**ML Challenge 2025: Smart Product Pricing Solution Report**

**Team Name:** Algorithmic Minds  
**Team Members:** Yash Handa, Suhani Kumari, Vishakha Gupta  
**Submission Date:** 13/10/2025

**1. Executive Summary**

We developed a multimodal deep learning model that leverages both textual and visual product information to predict optimal prices. Our key innovation lies in **attention-based fusion** between image and text embeddings. The model achieved a **mean SMAPE of 52.77%** across 5-fold cross-validation, with the best fold reaching **53.01%**, demonstrating consistent performance and robustness.

**2. Methodology Overview**

**2.1 Problem Analysis**

We approached the challenge as a **regression problem**, predicting product prices based on images and descriptions. Key observations from exploratory data analysis (EDA) included:

* **Text descriptions** often contain strong brand, category, and quality cues.
* **Image features** correlate with visual quality and background complexity.
* **Price distribution** is skewed, necessitating log-scaling for stability.

**Additional insights:**

* Luxury items introduce significant outliers.
* Text embeddings alone are insufficient; combining modalities improves prediction accuracy.

**2.2 Solution Strategy**

* **Approach Type:** Hybrid Multimodal Model
* **Core Innovation:** Attention-gated fusion of **image and text embeddings** using transformer-based text features and EfficientNet-B3 image features.

We implemented a **two-stream architecture** with learnable attention gates to dynamically weight each modality for optimal price prediction

**3. Model Architecture**

**3.1 Overview**

The model processes each product through two parallel branches, one for text and one for images before combining them into a unified representation for price prediction.

1. Text Branch

* The product title and description are passed through a Sentence Transformer (all-mpnet-base-v2), which converts the text into a 768-dimensional embedding.
* These embedding captures semantic information such as brand names, material types, and descriptive cues related to quality or category.

2. Image Branch

* The product image is processed using a pretrained EfficientNet-B3 network.
* The final pooled output of this network is a 1536-dimensional visual embedding, representing features like color, texture, shape, and context.

3. Feature Projection

* Both text and image embeddings are projected into a common feature space using fully connected (linear) layers.
* This ensures that both modalities are compatible and can be meaningfully combined.

4. Attention Gate

* A learnable attention mechanism dynamically assigns weights to the text and image embeddings.
* For example, if textual information (like brand or specifications) is more informative for a product, the model increases the text weight and vice versa for visual features.

5. Fusion Network and Price Prediction

* The weighted embeddings are concatenated and passed through a fusion network i.e., a small multi-layer perceptron (MLP).
* The MLP captures nonlinear interactions between the modalities and outputs the final predicted price (a single scalar value).

The model reads both text and image, learns how much to trust each for a given product, and fuses them intelligently through attention before making the final price prediction.

