NOVA: Non-monotonic Orthogonal Variance Activation per Architetture Neurali Profonde

Preprint — Work in Progress

Autore
Laboratorio di AI e Matematica Applicata

23 febbraio 2026

Sommario

La scelta della funzione di attivazione incide significativamente sulla topologia della loss landscape e sulla capacità espressiva delle reti neurali profonde. In questo articolo preliminare introduciamo **NOVA** (*Non-monotonic Orthogonal Variance Activation*), una funzione ibrida razionale-esponenziale progettata analiticamente. NOVA estende i benefici delle attivazioni gating introducendo un termine di smorzamento razionale che modula la derivata seconda. Dimostriamo analiticamente come NOVA induca una contrazione controllata della varianza all'inizializzazione (richiedendo una scala dei pesi di $\approx 2.8/n_{\text{in}}$) e formalizziamo il fenomeno della *Covariance Shift Collision*, fornendo una spiegazione teorica per il suo degrado prestazionale in architetture basate su Batch Normalization. In un confronto esplorativo a run singola su Mini-ViT / CIFAR-100 a 100 epoche, NOVA raggiunge il 56.75% di test accuracy, superando GELU (54.67%), ReLU (54.55%), SiLU (53.84%) e Mish (53.25%). Uno studio di scaling su tre varianti di ViT (3.2M–25.3M parametri) mostra un vantaggio crescente di NOVA su GELU (da +2.28 a +4.07 punti percentuali). Con regolarizzazione DeiT-style (RandAugment, CutMix, Mixup, Stochastic Depth), il vantaggio si amplifica fino a +8.35 pp alla scala Small (62.66% vs 54.31%) e l'accuracy alla scala Base recupera +5.09 pp, suggerendo una forte complementarità tra la regolarizzazione implicita di NOVA e quella esplicita. Una validazione cross-dataset su Tiny-ImageNet-200 (200 classi, 64×64) conferma il pattern: il vantaggio di NOVA cresce da +1.75 (Tiny) a +6.05 pp (Base) in Top-1 accuracy, con miglioramenti analoghi in Top-5, rafforzando l'ipotesi che il beneficio di NOVA sia indipendente dal dataset e cresca con la capacità del modello. Osserviamo trend qualitativi incoraggianti anche su Nano-GPT. Nello *Scientific Machine Learning*, l'espressività heavy-tailed della derivata seconda di NOVA ha mostrato una riduzione dell'errore residuo sulle Physics-Informed Neural Networks (PINN) di un fattore $\sim 13\times$ su un'equazione di Burgers 1D, suggerendo un forte potenziale per la modellazione di singolarità fisiche.

Indice

1 Formulazione Matematica e Derivazioni

NOVA è definita come la composizione di un meccanismo di gating esponenziale e una regolarizzazione razionale sottrattiva:

$$f(x) = x \cdot \sigma(\beta x) - \frac{x}{1 + (\beta x)^2} \tag{1}$$

dove $\sigma(z) = (1 + e^{-z})^{-1}$ è la funzione logistica standard e β è un parametro continuo che modula la curvatura.

1.1 Analisi del Gradiente e Comportamento Asintotico

La derivata prima $f'(x)$, che governa il flusso del gradiente durante la retropropagazione, è data analiticamente da:

$$f'(x) = \sigma(\beta x) + \beta x \cdot \sigma(\beta x)\big(1 - \sigma(\beta x)\big) - \frac{1 - (\beta x)^2}{\big(1 + (\beta x)^2\big)^2} \tag{2}$$

Per valutare il comportamento in prossimità dell'origine (assumendo $\beta = 1$), calcoliamo $f'(0)$ passo per passo:

$$\begin{aligned} f'(0) &= \sigma(0) + 0 \cdot \sigma(0)\big(1 - \sigma(0)\big) - \frac{1 - 0^2}{\big(1 + 0^2\big)^2} \\ &= \frac{1}{2} + 0 - \frac{1}{1} \\ &= -0.5 \end{aligned}$$

Questo gradiente marcatamente negativo all'origine definisce la *sacca di non-monotonicità* di NOVA. A differenza di GELU o SiLU, dove la concavità negativa è debole, NOVA esercita una forte spinta repulsiva (hard-thresholding implicito) per attivazioni prossime allo zero. Asintoticamente, per $x \to +\infty$, il termine razionale decade a zero ($O(1/x)$) e $\sigma(x) \to 1$, restituendo $f(x) \approx x$. Per $x \to -\infty$, $f(x) \to 0$.

