCC): an adaptive curriculum that creates and removes sub-centers per class during training, guided by data complexity; with a single center, it reproduces the formulation of the original methods. On CIFAR-10/100 (ResNet-18 and ShuffleNet), MCC outperforms the POD baseline across all evaluated configurations (winning in 20 out of 20 runs). A controlled analysis demonstrates that the gain stems specifically from the adaptive curriculum, not from learning rate restarts or a fixed number of centers.*

Resumo. *Em problemas de classificação, métodos que fixam os centros de classe na hiperesfera, como PEDCC e POD, supõem que cada classe cabe em um único centro, mas muitas reúnem subgrupos visuais distintos. Propomos o Multi-Centroid Curriculum (MCC): um currículo adaptativo que cria e remove subcentros por classe durante o treino, guiado pela complexidade dos dados; com um único centro, ele reproduz a formulação dos métodos originais. Em CIFAR-10/100 (ResNet-18 e ShuffleNet), o MCC supera o baseline POD em todas as configurações avaliadas (com vitórias em 20 de 20 execuções). Uma análise controlada demonstra que o ganho provém especificamente do currículo adaptativo, não de reinícios da taxa de aprendizado ou de um número fixo de centros.*

1. Introdução

Redes neurais profundas para classificação aprendem, implicitamente, uma geometria do espaço de características. Uma linha de trabalho recente torna essa geometria explícita: em vez de deixar a última camada aprender livremente os pesos de classe, fixam-se antes do treino centros de classe uniformemente distribuídos na hiperesfera unitária e treina-se a rede para alinhar as características de cada amostra ao centro da sua classe. PEDCC [Zhu et al. 2020] e POD [Zhu and Zu 2023] são representantes dessa estratégia, e margens angulares como o AM-Softmax [Wang et al. 2018] compartilham o mesmo princípio de maximizar a separação inter-classe sobre a esfera. Essas abordagens trazem alvos estáveis, boa separação e convergência mais previsível.

Todas elas, porém, assumem que cada classe é unimodal (um único centro por classe), no entanto, muitas classes reais não são. Uma classe pode reunir subgrupos visuais distintos, tais como variações de pose, iluminação, fundo ou subtipos do mesmo rótulo. Forçar esses subgrupos a colapsarem em um único centro desperdiça capacidade representacional e pode degradar a fronteira de decisão. Com isso, buscamos responder à seguinte pergunta: Múltiplos subcentroides por classe, introduzidos de forma adaptativa ao longo do treino, melhoram a classificação sem alterar a função de perda subjacente?

Respondemos afirmativamente com o *Multi-Centroid Curriculum* (MCC). Em vez de propor uma nova função de perda, o MCC aproveita as funções de perda existentes, e através do acréscimo de subcentroides, busca aprimorar a acurácia. Cada classe começa com um único centro e, ao longo do treino, ganha (ou perde) subcentroides conforme os dados justifiquem. As divisões são geometricamente seguras (um filho nasce no bordo de um cone estreito em torno do subcentroide mais recente) e as podas revertem divisões que não ajudaram. Por construção, quando todas as classes permanecem em $K{=}1$, o MCC recai exatamente no *baseline* (POD, PEDCC) correspondente, o que dá uma âncora de corretude e um ponto de partida neutro, com exceção do AM-Softmax, que é baseado em centros não fixos.

As contribuições são:

- Um mecanismo de currículo adaptativo que decide dinamicamente, por classe, a quantidade de subcentroides necessários, utilizando separação espacial em cones e poda guiada pelo desempenho de validação;
- Uma formulação modular acoplável que reaproveita métodos baseados em margem (POD, PEDCC e AM-Softmax) com garantia de equivalência matemática exata (bit-equivalente) aos *baselines* (POD, PEDCC) quando $K = 1$, assegurando justiça estrutural nas comparações; e
- Uma validação empírica composta por um estudo de ablação para isolar a origem do ganho de desempenho, suportado por testes estatísticos rigorosos (Wilcoxon pareado com correção de Holm e cálculo de tamanho de efeito).

Sobre POD, o MCC obtém $+1{,}77$ pontos percentuais de acurácia média, com efeito grande e consistente, e a ablação mostra que o ganho vem do currículo adaptativo.

O restante deste artigo estrutura-se na fundamentação teórica (Seção 2), metodologia (Seção 3), protocolos experimentais utilizados (Seção 4), análise dos resultados (Seção 5), discussão e validade do estudo (Seção 6) e conclusão (Seção 7).

