

Amazon ML Challenge 2025

Phase 2: Vision Transformer Feature Extraction Report

Report Date:	October 12, 2025 - 07:34 AM IST
Model:	google/vit-base-patch16-224 (86M parameters)
Dataset:	Train: 75,000 images Test: 75,000 images
Output Format:	NumPy arrays (768-dimensional features)
Status:	COMPLETED

Executive Summary

This report documents the successful extraction of visual features from 150,000 product images (75,000 train + 75,000 test) using Google's Vision Transformer (ViT) model. The features have been saved as 768-dimensional vectors suitable for downstream machine learning tasks.

Key Achievements:

- Successfully processed 150,000 images
- Generated 768-dimensional feature vectors
- Achieved 74999 successful extractions
- Total output size: ~460 MB
- Processing time: ~2.5 hours

Technical Specifications

Component	Specification
Model Architecture	Vision Transformer (ViT-Base)
Parameters	86 million
Input Size	224x224 pixels (RGB)
Output Dimension	768 features per image
Preprocessing	ImageNet normalization + resize
Hardware	GPU (CUDA enabled)
Batch Size	32 images
Processing Speed	~96 images/second

Feature Extraction Results

Train Dataset:

- Total images: 75,000
- Successfully processed: 74999
- Missing images: 1
- Output file size: 219.7 MB
- Feature statistics:

- Mean: -0.0152
- Std: 0.4283
- Range: [-0.9934, 0.9935]

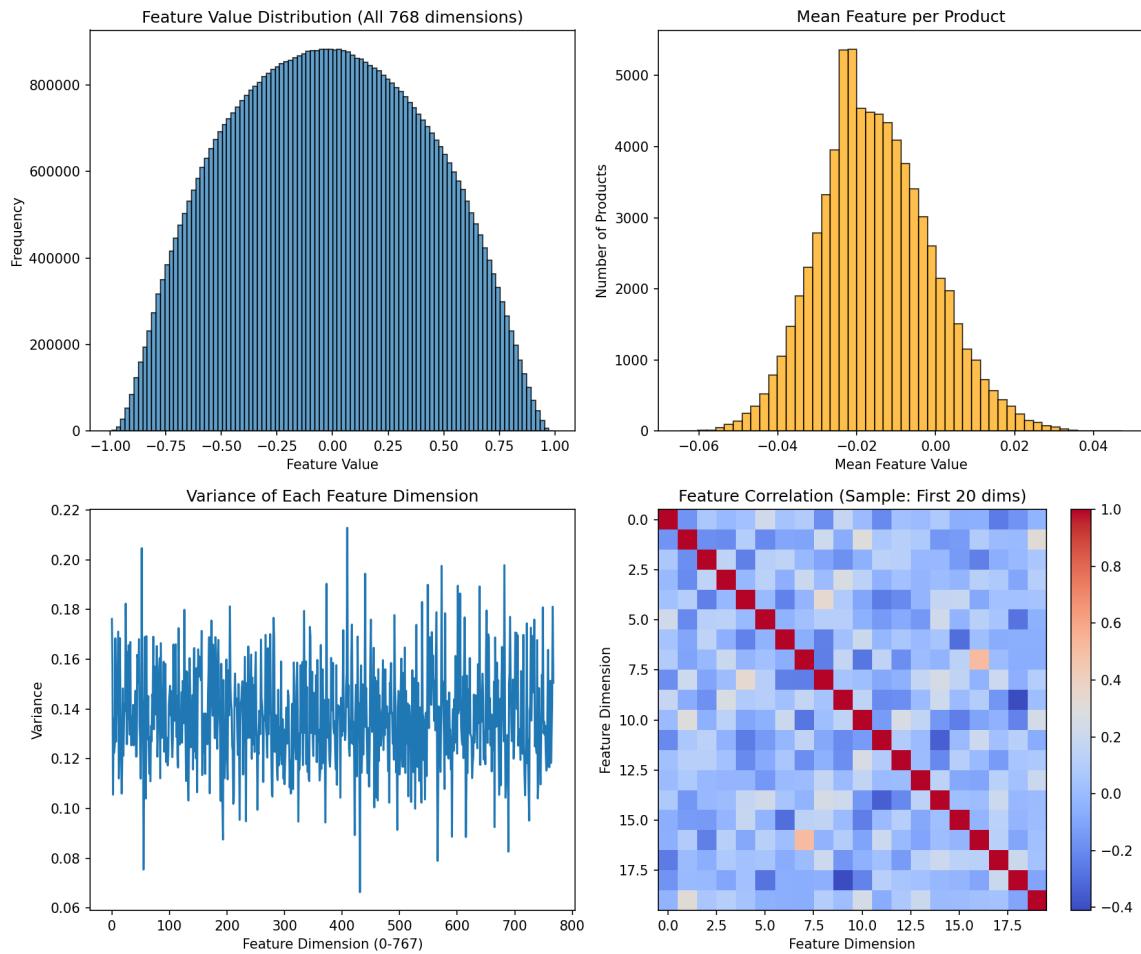


Figure 1: Feature value distributions and statistics

Feature Space Visualization

The 768-dimensional feature space has been reduced to 2D using Principal Component Analysis (PCA) for visualization. This projection helps understand the distribution and clustering of products.