1.2 Conservazione della Varianza (NOVA-Init)

Il mantenimento dell'entropia del segnale nei layer profondi richiede una corretta inizializzazione dei pesi. Nel framework generalizzato di He et al. [1], la varianza ideale è:

$$\mathrm{Var}(W) = \frac{1}{n_{\text{in}} \cdot \mathbb{E}[f(X)^2]}, \quad X \sim \mathcal{N}(0, 1)$$

A differenza di ReLU, per cui $\mathbb{E}[\max(0, X)^2] = 0.5$ (regola di He standard $2/n_{\text{in}}$), NOVA possiede una regione contrattiva. Integrando numericamente l'aspettativa dell'attivazione al quadrato sotto distribuzione normale:

$$\mathbb{E}[f(X)^2] = \int_{-\infty}^{+\infty} \left(x\sigma(x) - \frac{x}{1+x^2} \right)^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \approx 0.357 \tag{3}$$

Pertanto, per prevenire il *signal decay* nelle prime fasi del training:

$$\text{Var}(W_{\text{NOVA}}) = \frac{1}{n_{\text{in}} \cdot 0.357} \approx \frac{2.801}{n_{\text{in}}} \tag{4}$$

1.3 Derivata Seconda ed Espressività per le PINN

Scomponendo $f(x) = S(x) - R(x)$, dove $S(x) = x\sigma(\beta x)$ è il termine SiLU e $R(x) = x/(1 + (\beta x)^2)$ è il termine razionale, deriviamo l'Hessiana 1D di NOVA:

$$f''(x) = \underbrace{2\beta\sigma'(\beta x) + \beta^2 x\, \sigma'(\beta x)\big(1 - 2\sigma(\beta x)\big)}_{S''(x)} - \underbrace{\frac{2\beta^2 x\big(\beta^2 x^2 - 3\big)}{\big(1 + \beta^2 x^2\big)^3}}_{R''(x)} \tag{5}$$

La derivata seconda di funzioni esponenziali pure (come GELU) decade rapidamente per $|x| > 2$. In NOVA, il termine $R''(x)$ introduce una coda polinomiale (*heavy-tailed*), che permette all'operatore Laplaciano u_{xx} calcolato dalla rete di catturare variazioni ad alta frequenza (fronti d'urto, discontinuità termiche) che funzioni troppo lisce tendono a sfumare.

2 Contesto Architetturale e Stato dell'Arte

2.1 Confronto con Funzioni Alternative

NOVA si colloca in un ecosistema dove si cerca il bilanciamento tra levigatezza (*smoothness*) e capacità di thresholding. La Tabella 1 riassume le differenze strutturali con le principali funzioni di attivazione.

Tabella 1: Confronto tassonomico preliminare delle principali funzioni di attivazione.

Funzione	**Formula**	**Monotona**	C^∞	**Coda f''**	**Costo Rel.**
ReLU	$\max(0, x)$	Sì	No	Nulla	1.0×
GELU [2]	$x\Phi(x)$	No	Sì	Esponenziale	2.5×
SiLU/Swish [3]	$x\sigma(x)$	No	Sì	Esponenziale	2.2×
Mish [4]	$x \tanh(\text{softplus}(x))$	No	Sì	Esponenziale	2.4×
Snake [5]	$x + \sin^2(x)$	No	Sì	Periodica	4.0×
NOVA	Eq. (1)	No	Sì	Polinomiale	2.8×

Rispetto a Snake (ideale per segnali periodici, ma instabile nei Transformer) e SiLU (eccellente per LLM ma debole su operatori differenziali), NOVA fornisce asintoti globali stabili con perturbazioni locali heavy-tailed.