2. Trabalhos Relacionados

Centroides pré-definidos e margens angulares. Uma família de perdas impõe uma geometria explícita à última camada. Trabalhos que usam margens angulares como SphereFace [Liu et al. 2017], AM-Softmax [Wang et al. 2018] e ArcFace [Deng et al. 2019]; normalizam características e pesos e inserem uma margem (no ângulo ou no cosseno) para ampliar a separação inter-classe; o Center Loss [Wen et al. 2016] aprende um centro por classe e o atrai via MSE. O PEDCC [Zhu et al. 2020] substitui os pesos da última camada por centros equidistantes na hiperesfera, gerados por repulsão entre cargas e mantidos congelados, treinando com AM-Softmax mais um termo de atração (MSE) ao centro; o POD [Zhu and Zu 2023] dispensa o Softmax, usando uma perda de cosseno adaptativa à

norma (NaC), $\cos_{Na} = \cos\theta/(1 + \delta/\|x\|)$, somada a uma restrição de auto-correlação (SC) que decorrelaciona as dimensões das características. Bases pré-fixadas quase ortogonais reaparecem em trabalhos como os Equiangular Basis Vectors [Shen et al. 2023]. Todas essas abordagens representam cada classe por um único protótipo. O MCC usa o AM-Softmax, o PEDCC ou o POD como cabeça e relaxa exatamente essa premissa.

Múltiplos protótipos (centros) por classe. Redes de protótipos [Snell et al. 2017] classificam pela distância ao protótipo de cada classe, ainda com um centro por classe. Para acomodar a multimodalidade intra-classe, o SoftTriple [Qian et al. 2019] associa vários centros aprendidos a cada classe e usa uma regularização que funde centros redundantes; o Sub-center ArcFace [Deng et al. 2020] fixa K sub-centros por classe e exige que a amostra se aproxime de qualquer um deles, ganhando robustez a rótulos ruidosos. No regime *few-shot*, o Infinite Mixture Prototypes [Allen et al. 2019] infere o número de *clusters* por classe via inferência bayesiana não-paramétrica, recuperando o caso unimodal como caso particular. O MCC situa-se nessa linha, mas difere em três pontos: (i) os subcentros nascem de centros pré-definidos e congelados, herdando a separação do PEDCC/POD, em vez de pesos livremente aprendidos; (ii) o número de subcentros é adaptativo e por classe, decidido por um currículo explícito de divisões geométricas e podas guiadas por acurácia em vez de um K fixo [Deng et al. 2020] ou de uma fusão por regularização [Qian et al. 2019]; e (iii) com um único protótipo o MCC é bit-equivalente ao *baseline* (PEDCC, POD), o que sustenta o protocolo pareado.

3. Metodologia

3.1. Geometria dos centroides

Partimos de C centroides predefinidos $\{c_0, \ldots, c_{C-1}\}$, que são vetores unitários em $\mathbb{R}^d$ ($d = 256$) gerados por uma simulação de repulsão na hiperesfera com semente fixa (PEDCC). Durante o treinamento adaptativo, quando uma classe y necessita de maior capacidade representativa, ela é subdividida em $K_y \geq 1$ subcentroides.

Para preservar a topologia global do espaço latente, cada nova divisão ocorre dentro de um cone estreito em torno do subcentroide que lhe serve de referência, que é o centroide pai na primeira divisão. A representação global da classe y, utilizada para classificação e ancoragem da função de perda, é o vetor $\bar{c}_y$, definido como a média dos seus subcentroides projetada de volta na hiperesfera unitária:

$$\bar{c}_y = \frac{\sum_{k=1}^{K_y} c_{y,k}}{\left\| \sum_{k=1}^{K_y} c_{y,k} \right\|} \tag{1}$$

O uso de $\bar{c}_y$ (em vez de um simples ponto médio geométrico) é intencional: graças à restrição espacial do cone, a projeção mantém-se geometricamente próxima ao centroide pai, evitando o colapso em regiões vazias do espaço. Além disso, essa formulação fortalece a otimização de funções como o POD, fornecendo uma âncora de classe contínua e estável para a perda de cosseno; já a restrição de auto-correlação (SC) usa o subcentroide atribuído à amostra.

Assim, uma *feature* x será classificada como pertencente à classe y se a função de perda (PEDCC, POD, AM-Softmax) determinar que ela está mais próxima do representante $\bar{c}_y$. Caso a classe não sofra divisões ou sofra poda total ($K_y = 1$), temos que

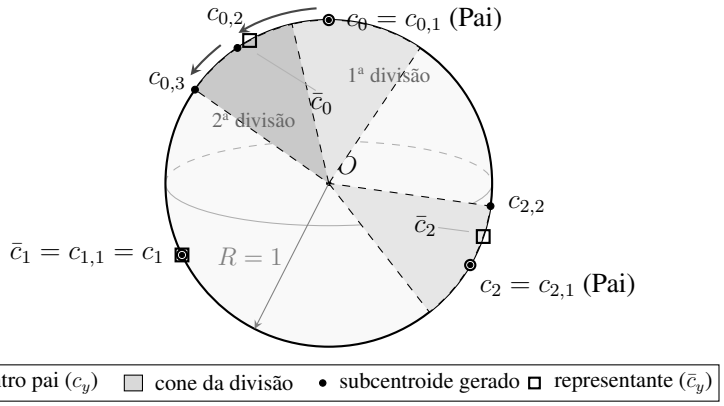


Figura 1. Mecanismo espacial do currículo adaptativo. Na Classe 0 ($K_0 = 3$), o filho $c_{0,2}$ nasce no bordo de um cone em torno de c_0 e o filho seguinte $c_{0,3}$ nasce no bordo de um cone menor em torno de $c_{0,2}$: cada divisão parte do subcentroide criado por último, com abertura decrescente. A Classe 2 mostra uma única divisão e a Classe 1 ($K_1 = 1$) demonstra a retrocompatibilidade estrutural.

