Amazon ML Challenge 2025
Team: CTRL + ALT + DEV
Amrita Vishwa Vidyapeetham, Coimbatore

Smart Product Pricing Challenge

Technical Report

Team Members: Vishal S, Krishna Leela Amritha Nandini, Suriya KP, SP Saran Dharshan
Date: 13 October 2025

Executive Summary

This report presents a comprehensive analysis of multiple machine learning approaches developed for the Amazon ML Challenge 2025, which focused on predicting grocery product prices from textual descriptions and images. Our best performing solution achieved a SMAPE score of 40.777% on the 25,000-entry public dataset using a sophisticated multi-modal fusion architecture that combines CLIP [1] and DistilBERT [2] models with contrastive learning techniques. The final submission represents an ensemble average of our top two models, demonstrating the effectiveness of multi-modal deep learning approaches for price prediction tasks.

1 Approach: Multi-Modal Fusion Architecture

Our solution employs a dual-backbone fusion architecture that leverages both visual and textual information through two parallel processing streams. The first stream utilizes OpenAI's CLIP model (clip-vit-large-patch14) [1] to extract 768-dimensional embeddings from both product images and text descriptions, creating a unified vision-language representation space. The second stream processes text through DistilBERT (distilbert-base-uncased-finetuned-sst-2-english) [2] with two augmented views to generate 256-dimensional projected embeddings. These multi-modal features are then normalized and concatenated before being fed into a multi-layer perceptron fusion head that predicts the final price.

The architecture's key innovation lies in its sophisticated training strategy that combines multiple loss functions. The primary regression loss uses Huber loss ($\delta = 1.0$) on log-transformed prices to handle the skewed price distribution and outliers robustly. This is augmented by two contrastive learning components: a CLIP image-text contrastive loss (InfoNCE with temperature $\tau = 0.07$ and weight $\alpha_{\text{clip}} = 0.20$) that aligns visual and textual representations, and a DistilBERT SimCSE-style contrastive loss (weight $\alpha_{\text{txt}} = 0.10$) that improves text representations through comparison of augmented views with 6% word masking probability. This multi-loss optimization strategy enables the model to learn rich, discriminative representations while maintaining robust price prediction capabilities.

The model was trained for 15 epochs with a learning rate of 2e-5, batch size of 16, and 6% warmup ratio, accumulating a total of 501,196,546 trainable parameters. Mixed precision training (FP16) was employed for memory efficiency, while gradient clipping (max norm 1.0) and weight decay (0.01) ensured stable optimization. Text inputs were tokenized with maximum lengths of 64 tokens for CLIP

and 192 tokens for DistilBERT, while images were resized to 224×224 pixels and processed through CLIP's vision encoder. A critical design decision was the zero-padding strategy for handling missing images, which preserved the full sample size without performance degradation, demonstrating the architecture's robustness to incomplete data.

2 Alternative Approaches and Performance Analysis

We systematically explored eleven different approaches spanning multiple paradigms, ranging from pure vision-language models to traditional machine learning with feature engineering. The CLIP-based approaches achieved competitive results, with CLIP-Large (10 epochs) scoring 41.541% and the base CLIP model (5 epochs) achieving 42.234% SMAPE. These single-modality solutions demonstrated strong performance through their unified vision-language understanding but lacked the sophisticated text processing capabilities of our winning dual-backbone architecture.

Text-only approaches using BERT variants performed moderately, with DistilBERT SST-2 achieving 45.429% and BERT-Large with contrastive learning scoring 48.624% SMAPE. While these models benefited from contrastive learning and two-stage training strategies, their inability to leverage visual information significantly limited their predictive power. The embedding-based approaches, utilizing models like E5-base-v2 (50.702%) and Qwen3-Embedding-0.6B (57.402%) coupled with LightGBM regressors, suffered from static embeddings without end-to-end learning capabilities. Traditional machine learning approaches with extensive feature engineering, including brand detection, size extraction, TF-IDF, and LDA features, achieved 51.491% SMAPE but could not capture the complex non-linear relationships that deep learning models naturally encode.

Table 1: Performance Comparison of Top Approaches

Rank	Approach	SMAPE (%)
1	CLIP + DistilBERT Fusion	**40.777**
2	CLIP Large (10 epochs)	41.541
3	CLIP + E5 Embeddings	42.006
4	CLIP Large (5 epochs)	42.234
5	DistilBERT SST-2	45.429
6	BERT Large + Contrastive	48.624
7	E5 Base v2 + LightGBM	50.702
8	Feature Engineering + LightGBM	51.491

3 Key Insights and Conclusions

Our experimental results reveal several critical insights for multi-modal price prediction. Multi-modal approaches consistently outperformed single-modality solutions, with all top four performers leveraging visual information alongside textual features. The superiority of the dual-backbone fusion architecture over pure CLIP implementations demonstrates that specialized text processing through DistilBERT provides complementary information that enhances prediction accuracy. Contrastive learning emerged as a universally beneficial technique, with all top-performing models incorporating some form of self-supervised representation learning to improve feature quality.

The analysis revealed that model capacity and training duration significantly impact performance, with larger models like CLIP-Large and BERT-Large generally outperforming their smaller counterparts, and extended training (15 epochs versus 5-10 epochs) yielding measurable improvements. However,

the success of our approach cannot be attributed solely to scale. The careful tuning of loss weights, temperature parameters, and architectural design choices proved equally crucial. The zero-padding strategy for missing images exemplifies how thoughtful engineering decisions can maintain dataset integrity while ensuring robustness.

Looking forward, our winning approach with its 40.777% SMAPE score provides a strong foundation for production deployment in e-commerce price prediction systems. The architecture's ability to handle missing data gracefully, combined with its sophisticated feature fusion and contrastive learning components, makes it particularly suitable for real-world applications where data quality may vary. Future improvements could explore additional data augmentation strategies, more sophisticated ensemble methods beyond simple averaging, alternative fusion architectures such as cross-attention mechanisms, and semi-supervised learning techniques to leverage unlabeled product data. The approximately 5-hour training time on modern hardware makes this approach computationally feasible for iterative development and production retraining cycles, though inference optimization would be necessary for real-time pricing applications at scale.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.