Multi-Model Merging via Spherical Barycenters (SLERP)

August 14, 2025

Girolamo Politanò Chiara Andreoni

Abstract

Il progetto prevede di applicare il Multi-Model Merging con baricentro sferico su modelli di reti neurali per trovare una media pesata di n modelli nello spazio non euclideo della (iper)sfera. I risultati vengono confrontati su due architetture, un semplice MLP (Multi-Layer Perceptron) e una regressione logistica, addestrati sul dataset MNIST per la classificazione di immagini.

1. Introduzione

Il termine SLERP (Spherical Linear Interpolation) si riferisce alla tecnica di interpolazione sulla superficie di una sfera unitaria. A differenza dell'interpolazione lineare, i pesi dei modelli vengono trattati come vettori in uno spazio n-dimensionale e l'interpolazione viene eseguita sulla superficie dell'ipersfera unitaria.

Nel progetto si estende la tecnica SLERP al caso di $n > 2$ modelli. L'approccio adottato è l'implementazione di un **baricentro sferico**, che generalizza l'idea della media pesata sull'ipersfera. È importante garantire la **coerenza ciclica**, ovvero che il risultato finale del merging sia unico, indipendentemente dall'ordine in cui i modelli vengono combinati.

I principali obiettivi di questo progetto sono:

- Implementare un metodo per il calcolo del baricentro sferico di n modelli;
- Valutare le prestazioni del modello unito su una task di classificazione sul dataset MNIST;
- Verificare la coerenza ciclica;
- Utilizzo del Riemannian Trust-Region, per il calcolo del baricentro.

Email: Chiara Andreoni, Girolamo Politanò <politano.2005893@studenti.uniroma1.it, andreoni.2087122@studenti.uniroma1.it>.

Machine Learning 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

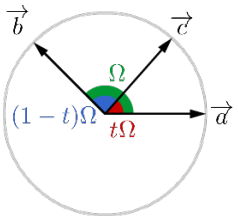

Figure 1. L'immagine illustra come Slerp non interpola linearmente i vettori nello spazio cartesiano, ma interpola linearmente l'angolo tra di essi. Questa è la caratteristica che distingue slerp rispetto dall'interpolazione lineare (lerp), che si muove in linea retta tra i due punti, non seguendo la superficie sferica.

2. Elementi di teoria

L'interpolazione SLERP tra due vettori unitari v_0 e v_1 è definita come:

$$\text{SLERP}(v_0, v_1, t) = \frac{\sin((1-t)\omega)}{\sin(\omega)} v_0 + \frac{\sin(t\omega)}{\sin(\omega)} v_1$$

dove $\omega = \arccos(v_0 \cdot v_1)$ è l'angolo geodetico tra i due vettori.

Se l'angolo ω della distanza geodetica è molto piccolo, possiamo utilizzare la formula dell'interpolazione lineare per calcolare la distanza tra i due punti:

$$LERP(v_0, v_1, t) = (1-t)v_0 + tv_1$$

Il concetto di **baricentro sferico** generalizza questa idea a un insieme di n vettori. In questo modo otteniamo un punto b sull'ipersfera che minimizza la somma pesata delle distanze geodetiche dagli altri vettori, in modo iterativo. Un vantaggio importante di questo approccio è che risolve intrinsecamente il problema della coerenza ciclica.

3. Dimostrazione della Coerenza Ciclica

La dimostrazione si basa sulla proprietà associativa dell'operazione di baricentro, che garantisce che il risultato finale di una fusione di più modelli sia indipendente dall'ordine delle fusioni a coppie.

Il baricentro sferico di N vettori unitari $\mathbf{v}_1, \ldots, \mathbf{v}_N$ con pesi $w_1, \ldots, w_N$ è definito come il vettore $\mathbf{v}_{\text{baricentro}}$ che minimizza la somma delle distanze geodetiche:

$$\mathbf{v}_{\text{baricentro}} = \underset{\mathbf{v} \in S^{n-1}}{\operatorname{argmin}} \sum_{i=1}^{N} w_i \cdot \text{dist}_{\text{angolare}}(\mathbf{v}, \mathbf{v}_i)$$

dove $\sum_{i=1}^{N} w_i = 1$.

La coerenza ciclica è una conseguenza diretta della proprietà associativa dell'operazione di baricentro. Per dimostrarla, si possono confrontare due percorsi di merging per tre vettori $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$.

