Multimodal Machine Learning: Housing Price Prediction using Images + Tabular Data

Summary:

In this project, I developed a regression model that predicts housing prices using multimodal machine learning, combining image data (house photos) with tabular data (like square footage, number of bedrooms, and year built). The goal was to fuse both numerical and visual information to improve prediction accuracy using deep learning.

What I Did:

- Created synthetic tabular data with key housing features for 1,000 samples.
- Simulated house images using random 64x64 RGB arrays.
- Preprocessed tabular data with StandardScaler, and split both modalities into training and test sets.
- Built a CNN to extract visual features from images.
- Designed a dense neural network for the tabular features.
- Fused both models at the feature level, followed by regression layers for price prediction.
- Trained the multimodal model using MSE loss and Adam optimizer in TensorFlow/Keras.
- Evaluated performance using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- Visualized loss curves to monitor training vs. validation performance.

Skills Gained:

- CNN architecture and feature extraction
- Multimodal input modeling in Keras
- Data preprocessing and feature scaling
- Loss function tuning for regression tasks
- Evaluation metrics for continuous prediction (MAE, RMSE)
- Feature fusion design (tabular + image)

What I Learned:

Multimodal models unlock richer representations than either tabular or image data alone. Designing models that handle different data types in parallel requires careful pipeline construction. I also learned the importance of choosing the right metrics, MAE is intuitive, while RMSE penalizes larger errors more harshly. Even with synthetic data, the exercise gave me a solid grasp of real-world deployment challenges, like preprocessing, model size, and interpretability.

Next Steps:

- Replace fake images with real house photos (e.g., from Zillow or Redfin datasets)
- Use pretrained CNNs like ResNet50 for better image feature extraction
- Add real location data (zip codes, school ratings, etc.)
- Deploy the model using Flask or FastAPI as a web service

Reflection:

This wasn't just about predicting prices, it was about teaching machines to *see* and *understand*. By combining vision and structure, I explored how future AI systems will reason not with one sense, but with many.