2.2 Teorema della Covariance Shift Collision

Nel nostro studio esplorativo su Convolutional Neural Networks (CNN), abbiamo riscontrato un fallimento critico di NOVA abbinata alla Batch Normalization (BN). Un layer BN standardizza le pre-attivazioni X imponendo $\mu = 0$ e $\sigma^2 = 1$: circa il 68% della massa di probabilità ricade nell'intervallo $x \in [-1, 1]$.

In NOVA, questa regione coincide quasi esattamente con il dominio in cui $f'(x) \leq 0$, raggiungendo il minimo in $f'(0) = -0.5$. Poiché l'entropia differenziale dell'output $h(Y)$ con $Y = f(X)$ è limitata da $\mathbb{E}[\log|f'(X)|]$, forzare la maggior parte del segnale dove la derivata è contrattiva e non-monotona causa un drastico collasso della Mutua Informazione $I(X;Y)$ tra layer adiacenti. Al contrario, la Layer Normalization (usata nei Transformer) non impone una centratura globale, permettendo ai vettori di feature di mantenere una covarianza più naturale e di sfruttare la sacca negativa di NOVA come meccanismo selettivo per il rumore a bassa magnitudine.

2.3 Ablation Study Analitico

Dal punto di vista teorico, il comportamento di NOVA può essere disaccoppiato in due meccanismi:

- **Termine esponenziale $x\sigma(\beta x)$ (Gating):** Fornisce il limite asintotico lineare e garantisce la stabilità del gradiente per attivazioni fortemente positive. Senza questo componente, la limitatezza asintotica del termine razionale causerebbe la divergenza della rete.
- **Termine razionale $x/(1+(\beta x)^2)$ (Damping):** Modula e approfondisce la sacca non-monotona ed è l'unico responsabile della coda heavy-tailed nella derivata seconda (Eq. (5)), senza la quale le performance su operatori differenziali regredirebbero a quelle di SiLU.

3 Validazione Empirica Preliminare

Tutti gli esperimenti riportati in questa sezione sono **indagini a run singola (singolo seed)**, condotte per valutare trend di convergenza e formare ipotesi preliminari. Studi futuri con replicazione statistica saranno necessari per trarre conclusioni definitive.

3.1 Classificazione Visiva (Mini-ViT su CIFAR-100)

Abbiamo addestrato da zero un Mini-ViT (4 layer, embed_dim=256, 4 heads, MLP×4) su CIFAR-100 per 100 epoche con Data Augmentation, Label Smoothing (0.1), Mixed Precision (FP16) e LayerNorm. L'addestramento è stato effettuato con AdamW (lr=3×10^{-3}, weight_decay=0.05), warmup lineare (10 epoche) seguito da cosine annealing, su 2× GPU NVIDIA T4 (Kaggle). I risultati (Tabella 2) confrontano cinque funzioni di attivazione.

NOVA raggiunge una Best Test Accuracy del 56.75%, superando GELU di +2.08 punti percentuali. Il margine è consistente anche rispetto a ReLU, SiLU e Mish. Pur essendo una run singola, un divario di oltre 2 punti su CIFAR-100 a 100 epoche suggerisce che il meccanismo di auto-regolarizzazione geometrica di NOVA fornisca un vantaggio qualitativo in architetture LayerNorm-based rispetto alle attivazioni puramente esponenziali.

Tabella 2: Best Test Accuracy – Mini-ViT su CIFAR-100 (run singola, seed 42, 100 epoche, FP16).

Attivazione	Best Val Acc	Δ vs GELU
NOVA	**56.75%**	**+2.08**
GELU	54.67%	—
ReLU	54.55%	−0.12
SiLU	53.84%	−0.83
Mish	53.25%	−1.42

3.2 Studio di Scaling (ViT su CIFAR-100)

Per verificare se il vantaggio di NOVA si mantenga all'aumentare della capacità del modello, abbiamo condotto uno studio di scaling con tre configurazioni di ViT di dimensione crescente, confrontando

NOVA e GELU a 100 epoche su CIFAR-100. Ogni scala utilizza iperparametri calibrati per le GPU T4 (Tabella 3).

