Multimodal Price Prediction with EfficientNet-B3 - Technical Approach

Overview

Automated product price prediction is crucial for e-commerce platforms, enabling competitive pricing, catalog management, and dynamic marketing. Traditional models primarily rely on textual catalog data (titles, descriptions, specifications), missing out on the rich visual cues inherent in product images. Modern deep learning makes it possible to fuse both sources—text and image—for more accurate and robust predictions. This document outlines the technical solution for enhancing price prediction by combining a DistilBERT text encoder with an EfficientNet-B3 image encoder, using advanced fusion strategies for optimal performance.

Motivation for EfficientNet-B3

EfficientNet-B3 is a state-of-the-art convolutional neural network (CNN) architecture that achieves better accuracy and efficiency compared to classic CNNs like ResNet-50. Its compound scaling method allows for improved performance without a proportional increase in computational resources. For retail datasets with diverse product images (ranging from food, electronics, luxury goods to household items), EfficientNet-B3 reliably extracts visual features relevant to pricing, such as product quality, packaging, size, and branding cues.

Detailed Solution Approach

- Data Modalities:
    - - Text: All catalog information (name, brand, description, specifications)
    - - Image: Product images from URLs (image_link)

Data Processing

1. 1. Text Pipeline:
    - Use DistilBERT tokenizer on concatenated catalog content.
    - Pad/truncate to max sequence length.
    - Output: 768-dimensional semantic feature vector per sample.
2. 2. Image Pipeline:
    - Download, center-crop, and resize images to 300x300 pixels (EfficientNet-B3 input size).

- Normalize with ImageNet statistics.
- Augment with random flips, brightness, and contrast changes during training.
- Feed processed image through EfficientNet-B3 (pre-trained).
- Extract global average pooled (GAP) feature (1,536 dims).
- Dense projection to 512 dims for fusion parity.

Multimodal Fusion

- Intermediate fusion approach:
  - Concatenate 768-dim text embedding + 512-dim image feature (1280-dim).
  - Feed through fusion block (dense layers, dropout, normalization, relu).
  - Optional: Add multi-head self-attention layer.

Regression Head

- Multi-layer perceptron (MLP) with skip connections:
  - Hidden sizes: 1280 → 640 → 128 → 1
  - Activation: ReLU, with dropout regularization.
  - Output: Scalar price prediction per product.

Training and Evaluation

- Loss: Symmetric Mean Absolute Percentage Error (SMAPE) suitable for skewed prices.
- Optimizer: AdamW (different learning rates for each branch).
- Scheduling: Warmup, progressive unfreezing.
- Batch size: 16–32 (mixed precision).
- Early stopping: Validation SMAPE.
- Error handling: For missing/broken images, fallback to text-only.

Implementation Roadmap

- Data Preparation: Download/cache product images. Build a multimodal dataset class.
- Model Construction: Assemble text and image encoders; implement fusion and regression.
- Training Infrastructure: Augment pipeline to process both modalities, train/validate.
- Interpretability: Attention visualization, feature importance analysis.

Performance Expectations

- 20–25% better accuracy vs text-only baseline (literature reference).
- Robustness for visually-similar products at different prices.
- Improved generalization across product categories where both modalities provide value.

Conclusion

Integrating EfficientNet-B3 for visual analysis within a multimodal fusion architecture positions the solution at the cutting edge of e-commerce AI. This approach not only improves price prediction accuracy but also enables explainable pricing decisions, scalable across diverse product categories.

References

- [21][27] EfficientNet literature and benchmarking studies
- [66] Leading multimodal fusion strategies for e-commerce
- [41][44] SMAPE in price regression problems