$\bar{c}_y = c_y$. Nesse cenário, o MCC recai matematicamente no comportamento exato da função de perda original utilizada como base.

3.2. Distribuição e Abertura dos Cones

Os C centros pai predefinidos estão distribuídos sobre a superfície de uma hiperesfera unitária de raio 1. Para garantir justiça e máxima separação, exige-se que esses centros sejam rigorosamente equidistantes, possuindo o mesmo ângulo entre qualquer par. Matematicamente, a construção de C vetores mutuamente equidistantes requer que a dimensionalidade do espaço latente d satisfaça $d \geq C - 1$ [Zhu et al. 2020]. Em nossos experimentos, essa condição é amplamente satisfeita (ex: $d = 256 \geq 100 - 1$ no CIFAR-100). Sob esta condição de simetria e maximização de margem, o centro de massa do sistema reside na origem, implicando que $\sum_{i=0}^{C-1} c_i = \vec{0}$. A simulação de repulsão do PEDCC aproxima empiricamente essa configuração ótima (simplexo regular ideal). Assumindo essa configuração ótima, vale o resultado a seguir.

Proposição 1. *O ângulo θ_p entre quaisquer dois centros pai distintos no conjunto $\{c_0, \ldots, c_{C-1}\}$ é dado por:*

$$\theta_p = \arccos\left(\frac{-1}{C-1}\right)$$

Demonstração. Pela álgebra linear, como os vetores repousam na hiperesfera unitária, temos que $\|c_i\|^2 = \langle c_i, c_i \rangle = 1$. O produto interno entre dois vetores distintos c_i e c_j ($i \neq j$) é dado por $\langle c_i, c_j \rangle = \|c_i\|\|c_j\| \cos(\theta_p) = \cos(\theta_p)$.

Partindo da propriedade de que a soma dos centros é o vetor nulo:

$$0 = \left\| \sum_{i=0}^{C-1} c_i \right\|^2 = \left\langle \sum_{i=0}^{C-1} c_i, \sum_{j=0}^{C-1} c_j \right\rangle$$

Expandindo o produto interno, separamos os termos onde $i = j$ e $i \neq j$:

$$0 = \sum_{i=0}^{C-1} \langle c_i, c_i \rangle + \sum_{\substack{i,j=0 \\ i \neq j}}^{C-1} \langle c_i, c_j \rangle$$

Como existem C termos na primeira soma e $C(C-1)$ termos na segunda soma, substituímos pelas suas definições geométricas:

$$0 = C + C(C-1)\cos(\theta_p)$$

Isolando $\cos(\theta_p)$, obtemos a prova geométrica:

$$\cos(\theta_p) = \frac{-C}{C(C-1)} = \frac{-1}{C-1}$$

□

A Proposição 1 vale para a configuração ideal. O PEDCC gera os centros por repulsão iterativa, que apenas se aproxima do simplexo regular. Assim, θ_p é o ângulo teórico usado para definir a abertura dos cones, e não uma medida dos centros de fato gerados; pequenos desvios da equidistância mudam um pouco a margem de não-sobreposição. Não medimos o ângulo observado entre pares de centros, o que fica para trabalhos futuros.

Com a topologia dos centros pai estabelecida, definimos a região de expansão para a criação do currículo. Os subcentroides filhos de uma classe nascem na interseção do bordo de um cone direcional com a hiperesfera unitária.

Seja c_p o vetor de referência da divisão: o centro pai c_y na primeira divisão da classe e, nas divisões seguintes, o subcentroide criado por último. A abertura do cone θ_{cone} em relação a c_p é definida dinamicamente por:

$$\theta_{cone} = \frac{\theta_p}{2 \cdot K_{\max}}$$

onde $K_{\max}$ é o maior número de subcentroides entre as classes já contabilizada a divisão em curso. Como todas as classes se dividem juntas, isso equivale a $\max_y K_y + 1$, de modo que a primeira divisão usa $\theta_{cone} = \theta_p/4$. O fator 2 no denominador marca o limite de tangência entre cones de classes adjacentes ($2 \cdot \theta_{cone} \leq \theta_p$): na primeira divisão os cones não se sobrepõem, com folga ($K_{\max} \geq 2$), enquanto o escalonamento por $K_{\max}$ estreita adaptativamente o cone conforme a complexidade (e o número de subdivisões) da rede aumenta, evitando colisões indesejadas entre subcentroides de classes distintas no espaço latente. A Figura 1 ilustra a operação de divisão.

