Rebasin ortogonale:una nuova prospettiva sulla connettività lineare tra modelli neurali

September 9, 2025

Valeria d'Orsi

1. Introduzione

La ricerca sull'allineamento e la fusione di reti neurali è attualmente in fase di sviluppo e riveste un ruolo cruciale in molteplici contesti applicativi.I metodi proposti in tale ambito mostrano il loro limite nell'esclusivo uso di permutazioni:per superare tale vincolo, si propongono tecniche basate su **matrici ortogonali generali** che, pur non assicurando una corrispondenza funzionale esatta dopo l'allineamento, risultano efficaci, con un errore riconducibile a non linearità modeste e tollerabili.

https://github.com/valedorsi-sapienza/04_Rebasin_ortogonale.git

2. Related work

Estendendo l'intuizione di Entezari (2021), Ainsworth et al. nell'articolo *Git Re-Basin* hanno dimostrato empiricamente che reti neurali con identica architettura, addestrate in modo indipendente, possono essere connesse tramite interpolazione lineare a bassa perdita, previo un opportuno riallineamento mediante l'applicazione di permutazioni tra le unità dei modelli in esame.(Ainsworth, 2023)

3. Metodi

3.1. Baselines

Ainsworth et al (*Git Re-Basin*) propongono due tecniche di allineamento mediante permutazioni che costituiscono le fondamenta dello sviluppo di Re-basin orthogonale.Si tratta dell' *Activation Matching*, in cui si cerca la matrice di permutazione **minimizzando la distanza tra le attivazioni** dei neuroni e risolvendo un *Linear Assignment Problem* , e il *Weight Matching*,in cui si ricava la permutazione ottimale **minimizzando la distanza tra i pesi** corrispondenti dei modelli, adoperando un algoritmo *greedy* basato su *coordinate descent*.

Email: Valeria d'Orsi <dorsi.2136945@studenti.uniroma1.it>.

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

3.2. Contributo

3.2.1. ANALISI DI PROCRUSTE

La ricerca di una trasformazione ortogonale che allinei le rappresentazioni interne di due reti neurali può essere formalizzata come un problema di *Procruste ortogonale*,in cui data una coppia di matrici $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d\times n}$, si cerca la migliore matrice ortogonale R^* che le allinei, minimizzandone la distanza in norma di Frobenius:

$$R^* = \min_{\mathbf{R}\in\mathbb{R}^{d\times d}} \|\mathbf{X} - \mathbf{R}\mathbf{Y}\|_F \quad \text{t.c.} \quad \mathbf{R}^\top\mathbf{R} = \mathbf{I}$$

La soluzione ottimale(*Schonemann1964*) è data dalla *decomposizione ai valori singolari* (SVD) del prodotto $\mathbf{X}\mathbf{Y}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ da cui si ricava $\mathbf{R}^* = \mathbf{U}\mathbf{V}^\top$

Si applica tale formulazione a diversi livelli della rete neurale. A livello delle attivazioni si cerca, per ogni layer l, una matrice di rotazione $\mathbf{R}_l \in \mathbb{R}^{d\times d}$ che allinei le attivazioni della rete B a quelle della rete A, risolvendo:

$$\min_{\mathbf{R}_l\in\mathbb{R}^{d\times d}} \left\|\mathbf{Z}_l^A - \mathbf{R}_l\mathbf{Z}_l^B\right\|_F \quad \text{t.c.} \quad \mathbf{R}_l^\top\mathbf{R}_l = \mathbf{I}$$

Al livello dei pesi si formula il problema come segue:

$$\min_{R_l} \left\|W_l^A - R_l W_l^B R_{l-1}^\top\right\|_F \quad \text{t.c.} \quad \mathbf{R}_l^\top\mathbf{R}_l = \mathbf{I}$$

considerando che i pesi di ciascun layer l sono influenzati da una trasformazione ortogonale $R_{l-1}^\top$, indotta dalla compensazione della rotazione applicata al layer precedente.

Si propone inoltre un'ulteriore tecnica in cui si applica lo step centrale del problema di Procruste alla matrice di correlazione lineare aggregata A, calcolata per il layer l a partire dalle matrici dei pesi corrispondenti dello stesso layer. L'obiettivo è cercare:

$$R^* = \arg\max_{T\in\mathcal{O}(n)} \mathrm{Tr}(R^\top A)$$

3.2.2. OTTIMIZZAZIONE SULLA VARIETA' DI STIEFEL

Le trasformazioni ortogonali appartengono alla *varietà di Stiefel* ($\mathrm{St}(n,p)$) definita come l'insieme delle ma-

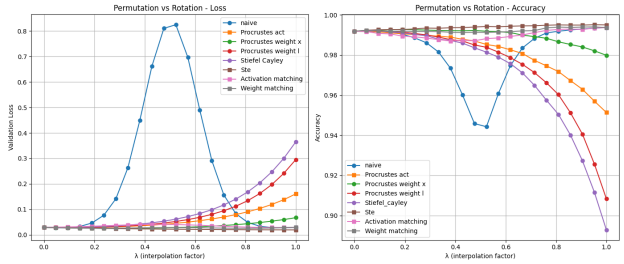

Figure 1. Permutation-based methods vs *Rotation-based methods*. Confronto tramite Validation Loss e Accuracy tra le interpolazioni lineari tra il modello A e il modello $B_{allineato}$ allineato tramite le diverse strategie.

