Spectrum-Guided Model Merging: Focused Fusion on Informative Layers

January 14, 2026

Mariastella Gioia La Rocca   **Mia Rodotà**

Sapienza University of Rome

Abstract

This study investigates the efficacy of Spectral Signal-to-Noise Ratio (SNR), derived from Random Matrix Theory (RMT), as a precise metric for merging deep learning models. By analyzing the eigenvalue distributions of weight matrices through the Marchenko-Pastur Law, we distinguish between informational ”signal” and random ”noise.” We compare traditional weight averaging against advanced spectral strategies—including ”Spectrum Base Reset” and ”SNR Winner Take All”—across BERT-based Transformers (NLP) and ResNet-18 (Vision). Our findings reveal that while spectral methods provide significant gains in structural NLP tasks, they exhibit high sensitivity and potential instability in convolutional architectures like ResNet.

1. Introduction

Model merging aims to create a multi-task expert by combining pre-trained weights without additional costly training. However, a ”Standard Merge” (simple arithmetic mean) often results in task interference, where the noise of one expert dilutes the critical signal of another. This report evaluates whether Spectral SNR can effectively act as a ”routing” mechanism to preserve high-information layers while discarding noisy or irrelevant parameters.

2. Methodology

2.1. Spectral SNR Calculation

1. **Covariance Matrix:** For a weight matrix W (reshaped to 2D if necessary), we compute the covariance matrix

$$C = \frac{1}{M} W^\top W \quad (\text{or } C = \frac{1}{N} W W^\top),$$

where M is the number of features.

2. **Eigenvalue Decomposition:** We extract the eigenvalues $\{\lambda_i\}$ of C. According to Random Matrix Theory (RMT), if a matrix contains only random noise, its eigenvalues remain within a *bulk* defined by a theoretical upper bound λ_+.

3. **Signal vs. Noise:** Eigenvalues satisfying

$$\lambda > \lambda_+$$

are interpreted as *signal* (components the model has effectively learned), whereas eigenvalues

$$\lambda \leq \lambda_+$$

correspond to *noise* (random fluctuations or uninformative weights).

4. **Signal-to-Noise Ratio (SNR):** In our NLP experiments, the SNR is defined as

$$\text{SNR} = \frac{\sum_{\lambda_i > \lambda_+} \lambda_i}{\sum_{\lambda_i \leq \lambda_+} \lambda_i}.$$

In the ResNet experiment, we adopt a simplified definition:

$$\text{SNR} = \frac{\lambda_{\max}}{\lambda_{\text{median}}},$$

where $\lambda_{\max}$ represents the peak signal and λ_{median} characterizes the noise bulk.

Email: Mariastella Gioia La Rocca <larocca.2115586@studenti.uniroma1.it>, Mia Rodotà <rodota.2148380@studenti.uniroma1.it>.

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

2.2. Merging Strategy Logic

Standard Merge: A simple arithmetic average ($0.5A + 0.5B$). It treats all layers as equally important.

Spectrum Base Reset: High-SNR layers are averaged. Low-SNR layers (noise) are reset to the original pre-trained weights of the base model. The goal is to "clean" the model by removing noisy fine-tuned weights.

SNR Winner (Top) / Mean (Low): For high-SNR layers, we keep the weights of the model with the higher SNR (the "winner"). For low-SNR layers, we fall back to a standard average.

`snr_lowA` / `snr_lowB`: If the combined signal is above the threshold, we average. If it is below (low signal), we replace the layer with the weights of Model A (or B) exclusively. This tests if "forcing" one expert's architecture in noisy regions helps stability.

`snr_max`: If the signal is low, instead of averaging, we take the weights of the model that has the higher SNR for that specific layer.

SNR Weighted Dynamic: Instead of a 50/50 split, each layer is merged using a weighted average where the weights are proportional to the relative SNR of each model (e.g., if $\text{SNR}_A > \text{SNR}_B$, Model A contributes more to that layer).

Table 1. Merging Strategies Logic.