- Percorso A:

  Calcoliamo il baricentro di $\mathbf{v}_1$ e $\mathbf{v}_2$ con pesi w_1 e w_2, ottenendo un vettore intermedio $\mathbf{v}_{12}$. Poi calcoliamo il baricentro di $\mathbf{v}_{12}$ e $\mathbf{v}_3$ con pesi $w_{12} = w_1 + w_2$ e w_3. Il risultato finale $\mathbf{v}_A$ minimizza la funzione costo:

  $$\text{Costo}_A = (w_1 + w_2) \cdot \text{dist}(\mathbf{v}, \mathbf{v}_{12}) + w_3 \cdot \text{dist}(\mathbf{v}, \mathbf{v}_3)$$

- Percorso B:

  Calcoliamo il baricentro di $\mathbf{v}_2$ e $\mathbf{v}_3$ con pesi w_2 e w_3, ottenendo un vettore intermedio $\mathbf{v}_{23}$ e poi calcoliamo il baricentro di $\mathbf{v}_{23}$ e $\mathbf{v}_1$ con pesi $w_{23} = w_2 + w_3$ e w_1. Il risultato finale $\mathbf{v}_B$ minimizza la funzione costo:

  $$\text{Costo}_B = (w_2 + w_3) \cdot \text{dist}(\mathbf{v}, \mathbf{v}_{23}) + w_1 \cdot \text{dist}(\mathbf{v}, \mathbf{v}_1)$$

Entrambi i percorsi di fusione portano alla minimizzazione della stessa funzione di costo, che può essere scritta come:

$$(\mathbf{v}, \mathbf{v}_1) + w_2 \cdot \text{dist}(\mathbf{v}, \mathbf{v}_2) + w_3 \cdot \text{dist}(\mathbf{v}, \mathbf{v}_3)$$

Poiché il punto di minimo di questa funzione è unico, si ha che $\mathbf{v}_A = \mathbf{v}_B$.

Questo dimostra che il risultato finale è indipendente dall'ordine delle fusioni, provando la coerenza ciclica del baricentro sferico.

4. Metodologia

Il progetto utilizza il dataset MNIST per un task di classificazione delle immagini. Per la sperimentazione sono state utilizzate due architetture di rete neurale:

- **SimpleMLP**: Una rete neurale semplice multi-strato con tre strati lineari e funzioni di attivazione ReLU. Il primo strato prende in input l'immagine 28x28 pixel e dà in output 128 neuroni.Il secondo layer prende in input i 128 neuroni e li riduce a 64. Infine, l'ultimo strato trasforma i 64 neuroni nei 10 parametri del modello.
- **LogisticRegression**: Un modello di regressione logistica, considerato un caso più semplice e lineare per il merging. Nonostante la regressione logistica non sia propriamente una rete neurale, essa può essere vista come una sua versione semplificata. Questa rete è composta di un solo strato che prende in input le immagini del dataset e resistuisce il vettore dei parametri.

Per ciascuna architettura abbiamo addestrato quattro modelli indipendenti per tre epoche utilizzando l'ottimizzatore Adam con un learning rate di 0.001 e la funzione di loss CrossEntropy. Il dataset MNIST è stato caricato e suddiviso in un set di addestramento (50000 immagini) e un set di validazione (10000 immagini).

Siamo poi passati al processo di merging:

1. Flattening dei pesi: i pesi di ciascun modello sono stati estratti, appiattiti in un singolo vettore 1D.
2. Normalizzazione: ogni vettore è stato normalizzato.
3. Calcolo del baricentro sferico.
4. Impostazione dei pesi: I pesi del baricentro risultanti sono stati utilizzati per inizializzare un nuovo modello, la cui accuratezza è stata valutata sul validation set.

In seguito, abbiamo anche implementato un ottimizzatore più avanzato, il **Riemannian Trust-Region**, per il calcolo del baricentro sferico, per studiare e verificare possibili differenze e miglioramenti sul modello.

5. Risultati Sperimentali

I risultati ottenuti dal merging differiscono significativamente tra le due architetture.

5.1. Merging su MLP

I quattro modelli MLP individuali hanno raggiunto le seguenti accuratezze:

- Modello 1: 96.33%
- Modello 2: 96.13%
- Modello 3: 96.77%
- Modello 4: 96.58%

Il modello MLP unito, calcolato tramite il baricentro sferico, ha mostrato un'accuratezza estremamente bassa: **Baricentro sferico**: 9.12%

Questo risultato evidenzia un fallimento del merging, probabilmente dovuto alla natura non convessa del landscape della loss per architetture più complesse, dove

l'interpolazione sulla ipersfera può portare a regioni di bassa performance.

Il risultato del baricentro sferico calcolato con l'ottimizzatore Riemannian Trust-Region ha dato un'accuratezza simile: **Baricentro sferico MLP (Riemannian Trust-Region):** 8.60%

Quindi possiamo concludere che non è l'ottimizzatore che determina la bassa accuratezza del modello con merging. La performance pessima del modello con merging è dovuta alla struttura di MLP. Due MLP addestrati separatamente possono avere neuroni "equivalenti" ma in posizioni diverse, quindi se sommati o mediati, i pesi, senza essere stati allineati prima, distruggono le rappresentazioni interne. Questa differenziazione accade perchè MLP può avere distribuzioni di pesi e bias molto diverse a seconda dell'inizializzazione, learning rate, e normalizzazione. La fusione lineare di modelli con scale diverse rompe l'equilibrio delle attivazioni.