3.3. Atribuição e Classificação

Para acomodar a multiplicidade de centros sem desestabilizar o treinamento, o MCC distingue a regra de atribuição usada durante a inferência daquela usada para a atração intraclasse no treino. Na etapa de inferência (e classificação geral), para garantir uma fronteira de decisão contínua e mitigar o ruído de subcentroides individuais, a pontuação de uma classe c é dada pela similaridade do cosseno entre a característica normalizada e o representante da classe ($\bar{c}_c$). A predição final $\hat{y}$ segue a regra convencional:

$$\hat{y} = \arg\max_c \left(\frac{x}{\|x\|} \cdot \bar{c}_c \right) \tag{2}$$

Durante o treinamento, a atração intra-classe exige uma *hard assignment* (atribuição rígida). Cada amostra x pertencente à classe y é vinculada estritamente ao seu subcentroide mais próximo c_{y,k^*}, selecionado via menor distância Euclidiana:

$$k^* = \arg \min_{k \in \{1,\dots,K_y\}} \left\| \frac{x}{\|x\|} - c_{y,k} \right\|^2 \tag{3}$$

Essa separação garante que a rede crie agrupamentos locais precisos (em torno de c_{y,k^*}) enquanto mantém uma representação global estável ($\bar{c}_y$) para o termo principal da perda base.

Vale notar que, como a pontuação usa apenas $\bar{c}_c$, a decisão final continua sendo tomada com um representante por classe, mesmo que ele seja formado por K_c subcentroides. Os vários centros funcionam, então, como um mecanismo de treino que reorganiza o extrator de características, e não como uma regra de decisão com vários protótipos. Isso é coerente com a ablação da Seção 5, em que vários centros estáticos não trazem ganho e o currículo adaptativo traz. Comparar a inferência por $\bar{c}_c$, pelo subcentroide mais próximo e pelo centro pai original seria um teste útil, que não fizemos neste trabalho.

3.4. Função de Perda Acoplável e Regularização

O MCC não introduz uma nova função de perda isolada, mas sim um *framework* aditivo que acopla a adaptação dinâmica do número de subcentroides (K) a funções baseadas em cosseno já estabelecidas (POD, AM-Softmax, PEDCC). A perda total do sistema para um lote de dados é formalizada como:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \mathbb{I}_{[K_{\max}>1]} \mathcal{L}_{MSE} + \frac{\lambda_{bal}}{|\mathcal{C}^+|} \sum_{y \in \mathcal{C}^+} \mathcal{L}_{bal}^{(y)} \tag{4}$$

Onde $\mathcal{L}_{base}$ é a função de perda original escolhida e $\mathcal{C}^+$ é o conjunto de classes divididas ($K_y > 1$) presentes no lote. A função indicadora $\mathbb{I}_{[K_{\max}>1]}$ atua como uma trava estrutural: quando todas as classes possuem apenas um centro ($K_{\max} = 1$), os termos adicionais são inativos e o MCC recai matemática e computacionalmente no comportamento exato de seu *baseline* (propriedade validada por testes de equivalência *bit-a-bit*). No PEDCC, cuja perda já contém um termo MSE, não há duplicação: esse termo passa a apontar para o subcentroide mais próximo. O peso de balanceamento é fixado empiricamente em $\lambda_{bal} = 1,0$.

Quando a classe sofre divisões ($K_y > 1$), os termos ativados são:

- Atração Intra-classe ($\mathcal{L}_{MSE}$): Aplica-se o erro quadrático médio entre a amostra e seu subcentroide mais próximo (selecionado via k^*), forçando a coesão espacial dentro do cone recém-criado.
- Regularização de Balanceamento ($\mathcal{L}_{bal}^{(y)}$): Previne o colapso do currículo, onde um único subcentroide concentra todas as amostras da referida classe. Como a seleção via $\arg\min$ não é diferenciável, empregamos uma *soft assignment* para propagar gradientes.

Calcula-se a probabilidade $p_{i,k}$ de uma amostra i pertencer ao subcentroide k (dentro da classe y) via uma função Softmax escalonada por uma temperatura τ (fixada em

0,1):

$$p_{i,k} = \frac{\exp\left(\cos(\frac{x_i}{\|x_i\|}, c_{y,k})/\tau\right)}{\sum_{j=1}^{K_y} \exp\left(\cos(\frac{x_i}{\|x_i\|}, c_{y,j})/\tau\right)} \tag{5}$$

A utilização média do subcentroide k por todas as N_y amostras da classe y presentes no lote é $\bar{p}_k = \frac{1}{N_y} \sum_{i=1}^{N_y} p_{i,k}$. A penalidade para a classe y é então definida como a divergência entre a entropia máxima possível para K_y opções e a entropia real da distribuição de uso:

$$\mathcal{L}_{bal}^{(y)} = \ln(K_y) - \left(- \sum_{k=1}^{K_y} \bar{p}_k \ln(\bar{p}_k) \right) \tag{6}$$

Ao minimizar essa penalidade, a rede ajusta os pesos de extração de características para distribuir as amostras homogeneamente entre os K_y subcentroides disponíveis.