Strategy	High SNR	Low SNR	Goal
Standard	Avg	Avg	Baseline
Reset	Avg	Base Init	Denoise
SNR Win	Winner	Avg	Specialize
snr_lowA	Avg	Force A	Stabilize
snr_max	Avg	Winner	Max Signal
Weighted	α-Mix	Avg	Soft Balance

3. Implementation

We tested three distinct experimental setups:

1. **NLP (SST-2 & Emotion):** Merging a Sentiment Analysis expert with an Emotion expert. The dataset for Emotion has 6 classes, unlike SST-2 that has two. While the "body" of the model was merged, the "head" (classifier) had to be re-initialized, introducing random noise. Despite this the merging worked well, but we also wanted to implement a mathematically "perfect" merge.
2. **NLP (SST-2 & CoLA):** Merging Sentiment with Linguistic Acceptability (Grammar). After the first setup we switched to CoLA because both SST-2 and CoLA are natively binary (2 classes), and therefore all layers, including the classifier, could be merged without re-initialization.
3. **Vision (FashionMNIST & KMNIST):** Merging two ResNet-18 experts specialized in different image domains.

4. Results and Analysis

4.1. Comparative Analysis of NLP Tasks

The behavior of spectral merging varies significantly based on the task pair:

- **SST-2 & Emotion:** The SNR Winner strategy achieved superior performance over the Standard Merge. For SST-2, the accuracy reached $\sim$ 0.885 at the Top 25% threshold, while Emotion reached $\sim$ 0.705 at the Top 70% threshold. This suggests that for semantically related tasks, choosing the "stronger" signal preserves specialized knowledge better than averaging. *See Figure 1*
- **SST-2 & CoLA:** Results were more divergent. For CoLA, the SNR Winner Strategy significantly outperformed the baseline, reaching $\sim$ 0.685. However, for SST-2, all spectral methods fell below the Standard Merge baseline ($\sim$ 0.695). This indicates that Sentiment Analysis (SST-2) may rely on distributed, lower-SNR information that is inadvertently discarded during selective merging, whereas Grammar (CoLA) is more "peak-dependent." *See Figure 2*

4.2. Vision Domain: ResNet-18 (FashionMNIST KMNIST)

The results for ResNet-18 reveal significant instability compared to Transformers (*See Figure 3)*:

- **Expert vs. Non-Expert Performance:** In Task C (FashionMNIST), the `snr_lowA` strategy performed best at higher thresholds, nearing the expert accuracy of $\sim$ 0.91. Conversely, `snr_lowB` caused accuracy to crash toward the non-expert baseline ($\sim$ 0.15) as the threshold increased.
- **Strategy Instability:** In Task D (KMNIST), most spectral strategies (`max`, `weighted_dynamic`) performed poorly, staying near or below the standard merge ($\sim$ 0.31) and far below the expert baseline ($\sim$ 0.98).