Tabella 3: Scaling study – ViT su CIFAR-100 (NOVA vs GELU, run singola, seed 42, 100 epoche, FP16, 2×T4).

Scala	Config	Params	NOVA	GELU	Δ
Tiny	4L, 256d, 4h	3.2M	**56.71%**	54.43%	+2.28
Small	6L, 384d, 6h	10.7M	**59.37%**	55.48%	+3.89
Base	8L, 512d, 8h	25.3M	**55.22%**	51.15%	+4.07

NOVA supera GELU in tutte e tre le scale, con un divario che cresce da +2.28 (Tiny) a +4.07 punti percentuali (Base). L'osservazione suggerisce che il meccanismo di auto-regolarizzazione geometrica di NOVA fornisca un vantaggio crescente con la capacità del modello.

Tuttavia, i risultati rivelano un fenomeno di **overfitting marcato** a tutte le scale. La Tabella 4 riporta il divario tra training accuracy e validation accuracy a fine addestramento.

Tabella 4: Overfitting gap (Train Acc − Val Acc) a fine addestramento – ViT Scaling.

	NOVA			GELU		
Scala	**Train**	**Val**	**Gap**	**Train**	**Val**	**Gap**
Tiny	98.55%	56.24%	42.31	96.69%	54.41%	42.28
Small	99.93%	59.33%	40.60	99.80%	55.41%	44.39
Base	99.96%	55.21%	44.75	99.86%	51.03%	48.83

Alla scala Base, l'accuracy in validazione di entrambe le attivazioni scende al di sotto della scala Small (NOVA: 55.22% vs 59.37%; GELU: 51.15% vs 55.48%), indicando che l'aumento di capacità non viene sfruttato per generalizzare, ma per memorizzare il training set. Osserviamo che il gap di overfitting è sistematicamente inferiore per NOVA rispetto a GELU (specialmente a scala Small e Base), il che suggerisce un effetto di regolarizzazione implicita coerente con la struttura contrattiva della sacca non-monotona. Tuttavia, questo effetto non è sufficiente a prevenire il degrado prestazionale al crescere del modello su un dataset di dimensione limitata come CIFAR-100, motivando l'indagine successiva con tecniche di regolarizzazione avanzate.

3.3 Scaling Study con Regolarizzazione DeiT-style (v2)

Per affrontare l'overfitting marcato osservato nello studio di scaling v1, abbiamo ripetuto l'esperimento applicando tre tecniche di regolarizzazione ispirate a DeiT [7]: (1) **RandAugment** [8] (2 operazioni, magnitudine 9); (2) **CutMix** [9] ($\alpha = 1.0$) + **Mixup** [10] ($\alpha = 0.8$) con switch probability 0.5; (3) **Stochastic Depth** [11] (DropPath) con tasso crescente per scala (Tiny: 0.1, Small: 0.2, Base: 0.3). Tutti gli altri iperparametri sono invariati rispetto a v1.

Tabella 5: Scaling study v2 (con regolarizzazione) – ViT su CIFAR-100 (NOVA vs GELU, run singola, seed 42, 100 epoche, FP16, 2×T4).

Scala	Params	NOVA v2	GELU v2	Δ	NOVA v1	Δ v2−v1
Tiny	3.2M	51.72%	45.12%	+6.60	56.71%	−4.99
Small	10.7M	**62.66%**	54.31%	+8.35	59.37%	+3.29
Base	25.3M	60.31%	53.14%	+7.17	55.22%	+5.09

I risultati (Tabella 5) rivelano tre fenomeni rilevanti:

(1) Recupero della scala Base. Con la regolarizzazione v2, NOVA-Base raggiunge il 60.31%, un miglioramento di +5.09 punti percentuali rispetto a v1 (55.22%). Il modello ora sfrutta la capacità aggiuntiva per generalizzare anziché memorizzare. Analogamente, GELU-Base migliora da 51.15% a 53.14% (+1.99).