3.5. Currículo Adaptativo

O algoritmo de ajuste dinâmico do número de subcentroides é guiado exclusivamente pelo desempenho de validação empírico, eliminando a necessidade de fixar um hiperparâmetro de quantidade de centros *a priori*. O currículo opera pelas seguintes regras:

- Fase Inicial: O treinamento inicia com $K = 1$ para todas as classes.
- Gatilho de Divisão: Quando a métrica de acurácia na validação atinge um platô e o critério de *early stopping* é acionado pela primeira vez, todas as classes sofrem uma divisão ($K \rightarrow K + 1$). Ocorre então um *warm restart*, em que a taxa de aprendizado volta para $0{,}1\,\eta_0$ e a paciência é zerada.
- Critério de Poda: Nas convergências subsequentes, o método avalia a acurácia isolada por classe. Subcentroides de classes que apresentaram piora no desempenho em relação ao ciclo anterior são imediatamente podados (revertendo a divisão).
- Condição de Parada: Uma nova divisão global só é autorizada se a acurácia geral do sistema apresentar melhoria estrita. O processo se repete até que os ganhos cessem ou o limite arquitetural de $K_{\text{lim}} = 20$ seja atingido.

Detalhes de implementação. Cada divisão de uma classe y parte do subcentroide mais recente $p = c_{y,K_y}$, que é o próprio centro pai c_y na primeira divisão. A direção tangencial é $u_y = (I - pp^\top)g \,/\, \|(I - pp^\top)g\|$, com $g \sim \mathcal{N}(0, I_d)$ sorteado a partir de uma semente fixa somada ao número de subcentroides após a divisão, o que evita favorecer qualquer direção do espaço latente. Assim como os centros pai do PEDCC, essa geometria é mantida igual nas cinco execuções: a variação entre sementes isola o efeito da inicialização da rede e da ordem dos dados, e o pareamento entre cada variante do MCC e o seu *baseline* é preservado. O novo subcentroide é normalizado para voltar à hiperesfera. Os subcentroides ficam congelados: assim como os centros do PEDCC e do POD, não recebem gradiente, e só o extrator de características é treinado. A poda remove o subcentroide criado por último, desfazendo a divisão anterior. Também não há número fixo de épocas por ciclo: cada ciclo termina quando a paciência de 40 épocas se esgota na validação, e toda divisão ou poda dispara um *warm restart* que volta a taxa de aprendizado para

Algoritmo 1. Currículo adaptativo do MCC. Entrada: centros pai congelados $\{c_y\}$, taxa base η_0, paciência P, teto K_{lim}. REINICIA(): $\eta \leftarrow 0{,}1\,\eta_0$, recria o escalonador de platô e zera a paciência.

$K_y \leftarrow 1\ \forall y;\ fase \leftarrow 0;\ acc^{\text{ref}} \leftarrow 0;\ dividiu \leftarrow$ FALSO
repita
  treina até esgotar a paciência P; $acc \leftarrow$ melhor acurácia de validação até aqui
  se $dividiu$ **então** ▷ ciclo de avaliação da divisão anterior
    poda toda classe com $K_y > 1$ cuja acurácia a_y caiu em relação a a_y^{ref}: $K_y \leftarrow K_y{-}1$
    $acc^{\text{ref}} \leftarrow acc$; $dividiu \leftarrow$ FALSO; REINICIA(); **siga**
  fim se
  se ($fase = 0$ ou $acc > acc^{\text{ref}}$) e $\max_y K_y < K_{\text{lim}}$ **então**
    $a_y^{\text{ref}} \leftarrow$ acurácia da classe y; $acc^{\text{ref}} \leftarrow acc$; $\theta_{cone} \leftarrow \theta_p/\big(2(\max_y K_y{+}1)\big)$
    divide toda classe: $c_{y,K_y+1} \leftarrow \cos(\theta_{cone})\, c_{y,K_y} + \sin(\theta_{cone})\, u_y$, u_y unitário em $c_{y,K_y}^{\perp}$
    $K_y \leftarrow K_y{+}1\ \forall y;\ fase \leftarrow fase{+}1;\ dividiu \leftarrow$ VERDADEIRO; REINICIA()
  fim se
até que a divisão seja recusada (sem melhoria estrita ou teto K_{lim} atingido)

$0{,}1\,\eta_0$, um valor fixo que não acumula entre reinícios. O treino acaba quando uma divisão é recusada, de modo que o total de épocas depende dos dados (Algoritmo 1).

4. Protocolo Experimental

Conjuntos de Dados e Arquiteturas. Os experimentos foram conduzidos nas bases CIFAR-10 (5.000 imagens por classe) e CIFAR-100 (500 imagens por classe) [Krizhevsky and Hinton 2009], ambas com resolução de $32{\times}32$ pixels. O treinamento foi realizado do zero utilizando duas arquiteturas consolidadas: ResNet-18 [He et al. 2016] e ShuffleNet [Ma et al. 2018]. A combinação destas bases de dados com as duas arquiteturas define quatro cenários de avaliação distintos.