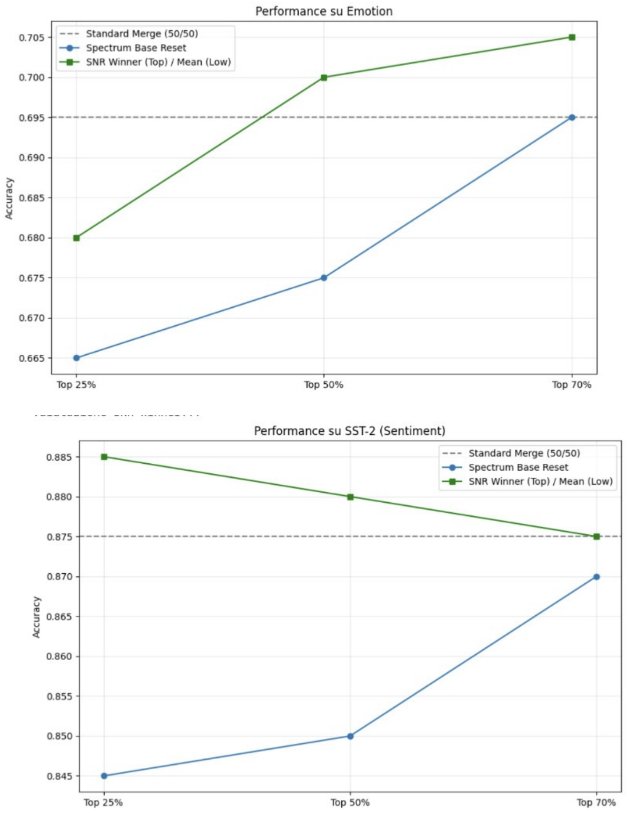

Figure 1. **NLP Task Analysis.** *Top:* Performance on Emotion. The *SNR Winner* strategy (green line) shows a progressive improvement, surpassing the *Standard Merge* baseline (dashed line) as the retained spectrum increases (50%-70%). *Bottom:* Performance on SST-2 (Sentiment). Here, the *SNR Winner* consistently outperforms the baseline across all percentages, highlighting robustness in sentiment tasks, while *Spectrum Base Reset* (blue line) requires higher retention rates to approach baseline performance.

- **Analysis:** Convolutional layers are highly sensitive to weight discontinuities. Selective merging in ResNet often breaks the hierarchical feature extraction pipeline, leading to “all-or-nothing” performance based on whether a specific critical layer was preserved or averaged.

4.3. Deep Dive into Merging Strategies

- **Standard Merge:** Provides a consistent, albeit mediocre, baseline. It is resilient but suffers from signal dilution.
- **SNR Winner / SNR Max:** These are high-reward but high-risk. They work best when task features are concentrated in specific layers (like CoLA) but can fail if the “winner” for one task is the “noise” for another.
- **Spectrum Base Reset:** Shows a steady upward trend in Transformers as more layers are included (Top 70%), but it is often too aggressive, discarding “quiet” but essential parameters.

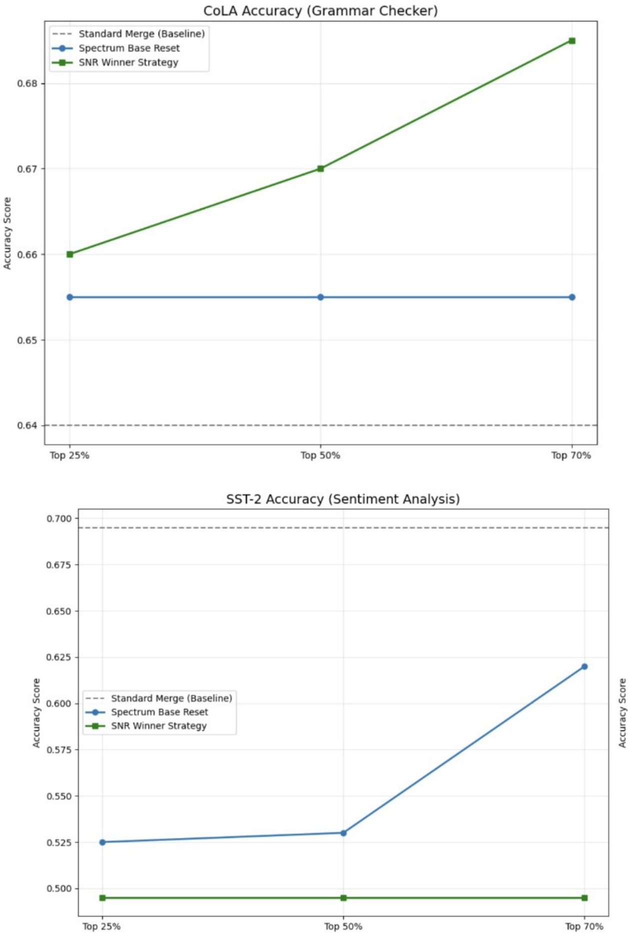

Figure 2. **NLP Task Analysis: CoLA vs. SST-2.** *Top:* On CoLA (Grammar), the *SNR Winner* strategy (orange bar) outperforms Standard Merge, proving that structural rules are localized in high-SNR weights. *Bottom:* On SST-2 (Sentiment), spectral strategies fail to beat the Standard Merge baseline (blue bar), indicating that sentiment information is distributed and damaged by aggressive weight selection.