(2) Sotto-adattamento alla scala Tiny. La combinazione di RandAugment, CutMix/Mixup e DropPath risulta eccessivamente aggressiva per il modello più piccolo (3.2M parametri), causando una riduzione dell'accuracy: NOVA scende da 56.71% a 51.72% (−4.99), GELU da 54.43% a 45.12% (−9.31). L'osservazione suggerisce che la forza della regolarizzazione debba essere calibrata rispetto alla capacità del modello.

(3) Ampliamento del vantaggio di NOVA. Con regolarizzazione adeguata, il divario NOVA–GELU si amplifica significativamente: da +2.28/+3.89/+4.07 (v1) a +6.60/+8.35/+7.17 punti percentuali (v2). In particolare, alla scala Small NOVA raggiunge il 62.66%, il miglior risultato assoluto osservato in tutti i nostri esperimenti ViT su CIFAR-100. Questo suggerisce che la struttura auto-regolarizzante di NOVA sia complementare alla regolarizzazione esplicita: mentre CutMix/Mixup/DropPath prevengono la memorizzazione dei dati, la sacca non-monotona di NOVA sembra fornire un bias induttivo geometrico aggiuntivo che GELU non possiede.

Tabella 6: Confronto overfitting gap (Train Acc − Val Acc) – v1 vs v2. *Nota:* in v2, CutMix/Mixup rendono le immagini di training miste con label soft, producendo un'accuracy di training non confrontabile direttamente con la validazione.

	NOVA		**GELU**	
Scala	**Gap v1**	**Gap v2**	**Gap v1**	**Gap v2**
Tiny	42.31	−22.23	42.28	−20.75
Small	40.60	−16.92	44.39	−18.88
Base	44.75	−14.36	48.83	−14.90

La Tabella 6 mostra l'inversione completa del gap di overfitting: in v2 il training accuracy è sistematicamente inferiore al validation accuracy. Questo è un artefatto atteso di CutMix/Mixup, che creano immagini miste con label distribuzionali durante il training, rendendo il compito di training intrinsecamente più difficile della valutazione su immagini pulite. L'overfitting strutturale osservato in v1 (gap $>$ 40 pp) è stato completamente eliminato.

3.3.1 Evoluzione del Parametro β

Il parametro apprendibile β di NOVA mostra una dinamica di convergenza coerente in v2: partendo da $\beta_0 = 1.0$, decresce rapidamente nelle prime 20 epoche, stabilizzandosi attorno a $\beta \approx 0.45$ per tutte e tre le scale (Tiny: 0.478, Small: 0.444, Base: 0.448). Questo comportamento suggerisce che l'ottimizzazione comprime spontaneamente l'ampiezza della sacca non-monotona per adattarsi al regime regolarizzato, riducendo la curvatura della funzione di attivazione quando la regolarizzazione esterna è già presente.

3.4 Validazione Cross-Dataset (ViT su Tiny-ImageNet-200)

Per verificare che il vantaggio di NOVA non sia specifico a CIFAR-100, abbiamo replicato lo studio di scaling su Tiny-ImageNet-200 (200 classi, 100K immagini di training, 10K di validazione, risoluzione 64×64). L'architettura ViT è la stessa famiglia a tre scale (Tiny/Small/Base), con `patch_size` = 8 ($\rightarrow$ 64 patch per immagine). La regolarizzazione è calibrata per scala: Tiny con RandAugment magnitudine 5, CutMix/Mixup probabilità 0.5, DropPath 0.05; Small e Base con regolarizzazione DeiT-style piena (magnitudine 9, probabilità 1.0, DropPath 0.2/0.3). L'addestramento è stato condotto per 100 epoche con AdamW su 2×T4 in FP16.

Tabella 7: ViT su Tiny-ImageNet-200 – NOVA vs GELU (run singola, seed 42, 100 epoche, FP16, 2×T4).