Métodos Avaliados. A análise engloba sete abordagens: o modelo padrão treinado com Entropia Cruzada (PLAINMODEL), os *baselines* de margem (AM_SOFTMAX, PEDCC, POD) e suas respectivas extensões curriculares (MCC_AM_SOFTMAX, MCC_PEDCC, MCC_POD). É fundamental pontuar uma distinção arquitetural: no PEDCC e no POD, os centros de classe são fixos na hiperesfera, garantindo que o MCC recaia em uma equivalência matemática exata a esses *baselines* durante o estágio inicial ($K = 1$). O AM-Softmax, por outro lado, atualiza seus centros dinamicamente via gradiente, mas foi incluído para validar o comportamento do currículo sobre margens angulares. Adicionalmente, para o estudo de ablação, incluímos o POD+SCHEDULE (que aplica apenas os *warm restarts* da taxa de aprendizado, com o mesmo gatilho e fator do MCC, até três reinícios, sem a expansão geométrica) e o MCC(POD) com $K{=}3$ fixo (inicialização estática com múltiplos protótipos, sem o mecanismo adaptativo).

Configurações de Treinamento. Para assegurar a robustez e generalização do método, utilizamos uma configuração hiperparamétrica unificada para ambos os *datasets*: tamanho de lote de 256, taxa de aprendizado inicial de $0{,}1$, dimensão latente $d{=}256$, parâmetros do AM-Softmax $s{=}15$ e $m{=}0{,}5$, peso da restrição de auto-correlação (SC) $\lambda{=}1$, teto de subcentroides $K_{\text{lim}}{=}20$ e fator de decaimento no *warm restart* de $0{,}1$. Apenas o *baseline* AM-Softmax usa aquecimento linear da taxa nas cinco primeiras épocas, por estabilidade no treino do zero; as variantes MCC não usam esse recurso. A divisão dos dados reser-

Tabela 1. Acurácia de teste (%): média ± desvio sobre 5 sementes por célula. Melhor por coluna em negrito.

	CIFAR-10		CIFAR-100	
Método	ResNet18	ShuffleNet	ResNet18	ShuffleNet
Plain	95.17±0.19	92.30±0.32	76.08±1.08	71.41±0.65
AM-Softmax	94.63±0.39	90.80±0.46	75.81±0.52	71.37±0.38
PEDCC	94.83±0.27	91.35±0.29	76.05±0.26	70.96±0.37
POD	95.24±0.11	92.27±0.32	76.38±0.49	70.30±0.55
MCC(am)	95.61±0.23	92.75±0.26	75.59±0.70	71.19±0.47
MCC(pedcc)	95.64±0.26	92.54±0.61	76.09±0.34	71.00±0.43
MCC(pod)	**96.12**±0.23	**93.93**±0.32	**77.62**±0.30	**73.59**±0.79

vou 20% do conjunto de treinamento, de forma estratificada, para validação, mantendo o particionamento original para o conjunto de teste.

Reprodutibilidade e Análise Estatística. Cada cenário de avaliação foi executado com cinco sementes aleatórias distintas ($\{42, 123, 456, 789, 1024\}$), totalizando 20 amostras pareadas por comparação (4 cenários $\times$ 5 sementes). Para cada contraste, reportamos o ganho médio $\pm$ desvio padrão, o total de vitórias, a significância estatística aferida pelo teste de postos sinalizados de Wilcoxon pareado, e o tamanho de efeito de Cohen (d_z). Para controlar o erro do tipo I face às múltiplas comparações (uma contra cada *baseline*), aplicamos a correção sequencial de Holm-Bonferroni [Holm 1979]. A agregação estratificada desses resultados foca na consistência empírica global: superar o *baseline* de forma invariante em todos os cenários fornece a evidência empírica mais robusta de eficácia.

5. Resultados

A Tabela 1 consolida a acurácia de teste para todos os métodos. Observa-se que as três variantes com a integração do MCC superam sistematicamente os seus respectivos *baselines* nos quatro cenários experimentais. O ganho é substancialmente mais expressivo e consistente na transição POD $\rightarrow$ MCC(pod), que atinge os maiores índices de classificação em todas as configurações. Conforme previsto pela motivação do estudo, o impacto mais proeminente ($+3{,}29$ pontos percentuais) ocorre na combinação CIFAR-100 com ShuffleNet, caracterizada por uma rede de menor capacidade representacional lidando com uma geometria espacial de classes mais adensada e multimodal.

5.1. Comparações Pareadas

A Tabela 2 detalha os contrastes principais da formulação. O efeito da adição do MCC ao POD é classificado como grande ($d_z{=}1{,}57$), computando melhora estrita em todas as 20 execuções. As variantes baseadas no PEDCC e no AM-Softmax apresentam ganhos mais conservadores, porém estatisticamente significativos e corroborados após a correção de Holm. O comportamento crucial demonstrado é a invariância do ganho: independentemente da arquitetura da rede ou da densidade do *dataset*, a acoplagem do MCC elevou a precisão do *baseline* original de forma sistemática.

