Robustness and Geometry in Low-Rank Multimodal Fusion or *Investigating Uncertainty and robustness in Tensor-Based Multimodal Learning.*

A project in development by
Parth Sinha
MSc AI: School of EECS
Queen Mary University of London

Problem Statement:

The main problem that is being tackled is the strong and stable combination of heterogeneous data modalities, including image and text, into a single machine learning system that can resist adversarial perturbations. The traditional approaches are usually plagued with disjoint latent spaces of semantic relations being geometrically misaligned. This mismatch is not just the cause of poor generalization but also contributes to brittle models to distribution shifts and adversarial attacks, which usually causes poor uncertainty estimates in safety-critical applications. Moreover, the standard application of high-dimensional features incurred by fusing is often associated with colossal over-parametrization. This dimensionality curse does not only add to the cost of computation, but also adds the adversarial attack surface, making the model vulnerable to noise and spurious correlations.

It is aimed at developing a theoretically-based and strong architecture. I suggest the application of semantic continuity with Joint Wasserstein geometric regularization to guarantee distributional stability and at the same time, Low-Rank Tensor Factorization. Our hypothesis is that this low-rank regularization will serve as a spectral regularization procedure that prevents the high-frequency noise and non-robust features to improve adversarial robustness and uncertainty quantification without incurring prohibitive computational costs.

Abstract:

This report synthesizes a theoretical and practical framework for advanced multimodal machine learning, derived from a specific collection of code notebooks, handwritten mathematical notes, and academic literature. The core investigation establishes a composite architecture integrating Joint Wasserstein Autoencoders (jWAE) with Low-Rank Multimodal Fusion (LMF). The findings demonstrate that jWAE acts as a critical geometric conditioner: by enforcing a shared Gaussian prior on latent embeddings, it ensures semantic continuity and alignment across modalities (e.g., image and text). This regularization "linearizes" the latent manifold, thereby satisfying the mathematical prerequisites required for the LMF model to efficiently approximate high-order tensor interactions using low-rank factorizations. Furthermore, the report incorporates principles from Multiple Kernel Learning (MKL), utilizing Dual Lagrangian optimization to

provide a theoretical basis for weighting and connecting disparate feature kernels. The overarching conclusion is that bridging jWAE's representation learning with LMF's efficient fusion creates a scalable, mathematically robust pipeline for multimodal tasks.

We hypothesize that the rank-constrained tensor product does not just compress data, but 'disentangles' it. We aim to prove that specific rank factors correspond to specific semantic interactions (e.g., Rank 1 captures text sentiment, Rank 2 captures visual context), offering a mathematical basis for explainability.

Methodology

Theoretical Foundation: Additivity in Multiple Kernel Learning (MKL)
Our theoretical framework begins with the principles of Multiple Kernel Learning (MKL). As explored in our preliminary analysis, MKL extends standard Support Vector Machines (SVM) by defining a combined kernel K_β as a convex linear combination of M base kernels K_m

$$K_\beta(x, x') = \sum_{m=1}^{M} \beta_m K_m(x, x')$$

The optimization of this framework relies on the Dual Lagrangian, where the objective is to maximize the margin by solving for the optimal kernel weights β alongside the Lagrange multipliers α.

Crucial Observation: The defining characteristic of MKL is that the fusion of heterogeneous data representations occurs via **addition** in the Reproducing Kernel Hilbert Space (RKHS). This additive property suggests that if we can project multimodal data into a compatible geometric space, their fusion can be modeled as a weighted summation of their feature maps. This "additivity" forms the central thesis for our tensor approach.

The Tensor Transition and the Context Problem

While kernels offer a powerful theoretical basis, modern multimodal tasks often require explicit modeling of feature interactions (e.g., how a visual smile changes the context of spoken text). This necessitates the use of Tensors.