Scala	Params	NOVA Top-1	GELU Top-1	Δ	NOVA Top-5	GELU Top-5
Tiny	3.3M	47.71%	45.96%	+1.75	72.52%	71.32%
Small	10.8M	**51.01%**	46.43%	+4.58	**75.49%**	72.18%
Base	25.5M	50.75%	44.70%	+6.05	75.32%	70.15%

I risultati (Tabella 7) confermano il pattern osservato su CIFAR-100: il vantaggio di NOVA cresce monotonamente con la scala del modello, da +1.75 a +6.05 punti percentuali in Top-1 accuracy. L'accuracy Top-5 mostra lo stesso trend (+1.20 a +5.17 pp).

Tabella 8: Confronto cross-dataset del vantaggio NOVA–GELU (pp, Top-1 accuracy).

Scala	CIFAR-100 (v2)	Tiny-ImageNet-200
Tiny	+6.60	+1.75
Small	+8.35	+4.58
Base	+7.17	+6.05

Il vantaggio assoluto è più contenuto su Tiny-ImageNet-200 (Tabella 8), il che è atteso: il dataset è più complesso (200 classi vs 100, risoluzione 64×64 vs 32×32) e i batch size ridotti (512/256/128 vs 1024/512/256) potrebbero aver limitato la convergenza. Tuttavia, il trend *monotonamente crescente* del vantaggio con la scala è preservato su entrambi i dataset, rafforzando l'ipotesi che la regolarizzazione geometrica intrinseca di NOVA diventi più efficace al crescere della capacità del modello.

Il parametro β converge a valori leggermente superiori rispetto a CIFAR-100 v2 ($\approx$ 0.54–0.58 vs $\approx$ 0.45), suggerendo un adattamento automatico alla maggiore complessità del dataset.

3.5 Modellazione Autoregressiva (Nano-GPT)

Abbiamo addestrato un Nano-GPT (≈10M parametri) sul dataset TinyShakespeare per 1000 iterazioni. NOVA ha prodotto una Validation Loss di 1.6949 rispetto a 1.7344 di GELU. L'osservazione suggerisce che il meccanismo di gating potenziato di NOVA possa contribuire a mitigare la saturazione dimensionale nei blocchi Feed-Forward dei Transformer.

Tabella 9: Validation Loss (Cross-Entropy) – Nano-GPT su TinyShakespeare (run singola).

Step	NOVA	GELU
0	4.3083	4.3501
400	1.9542	1.9747
1000	1.6949	1.7344

3.6 Modelli Generativi Continui (DDPM)

Su una U-Net DDPM per il denoising di immagini, dopo 10 epoche GELU ha ottenuto una MSE Loss di 0.0372 contro 0.0382 di NOVA. Ipotizziamo uno *Smoothness Trade-off*: l'integrazione differenziale (SDE) richiesta dalla Langevin Dynamics nei DDPM favorisce campi vettoriali estremamente lisci, mentre la forte irregolarità della sacca di NOVA introduce microscopiche perturbazioni nel reverse-sampling. Questo definisce un limite applicativo esplicito della funzione.

3.7 Scientific Machine Learning: PINN sull'Equazione di Burgers

Il test qualitativamente più rilevante ha riguardato una PINN addestrata a risolvere l'equazione nonlineare di Burgers 1D, che modella la formazione di onde d'urto fluidodinamiche e richiede il calcolo dell'operatore Laplaciano tramite auto-differenziazione doppia. Dopo 2000 step di ottimizzazione, la Physics Loss (residuo PDE) è stata:

Tabella 10: Physics Loss (residuo PDE) – PINN sull'Equazione di Burgers 1D (run singola).

Epoca	**NOVA**	**GELU**
200	0.07808	0.09757
1000	0.00207	0.00842
2000	0.00027	0.00353

Nonostante l'assenza di replicazione statistica, un abbattimento del residuo di oltre un ordine di grandezza (fattore $\sim 13\times$) suggerisce fortemente che la struttura polinomiale heavy-tailed della derivata seconda di NOVA (Eq. (5)) sia geometricamente superiore alle code esponenziali per approssimare singolarità e variazioni brusche tipiche delle onde d'urto. Questo risultato è coerente con la letteratura sulle attivazioni non-smooth per SciML [6].