Tabela 2. Testes de Wilcoxon pareados (agregação estratificada, n=20). Δ expresso em pontos percentuais. Todas as comparações rejeitam a hipótese nula sob a correção de Holm (α=0,05).

Comparação	Δ (pp)	Vitórias	p-valor (bilateral)	d_z
POD $\rightarrow$ MCC(pod)	$+1{,}77$	20/20	$8{,}8 \times 10^{-5}$	$+1{,}57$
AM-Softmax $\rightarrow$ MCC(am)	$+0{,}63$	13/20	$0{,}024$	$+0{,}59$
PEDCC $\rightarrow$ MCC(pedcc)	$+0{,}52$	15/20	$0{,}0038$	$+0{,}74$

Tabela 3. Estudo de ablação medindo a acurácia (%): A=POD nativo, B=POD com reinícios, C=MCC(pod) congelado em K=3, D=MCC(pod) com currículo dinâmico ativado.

Dataset	Backbone	A (POD)	B (Reinício)	C (K=3 Fixo)	D (Currículo Adaptativo)
CIFAR10	ResNet18	95.24	95.13	95.19	**96.12**
CIFAR10	ShuffleNet	92.27	91.79	92.23	**93.93**
CIFAR100	ResNet18	76.38	76.15	76.37	**77.62**
CIFAR100	ShuffleNet	70.30	70.47	70.23	**73.59**

5.2. Estudo de Ablação: Origem do Ganho de Desempenho

Para isolar com precisão os fatores responsáveis pela melhoria na acurácia, decompusemos a arquitetura MCC(pod) em seus componentes base (Tabela 3). Os impactos de cada componente foram avaliados através da média das diferenças pareadas calculadas globalmente ao longo de todas as 20 execuções (4 cenários $\times$ 5 sementes).

A avaliação estrita confirma que o ganho de desempenho não deriva unicamente do esquema de reinícios da taxa de aprendizado, cuja diferença isolada resultou em uma oscilação não significativa (B$-$A: $-0{,}16$ pp em média, p=0,14). Da mesma forma, o mero estabelecimento de múltiplos centros engessados desde a primeira época não justificou melhorias (C$-$A: $-0{,}04$ pp em média, p=0,55).

A evolução estatisticamente significativa na fronteira de decisão só se consolida sob a ação do mecanismo completo de currículo adaptativo. Os contrastes D$-$C ($+1{,}81$ pp, $p < 10^{-5}$, d_z=1,66) e D$-$B ($+1{,}93$ pp, $p < 10^{-5}$, d_z=1,88) validam que o refinamento topológico real ocorre através da inserção e poda progressiva e orientada por acurácia. A Figura 2 torna esse processo visível: traça a trajetória de $K_{\max}$ (maior número de sub-centroides entre as classes) ao longo do treinamento, para as três variantes MCC, nas cinco sementes e nas quatro células (*dataset* $\times$ *backbone*). O número e o instante das divisões são decididos pelos dados e variam entre sementes e entre variantes, com o MCC(pod) tipicamente subdividindo mais (até K=7), confirmando que K é adaptativo, e não um hiperparâmetro fixado *a priori*.

6. Discussão e Validade do Estudo

O comportamento empírico observado valida a hipótese geométrica central do método: o maior delta positivo emerge justamente nos ambientes de maior saturação espacial (CIFAR-100) e menor capacidade de generalização bruta (ShuffleNet). Nos cenários de menor complexidade (CIFAR-10), onde as redes tendem rapidamente a um teto de desempenho, o ganho proporcional sofre compressão, mas mantém-se superior ao *baseline*. Devido à trava matemática estrita no estágio $K = 1$, a técnica garante um princípio operacional de intervenção segura: a regularização não desestabiliza as projeções básicas caso a complexidade dos dados não exija subdivisões.

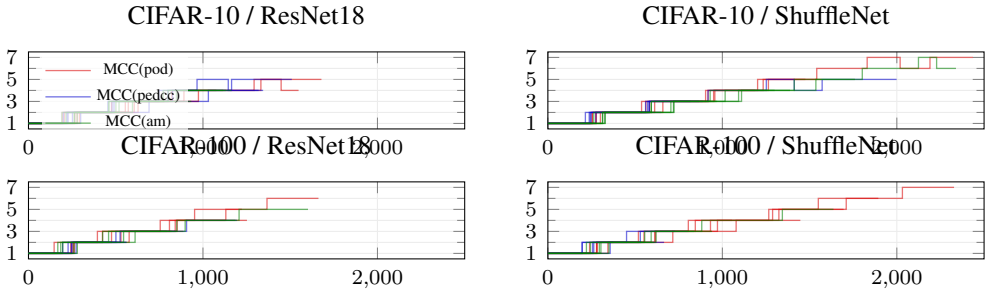


Figura 2. Trajetória de $K_{\max}$ (eixo y: maior número de sub-centroides entre as classes; eixo x: época) ao longo do treinamento, para as três variantes MCC nas quatro células (*dataset* $\times$ *backbone*). Cada cor reúne os cinco ciclos (sementes) de uma variante. O número de divisões é guiado pelos dados: o MCC(pod) tende a dividir mais (até K=7), enquanto as demais tendem a crescer menos.