A naive approach might attempt to apply the additive logic of MKL directly to input tensors. However, simply adding raw tensors from different modalities (e.g., $Z_{video} + Z_{audio}$) fails because it treats modalities as independent channels, leading to a loss of **contextual interaction**. To capture the correlation between modalities (e.g., sarcasm detection where text and audio contradict), we must theoretically utilize the tensor outer product \$$\mathcal{Z} = z_1 \otimes z_2 \otimes \cdots \otimes z_m.$

The computational cost of this full tensor product is exponential, $O(\prod_{m=1}^{M} d_m)$, rendering it intractable for high-dimensional data. This creates a conflict: we need the **interaction capture** of tensors with the **additive efficiency** of MKL.

The Solution: Rank-Based Decomposition

To resolve this, we employ the concept of **Low-Rank Multimodal Fusion (LMF)**. Instead of computing the high-dimensional tensor $\mathcal{Z}$ explicitly, we decompose the weight tensor $\mathcal{W}$ into a set of low-rank factors. This allows us to perform the fusion in a lower-dimensional "rank space".

The importance of the **Rank Function** (r) cannot be overstated. It acts as the bottleneck that forces the model to learn the most salient interactions. By decomposing the weights into modality-specific factors $w_m^{(i)}$, we can rewrite the fusion operation. Instead of a massive outer product, the fusion becomes a summation of element-wise products across the rank dimension:

$$h = \sum_{i=1}^{r} \left(\bigodot_{m=1}^{M} (w_m^{(i)} \cdot z_m) \right)$$

Here, $\bigodot$ represents the Hadamard (element-wise) product. This equation creates a mathematical bridge back to MKL: we are effectively performing an **additive** combination $\sum_{i=1}^{r}$ of interaction terms, but we do so within a compressed, low-rank field that preserves multimodal context without the computational explosion.

Definition: The **Rank** (R) is the number of distinct "factors" or "components" used to approximate the complex interaction tensor.
Intuition: Think of the full interaction between modalities as a complex 3D shape. The "Rank" is the number of simple cylinders or blocks one can use to build an approximation of that shape.
Role: A higher rank (R) allows the model to capture more complex, nuanced interactions (more "blocks" to build the shape). A lower rank forces the model to learn only the most dominant, critical interactions, acting as a form of regularization that prevents overfitting and reduces computational cost.

Derivation: Integrating jWAE with Low-Rank Fusion

The final component of our methodology integrates Joint Wasserstein Autoencoders (jWAE) to condition the input space for this low-rank fusion. The jWAE enforces a shared Gaussian prior on the latent embeddings (z_v, z_t), ensuring that the input vectors lie on a smooth, continuous manifold. This "linearization" of the latent space is what makes the subsequent linear low-rank projection effective

Based on our mathematical findings , we derive the end-to-end architecture as follows:

1. **Input Conditioning:** The jWAE encoders (f_v, f_t) map raw inputs to latent vectors z_v, z_t.
2. Bias Augmentation: To account for unimodal biases within the tensor product, we append a 1 to the latent vectors:

$$\tilde{z}_v = [z_v; 1], \quad \tilde{z}_t = [z_t; 1]$$

3. **Low-Rank Projection:** We learn rank-specific projection matrices $W_v, W_t \in \mathbb{R}^{r \times (d+1)}$.
4. **The Fusion Equation:** The final derived forward propagation combines these components into a single differentiable function:

$$\hat{y} = \Theta_{pred} \cdot \left(\underbrace{(W_v \cdot \tilde{z}_v)}_{\text{Visual Factor}} \odot \underbrace{(W_t \cdot \tilde{z}_t)}_{\text{Text Factor}} \right)$$

This formulation proves that by using jWAE to regularize the input geometry, we can utilize a low-rank tensor framework to capture complex interactions additively, satisfying both the theoretical elegance of MKL and the contextual requirements of deep multimodal learning.

Problems issued and addressed:

Loss of Context Handled by Hadamard Product with Rank Integration

Standard fusion methods (like concatenation) often treat modalities as independent channels, failing to capture how one modality modifies the meaning of another (e.g., sarcasm in tone changing the meaning of text). To capture these multiplicative interactions without the prohibitive cost of a full tensor product, this framework uses **Low-Rank Fusion**.