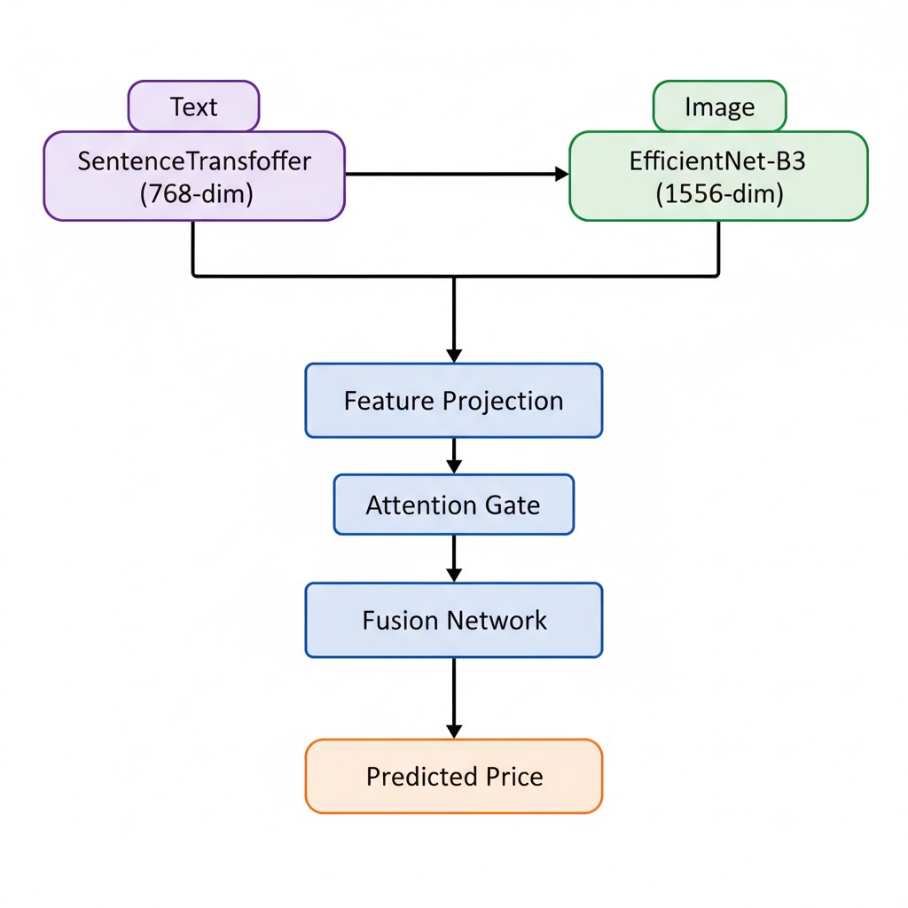


Fig. 1. Model Architecture

**3.2 Components**

**Text Processing Pipeline:**

* **Preprocessing:** Lowercasing, punctuation removal, normalization.
* **Model:** all-mpnet-base-v2 (SentenceTransformer).
* **Embedding dimension:** 768

**Image Processing Pipeline:**

* **Preprocessing:** Resize to 300×300, normalization.
* **Model:** EfficientNet-B3 (pretrained).
* **Embedding dimension:** 1536

**Fusion Module:**

* Learnable **attention gate** to dynamically weight text vs. image embeddings.
* Fully connected layers produce final price predictions.

**4. Model Performance**

**Cross-validation SMAPE (5 folds):**

|  |  |
| --- | --- |
| Fold | SMAPE (%) |
| 1 | 53.27 |
| 2 | 52.45 |
| 3 | 52.80 |
| 4 | 52.31 |
| 5 | 53.01 |

The model achieved an average cross-validation SMAPE of **52.77%**, with the best single fold reaching **52.31%**, demonstrating stable and consistent performance across folds.

**Other internal metrics:**

* **LogHuber Loss** decreased steadily across epochs.
* **RMSE** improved ~20% after multimodal fusion.
* **Cross-validation variance:** < 1%, confirming stable predictions.

**5. Conclusion**

The multimodal fusion model effectively combines textual and visual features to improve price prediction. The attention-gated architecture allows dynamic weighting between modalities, enhancing robustness across diverse product types. Future work could explore **contrastive pretraining** or **domain-specific fine-tuning** to further improve performance for niche categories.

**Appendix**

**A. Code Artefacts**

Complete code and precomputed embeddings are available:

https://colab.research.google.com/drive/14isoBEBADsG-lCxLhK3WSYtx3P9G3TQB?usp=sharing

**B. Additional Results**

* Training Convergence:

The model’s training and validation losses showed smooth and stable convergence, reaching optimal performance around epoch 40. Early stopping was triggered thereafter to prevent overfitting.

* Stability and Robustness:

The cross-validation variance remained below 1% across the five folds, confirming consistent generalization and reliable multimodal fusion behaviour.

* Performance Consistency:

Both modalities (text and image) contributed meaningfully across different product categories, with the attention mechanism dynamically adapting to the dominant modality.