5.2. Merging su Regressione Logistica

I quattro modelli di regressione logistica hanno raggiunto le seguenti accuratezze:

- Modello 1: 91.51%
- Modello 2: 91.45%
- Modello 3: 91.18%
- Modello 4: 91.52%

In questo caso, il modello unito ha mantenuto un'ottima performance, con un'accuratezza paragonabile a quella dei modelli individuali: **Baricentro sferico Regressione Logistica:** 91.46%

Questo suggerisce che il merging tramite baricentro sferico è efficace per architetture con un landscape di loss più convesso. Poichè la regressione logistica ha uno strato con mapping diretto delle feature, non abbiamo nessun problema di permutazione come avevamo in MLP.

5.3. Test di Coerenza Ciclica

Per dimostrare la coerenza ciclica, è stato eseguito un test di merging su quattro modelli di regressione logistica, combinandoli in due ordini diversi. Abbiamo scelto una soglia di tolleranza per valutare la distanza tra i due vettori ottenuti con ordini di merge diversi.

La distanza tra i vettori è minore della tolleranza di $1e-6$ utilizzata, indicando che il test conferma e dimostra la proprietà del baricentro sferico di coerenza ciclica.

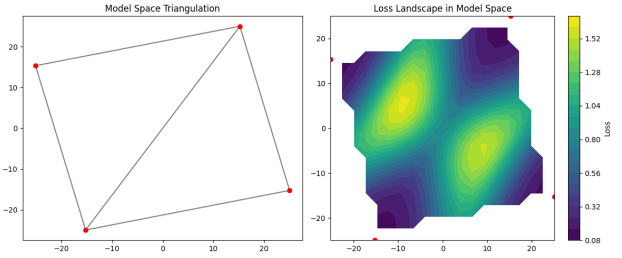

Figure 2. Loss landscape MLP

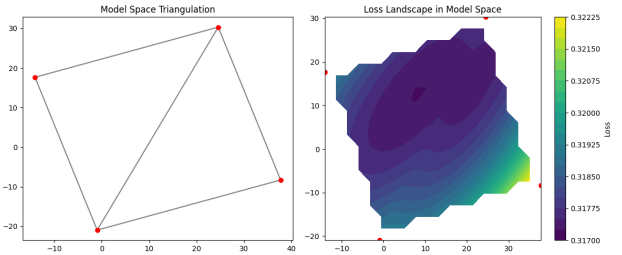

Figure 3. Loss landscape regressione logistica

5.4. Analisi della loss landscape

Abbiamo generato dei grafici bidimensionali sulla loss dei modelli, per evidenziare se ci fossero delle aree in cui fosse indicativamente più alta.

Gli assi X e Y rappresentano le due componenti principali ottenute dalla riduzione di dimensionalità dei vettori di peso dei modelli addestrati. Ogni punto in questo spazio 2D corrisponde a una combinazione di pesi del modello.

I punti rossi indicano le posizioni, nello spazio 2D ridotto, dei modelli individuali che abbiamo addestrato.

Nel grafico della loss landscape di MLP, si può notare la presenza di due aree in cui la loss è molto più alta del valore medio, questo evidenzia che il modello unito è instabile nelle aree intermedie tra i modelli che lo compongono. Invece, per quanto riguarda il grafico della regressione logistica, non ci sono evidenti picchi di loss, questo perchè il modello è molto più stabile quando vi si applica il merging.

6. Conclusioni

Il progetto dimostra che il concetto di baricentro sferico può essere applicato con successo per unire più modelli di regressione logistica. Al contrario, il merging su un MLP ha portato a un crollo totale della performance. Questo risultato suggerisce che la topologia dello spazio dei pesi dei modelli più complessi è altamente non-convessa e che una semplice interpolazione geodetica potrebbe non essere sufficiente per trovare una regione di minima promettente.

Bibliografia

References

[1] https://www.coinfeeds.ai/ai-blog/slerp-model-merging-primer

[2] Shoemake, K. (1985). ”Animating rotation with quaternion curves.” In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*.

[3] Crisostomi, D., Fumero, M., Baieri, D., Bernard, F., Rodola, E. (2024). ”C 2 M 3 : Cycle-Consistent Multi-Model Merging.” *Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[4] Arias-Castro, E., Donoho, D. L. (2007). ”Optimal estimation of the spherical barycenter.” *Electronic Journal of Statistics* -