The Mechanism (Hadamard Product): Instead of computing a massive outer product tensor $\mathcal{Z} = z_v \otimes z_t$, the model projects each modality into a set of R modality-specific factors (w_v, w_t). The fusion is then calculated as the **element-wise product (Hadamard product)** of these factors.This operation mathematically reconstructs the interactions in a compressed latent space. It preserves the "context" of interaction (i.e., how vectors align) but does so efficiently.

Dimensionality Problem Solved by jWAE

- **The Problem:** Raw data (pixels, audio waves) exists in extremely high-dimensional, non-linear spaces. Directly applying low-rank fusion to this raw data fails because the "interaction surface" is too complex to be approximated by a small number of factors (Rank R).
- **The jWAE Solution:** The Joint Wasserstein Autoencoder solves this not just by reducing dimensions (compression), but by **geometric regularization**. By forcing the latent embeddings (z_v, z_t) to match a shared **Gaussian prior** , jWAE "linearizes" the latent manifold.
- **Result:** It transforms a complex, rugged data landscape into a smooth, continuous bowl shape. This makes the data mathematically compatible with the linear, low-rank operations of the LMF, effectively bypassing the "curse of dimensionality" that would otherwise require an infinite rank to model accurately.

Problems Still Persisting

Data

- Issue: This architecture requires aligned, paired multimodal data (e.g., tuples of {image, text, label}). "Proper" data means data where the modalities actually interact (e.g., the text is ambiguous without the image). If the modalities are redundant (text describes the image perfectly), the fusion layer becomes unnecessary.
- Constraint: Unlike massive foundation models trained on unpaired internet scraps, this supervised pipeline typically requires curated datasets like CMU-MOSI (sentiment analysis) or MUSTARD (sarcasm detection) . Without such structured data, the joint training of jWAE and LMF is unstable or impossible.

Loyalty vs. Abstraction Trade-off

- **What We Compromise**: We pointedly admit that this architecture puts more emphasis on faithfulness, which is the capacity to rationalize the reasoning process of the model, as opposed to the sheer complexity and raw performance of foundation models that are huge.

- State-of-the-art autoregressive Transformers (e.g., GPT-4, LLaVA) use dynamic Self-Attention to put relative importance on each token and each image patch. Although this has the benefit of modeling complex sequential effects, it leads to a highly complex, non-linear, black box in which it is impractical to compute the contribution of each particular modality to the final decision.

- **Faithfulness (jWAE-LMF):** Non-invasive jWAE-LMF. Although, in contrast to the original model, there is a factorized embedding. Reducing the interaction to explicit Low-Rank factors results in a so-called fixed snapshot of context. This is not just a weakness but an element: this kind of stateness makes the model true to its purpose. It enables us to mathematically isolate and confirm precisely how the audio tensor has changed the text tensor, giving some sort of mechanistic transparency that is unavailable to dynamic Transformers.

- **The Acceptable Gap:** Although non-interpretable Multimodal Transformers (MulT) can have higher raw accuracy at leaderboards, it does so at the expense of explainability. This project will accommodate this performance gap as the cost needed to have a trustworthy, interpretable architecture in which the internal logic is visible and verifiable.

Reference and citations

I. Gurevych, S. Mahajan, T. Botschen, and S. Roth, "Joint Wasserstein Autoencoders for Aligning Multimodal Embeddings," *arXiv preprint arXiv:1909.06635*, 2019.

Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient Low-rank Multimodal Fusion with Modality-Specific Factors," *arXiv preprint arXiv:1806.00064*, 2018.

S. Varshneya, A. Ledent, P. Liznerski, A. Balinskyy, P. Mehta, W. Mustafa, and M. Kloft, "Interpretable Tensor Fusion," *arXiv preprint arXiv:2405.04671*, 2024.

Github Link: The minor implementations done on pipeline structure and not data, this does not include the jWAE implementation