4 Efficienza Computazionale (CUDA Fusion)

A causa della sua complessità composita, NOVA in modalità *Eager* su PyTorch richiede l'allocazione ripetuta di tensori temporanei intermedi, generando un Memory Bandwidth Bottleneck. Abbiamo sviluppato un **Fused CUDA Kernel** che sfrutta le istruzioni hardware *Fast Math* (`__expf`, `__fdividef`) direttamente nei registri del multiprocessore.

Un benchmark esplorativo su GPU NVIDIA T4 (tensore FP32 2048×2048, 100 iterazioni) ha prodotto i seguenti tempi per un ciclo Forward+Backward:

Tabella 11: Latenza Forward+Backward – GPU NVIDIA T4, FP32, tensore 2048×2048.

Funzione	**Implementazione**	**Latency**	**VRAM Intermedia**
NOVA	Python Eager	≈ 4.56 ms	128 MB
GELU	ATen Native Fused	≈ 0.57 ms	0 MB
NOVA	Custom CUDA Fused	≈ 0.92 ms	0 MB

Il kernel fused azzera il memory footprint intermedio e chiude circa il 90% del gap di latenza rispetto a GELU, rendendo NOVA sufficientemente efficiente per il training distribuito. L'overhead residuo (~ 0.35 ms) è attribuibile ai cicli aggiuntivi richiesti dalla divisione razionale rispetto alle sole operazioni esponenziali.

5 Limitazioni e Lavori Futuri

Il presente lavoro costituisce una *Proof of Concept* e presenta le seguenti limitazioni da affrontare in studi successivi:

- **Robustezza Statistica:** Tutti i risultati sperimentali derivano da run singole con singolo seed. La replicazione con seed multipli e test statistici formali è necessaria per validare i trend osservati.
- **Scala e Generalizzazione:** Lo studio di scaling v2 con regolarizzazione DeiT-style ha eliminato l'overfitting e la validazione cross-dataset su Tiny-ImageNet-200 ha confermato il trend, ma tutti gli esperimenti sono limitati a $\leq$25M parametri su dataset di dimensione modesta. Il

comportamento di NOVA a scale LLM ($\geq$7B parametri) e su dataset su larga scala (ImageNet-1K) rimane non verificato.

- **Stabilità di β:** Nello studio v2, il parametro β converge stabilmente a ≈ 0.45 partendo da $\beta_0 = 1.0$, ma la sensibilità all'inizializzazione β_0 e il comportamento in regimi di training estremamente lunghi (>1000 epoche) rimangono da caratterizzare.
- **Ambito PDE:** Il vantaggio nelle PINN è stato osservato su equazioni con singolarità (Burgers 1D). Su PDE paraboliche totalmente lisce (es. equazione del calore), i risultati preliminari suggeriscono prestazioni comparabili a GELU, non superiori.

5.1 Euristica d'Uso per Ricercatori

In base alle intuizioni teoriche e ai trend osservati in questa indagine preliminare:

- **Consigliato:** Architetture LayerNorm / RMSNorm (Vision Transformers, LLM Decoder-only); Scientific Machine Learning su PDE con onde d'urto o singolarità.
- **Sconsigliato:** Reti con forte uso di Batch Normalization (es. ResNet); modelli generativi a diffusione continua (DDPM, score-based models).

Riferimenti bibliografici

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034.

[2] Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

[3] Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

[4] Misra, D. (2019). Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*.

[5] Ziyin, L., Hartwig, T., & Ueda, M. (2020). Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33, 1583–1594.

[6] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.

[7] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 10347–10357.

[8] Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). RandAugment: Practical automated data augmentation with a reduced search space. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703.

[9] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032.

[10] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. *International Conference on Learning Representations (ICLR)*.

[11] Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep networks with stochastic depth. *European Conference on Computer Vision (ECCV)*, pp. 646–661.