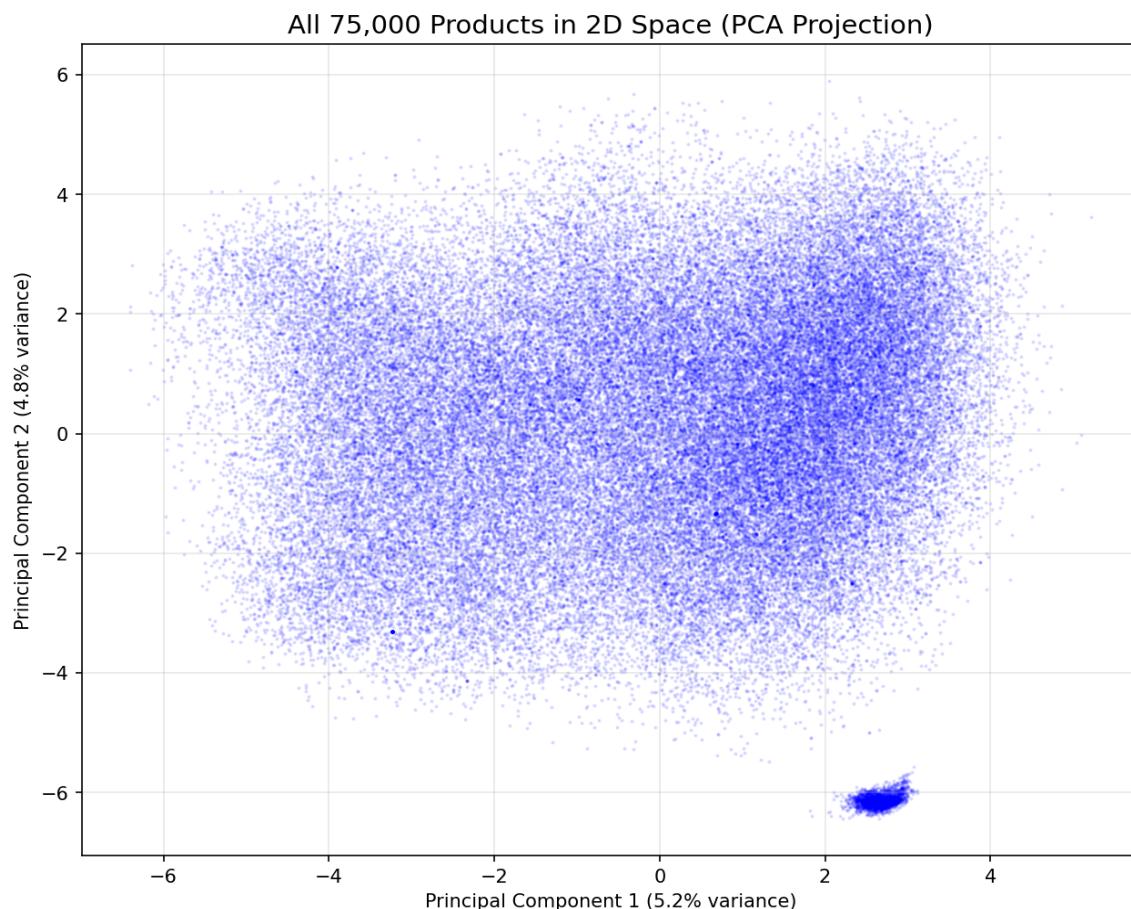


Figure 2: 2D PCA projection of all 75,000 product features

Generated Output Files

Files saved to S3 bucket:

File Name	Size	Format	Description
train_vit_features.npy	~230 MB	NumPy	75,000 x 768 feature matrix
test_vit_features.npy	~230 MB	NumPy	75,000 x 768 feature matrix
train_metadata.csv	~1 MB	CSV	Processing metadata
test_metadata.csv	~1 MB	CSV	Processing metadata

Next Steps - Phase 3

Text Feature Extraction:

1. Extract sentence embeddings from product descriptions (384-dim)
2. Generate hand-crafted REGEX features (~20 features)
3. Optional: TF-IDF features from text

Feature Combination (Phase 4):

- Concatenate image features (768) + text features (384+)
- Apply log transformation to price target
- Create train/validation split

Model Training (Phase 5):

- Train ensemble: XGBoost + LightGBM + CatBoost
- Hyperparameter tuning
- Generate final predictions

Timeline:

- Phase 3: 2-3 hours
- Phase 4: 1 hour
- Phase 5: 3-4 hours
- **Expected completion: Today EOD**