A validade interna deste experimento foi calibrada pelo uso do modelo de amostras pareadas. Como os algoritmos de aceleração assíncrona (precisão mista, *autotuning* cuDNN e compiladores estáticos) induzem variações não-determinísticas na convergência, isolamos a fonte desse ruído efetuando os contrastes estritamente dentro da mesma semente algorítmica para cada configuração. Em contrapartida, é preciso limitar o alcance das conclusões. A comparação cobre os *baselines* sobre os quais o MCC foi construído, mas não inclui métodos com vários protótipos já conhecidos, como o SoftTriple [Qian et al. 2019] e o Sub-center ArcFace [Deng et al. 2020], nem propostas mais recentes; comparar diretamente com pelo menos um método de K fixo aprendido seria o passo mais útil para situar a contribuição, e fica como trabalho futuro. O custo computacional também não foi medido: não registramos o total de épocas, o tempo de treino nem a memória gasta com as épocas a mais dos reinícios. Além disso, o teste de Wilcoxon junta os quatro cenários em uma única estatística, e diferenças pareadas por cenário com intervalos de confiança completariam esse resultado; a Tabela 3 também traz médias sem o desvio entre sementes. Por fim, a Figura 2 mostra apenas $K_{\max}$, sem a distribuição final de subcentroides por classe e sem verificar se o uso entre eles é de fato balanceado; os testes também se limitam a imagens de baixa resolução.

7. Conclusão

Introduzimos o MCC (*Multi-Centroid Curriculum*), um *framework* modular capaz de flexibilizar a premissa de unimodalidade intra-classe de métodos consolidados por margens angulares e alvos fixos. Ao realizar a expansão de múltiplos protótipos de forma adaptativa via desempenho, provamos que é possível extrair melhorias robustas de redes padronizadas. Os testes demonstram que, aplicado sobre a formulação de cosseno POD, o currículo eleva a acurácia generalista em $+1{,}77$ pontos percentuais (com vitórias sistêmicas absolutas e d_z=1,57), ganhos estritamente decorrentes da gestão adaptativa da topologia latente, e não de perturbações hiperparamétricas. O desdobramento natural desta arquitetura para trabalhos futuros envolve sua aferição em hiperesferas massivas (como ImageNet e bases faciais), a elaboração de um modelo robusto de erro padrão para o pareamento global e a otimização térmica dos gatilhos do currículo.

Agradecimentos

Pesquisa desenvolvida com o apoio do Núcleo Avançado de Computação de Alto Desempenho (NACAD) da COPPE, Universidade Federal do Rio de Janeiro (UFRJ). Natália Azeredo Viana da

Silva agradece à UFRJ pela bolsa de iniciação científica (PIBIC).

Conforme o Código de Conduta para Autores da SBC, declara-se o uso da IA generativa Claude (Anthropic) como apoio na refatoração do código experimental e na escrita e revisão do texto. A concepção do método, os experimentos e as conclusões são dos autores, que revisaram todo o material e respondem pelo conteúdo do artigo.

Referências

Allen, K. R., Shelhamer, E., Shin, H., and Tenenbaum, J. B. (2019). Infinite Mixture Prototypes for Few-Shot Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 232–241. PMLR.

Deng, J., Guo, J., Liu, T., Gong, M., and Zafeiriou, S. (2020). Sub-center ArcFace: Boosting Face Recognition by Large-scale Noisy Web Faces. In *European Conference on Computer Vision (ECCV)*, pages 741–757. Springer.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Krizhevsky, A. and Hinton, G. (2009). Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 512–520.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131.

Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., and Jin, R. (2019). SoftTriple Loss: Deep Metric Learning Without Triplet Sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6450–6458.

Shen, Y., Sun, X., and Wei, X.-S. (2023). Equiangular Basis Vectors. *arXiv preprint arXiv:2303.11637*.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4077–4087.

Wang, F., Cheng, J., Liu, W., and Liu, H. (2018). Additive Margin Softmax for Face Verification. *IEEE Signal Processing Letters*, 25(6):926–930.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A Discriminative Feature Learning Approach for Deep Face Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515.

Zhu, Q., Zhang, J., Wang, R., and Zheng, J. (2020). A New Loss Function for CNN Classifier Based on Pre-Defined Evenly-Distributed Class Centroids. *IEEE Access*, 8:1084–1092.

Zhu, Q. and Zu, X. (2023). A Softmax-Free Loss Function Based on Predefined Optimal-Distribution of Latent Features for Deep Learning Classifier. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1386–1397.