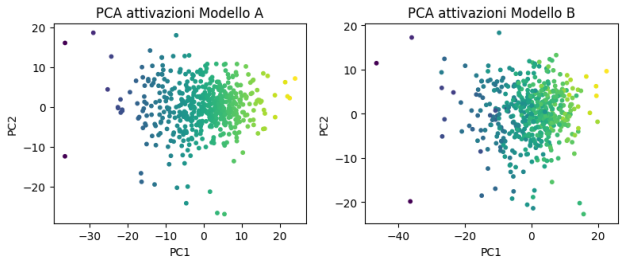

Figure 2. Confronto delle matrici di attivazione di due modelli allineati tramite Procruste mediante l'analisi delle componenti principali (PCA) su due dimensioni. L'immagine mostra la PCA delle attivazioni del layer1.

trici $X \in \mathbb{R}^{n\times p}$ le cui colonne sono ortonormali, cioè $\mathrm{St}(n,p) = \left\{X \in \mathbb{R}^{n\times p} \;\middle|\; X^\top X = I_p\right\}$

Nel contesto dell'allineamento tra le matrici dei pesi di due reti neurali, la ricerca della migliore trasformazione ortogonale si può esprimere come un problema di ottimizzazione vincolata su varietà di Stiefel:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p}} F(\mathbf{X}) \quad \text{t.c.} \quad \mathbf{X}^\top\mathbf{X} = \mathbf{I} \tag{1}$$

dove la funzione obiettivo $F(\mathbf{X})$ con $\mathbf{X} \in \mathrm{St}(d,d)$, è definita come $F(\mathbf{X}) = \left\|\mathbf{W}_l^A - \mathbf{X}_l\mathbf{W}_l^B\mathbf{X}_{l-1}^\top\right\|_F^2$, cioè la distanza tra i pesi del layer l della rete A e della rete B, tenendo conto della rotazione $X_{l-1}^\top$ applicata per compensare quelle precedenti.

L'ottimizzazione viene condotta tramite metodo iterativo basato sulla trasformazione di Cayley: a ogni iterazione, si proietta il gradiente sul piano tangente della varietà e si utilizza per definire una curva di ricerca, lungo la quale viene effettuata una discesa curvilinea. (Tagare, 2011).

4. Risultati sperimentali

4.1. Interpolazione

Si interpolano i modelli A e $B_{allineato}$ mediante LERP e SLERP. In entrambi i casi si osserva una buona linear connectivity, evidenziata soprattutto dai valori della validation loss generalmenti bassi (Figure 1) e dai ridotti valori della loss barrier, di seguito riportati.

Loss Barrier	
Naive	0.846
Stiefel	0.146
Procrustes activation	0.143
Procrustes weight l	0.129
Procrustes weight x	0.017

4.2. Errore di disallineamento

L'errore di disallineamento tra le attivazioni

$$E_{\text{all}}^{(l)} = \mathbb{E}_x\left[\left\|T\left(h_l^{(1)}(x)\right) - h_l^{(2)}(x)\right\|_2^2\right]$$

risulta tollerabile con valori il cui ordine di grandezza varia da 10^{-2} fino a 10^{-1}, raggiungendo ordine 1 solo in corrispondenda dell'ultimo layer. Si tratta di un buon risultato, che si evince anche graficamente. (Figure 2)

4.3. Consistenza ciclica

Def: Date due reti neurali f_A e f_B, e due trasformazioni $T_{A\to B}$ e $T_{B\to A}$ che mappano rispettivamente il modello f_A nello spazio di f_B e viceversa, si definisce *consistenza ciclica* la proprietà per cui:

$$T_{B\to A}(T_{A\to B}(f_A)) \approx f_A \quad \text{e} \quad T_{A\to B}(T_{B\to A}(f_B)) \approx f_B.$$

Def: Errore ciclico: $E_{\text{ciclico}} = \|T_{A\to B}T_{B\to A} - I\|_F$

La consistenza ciclica risulta pienamente soddisfatta quando l'allineamento avviene tramite permutazioni, tuttavia non è sempre perfettamente garantita quando si adoperano rotazioni ortogonali.I valori dell'errore ciclico oscillano a seconda della tecnica di allineamento adoperata come emerge dagli ordini di grandezza dell'errore.

Ordine errore ciclico	
Activation matching/Weight matching	0.0
Stiefel	10^1
Procrustes activation	10^1
Procrustes weight	10^{-3}

5. Conclusione

Come si evince dalla Figure 1 i risultati ottenuti a livello di interpolazione sono ottimistici, tuttavia sarebbe opportuno individuare altre strategie di allineamento che possano garantire una consistenza ciclica maggiore.

Bibliography. (Ainsworth, 2023) (Tagare, 2011) (Kornblith et al., 2019) (Akira Ito1, 2025)

References

Ainsworth, Hayase, S. Git re basin:merging models modulo permutation symmetries, 2023.

Akira Ito1, M. Y. . A. K. Analysis of linear mode connectivity via permutation-based weight matching:with insights into other permutation search methods, 2025. https://arxiv.org/pdf/2402.04051.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited, 09–15 Jun 2019. https://colab.research.google.com/github/google-research/google-research/blob/master/representation_similarity/Demo.ipynb.

Tagare. Notes on optimization on stiefel manifolds, 2011. https://cseweb.ucsd.edu/classes/sp24/cse291-e/papers/StiefelManifold/StiefelNotes.pdf.