An Independent Exploration of Multimodal Fusion: From Multiple Kernel Learning to Low-Rank Tensors

Parth Sinha
Queen Mary University of London

September 24, 2025

Abstract

This work documents an independent investigation into methods for multimodal data fusion. The research begins with a theoretical exploration of Multiple Kernel Learning (MKL), where key concepts such as additive kernels and Lagrangian expansion are discussed. The project then transitions to a practical implementation of a modern fusion technique, Low-Rank Multimodal Fusion (LMF), developed from scratch in PyTorch. The LMF model demonstrates a concrete method for fusing features from simulated visual, audio, and text modalities into a coherent, low-rank representation.

1 Introduction

The challenge of effectively combining data from multiple modalities (e.g., audio, visual, text) is a central problem in modern artificial intelligence. This research explores the hypothesis that the connectivity between tensors can be achieved through additive fusion mechanisms. The initial approach uses the principles of Multiple Kernel Learning (MKL) as a theoretical stepping stone. MKL provides a principled way to linearly combine multiple kernel functions, each operating on a different data representation, to find a more powerful vector space. This study leverages MKL as an inspiration for understanding how a constant or weighting scheme could be derived to quantitatively combine tensors.

2 Methodology

2.1 Theoretical Framework: Multiple Kernel Learning

The investigation starts with a standard SVM formulation with weights and biases. The core idea of MKL is to create a combined kernel, $K_\beta(x, x')$, as a weighted sum of base kernels, where each kernel K_m corresponds to a specific modality or feature representation. The optimization problem is to learn both the SVM parameters (w_m) and the kernel weights (β_m) simultaneously. This is expressed in the following minimization problem:

$$\min_{w_1,\ldots,w_M,\beta\in R^M} \left(\sum_{i=1}^{n} l\left(\sum_{m=1}^{M} \sqrt{\beta_m} \langle w_m, \Psi_m(x_i^m) \rangle_{\mathcal{H}^m} + b, y_i \right) + \frac{\lambda}{2} \sum_{m=1}^{M} \|w_m\|^2_{L^2(\mathcal{H}^m)} \right) \quad (1)$$

subject to $\beta_m \geq 0$ and $\|\beta\|_p \leq 1$.

To solve such a constrained problem, **Lagrangian expansion** is employed. This technique de-constraints the equation by introducing multiplicative constraints, allowing for a clearer projection of the vectorized equation's limits.

2.2 Practical Implementation: Low-Rank Multimodal Fusion (LMF)

Moving from theory to practice, a Low-Rank Multimodal Fusion (LMF) model was implemented from scratch using PyTorch. This model is specifically designed to handle multimodal data. The implementation fuses three simulated modalities:

- **Visual**: 128-dimensional feature vector.
- **Audio**: 64-dimensional feature vector.
- **Text**: 300-dimensional feature vector.

The LMF model projects each modality-specific feature vector into a lower-dimensional space defined by a fusion rank. These projections are then combined through element-wise multiplication (Hadamard product) to create a fused representation. This fused tensor is then passed through a final linear layer to produce a prediction. This method efficiently captures the complex interactions between modalities in a low-rank tensor space.

3 Experiments and Results

The theoretical MKL framework was visualized by applying Principal Component Analysis (PCA) to the data. The results demonstrate that the combined kernel can find a suitable vector space that improves class separability.

The LMF implementation was successfully trained on dummy data for a regression task. A simple training loop using an Adam optimizer and MSE loss showed a consistent decrease in training loss over 20 epochs, confirming that the model implementation is correct and capable of learning.

4 Conclusion and Future Work

This independent study successfully connected the theoretical underpinnings of data fusion through MKL with the practical implementation of a modern LMF model. It shows that additive principles can be a powerful starting point for multimodal fusion. The ultimate goal of this research is to build upon these findings to develop a constant or method for quantitatively combining tensors, drawing inspiration from the paper "Interpretable Tensor Fusion" by Varshneya et al.