- **SNR Weighted Dynamic:** Attempted to balance both signals but often reverted to the mean, showing limited gains over standard merging in the ResNet experiments.

5. Conclusions

Spectral SNR merging proves to be a powerful tool for Transformer-based architectures, particularly for structural tasks like grammar checking where information is concentrated. However, its application in Convolutional Neural Networks (ResNet) is currently limited by the high sensitivity of convolutional hierarchies to layer-wise weight se-

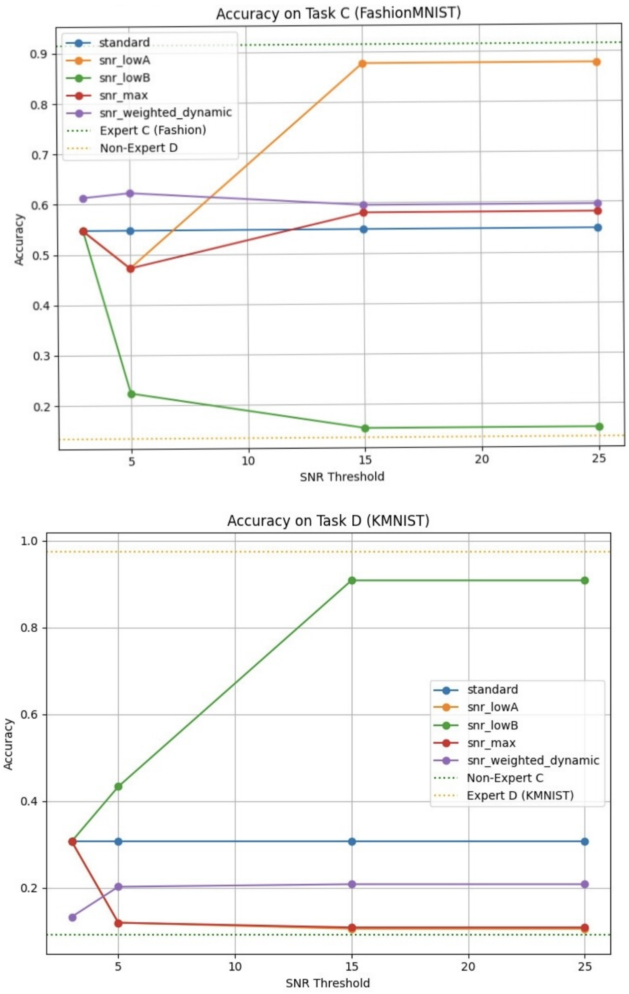

Figure 3. **ResNet-18 Stability Analysis.** Top: Accuracy on Task C (FashionMNIST). Note how `snr_lowB` (green line) causes a performance crash. Bottom: Accuracy on Task D (KMNIST), showing the opposite trend where `snr_lowA` fails. This symmetry highlights the extreme sensitivity of convolutional layers to the ”wrong” expert weights.

lection. The primary lesson from these experiments is that one size does not fit all:

1. Structural tasks (CoLA) benefit from aggressive SNR winning strategies.

2. Distributed semantic tasks (SST-2) are better served by Standard or Weighted merging to avoid losing subtle information.

3. Vision tasks (ResNet) require more sophisticated, perhaps smoother, transition mechanisms between layers to avoid the performance collapses observed in the `snr_lowB` and `snr_max` trials.

References

[1] V. A. Marchenko and L. A. Pastur, “Distribution of eigenvalues for some sets of random matrices,” 1967.

[2] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018.

[3] M. Wortsman et al., “Model soups: averaging weights of multiple fine-tuned models improves accuracy,” 2022.

[4] Eric Hartford, Lucas Atkins, Fernando Fernandes Neto, and David Golchinfar, “Spectrum: Targeted Training on Signal to Noise Ratio.”