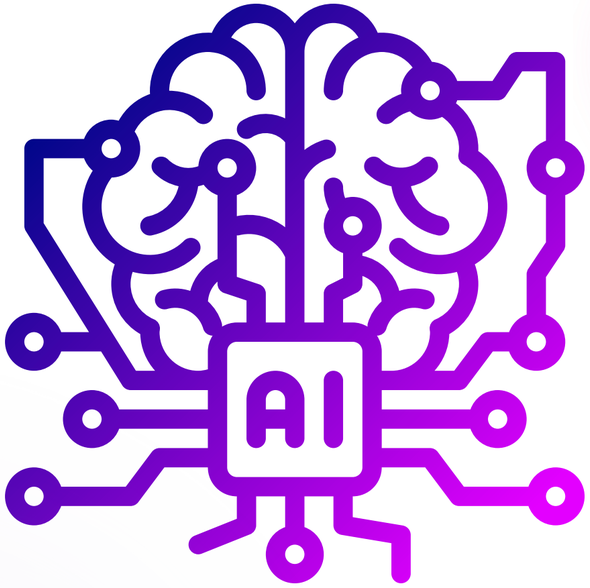


FASHION PRODUCT RECOMMENDATION USING MULTIMODAL LEARNING

Names: Marina Reda (221101235), Omar Adly (221101398)

Course: AIE417: Selected Topics in Artificial Intelligence 1

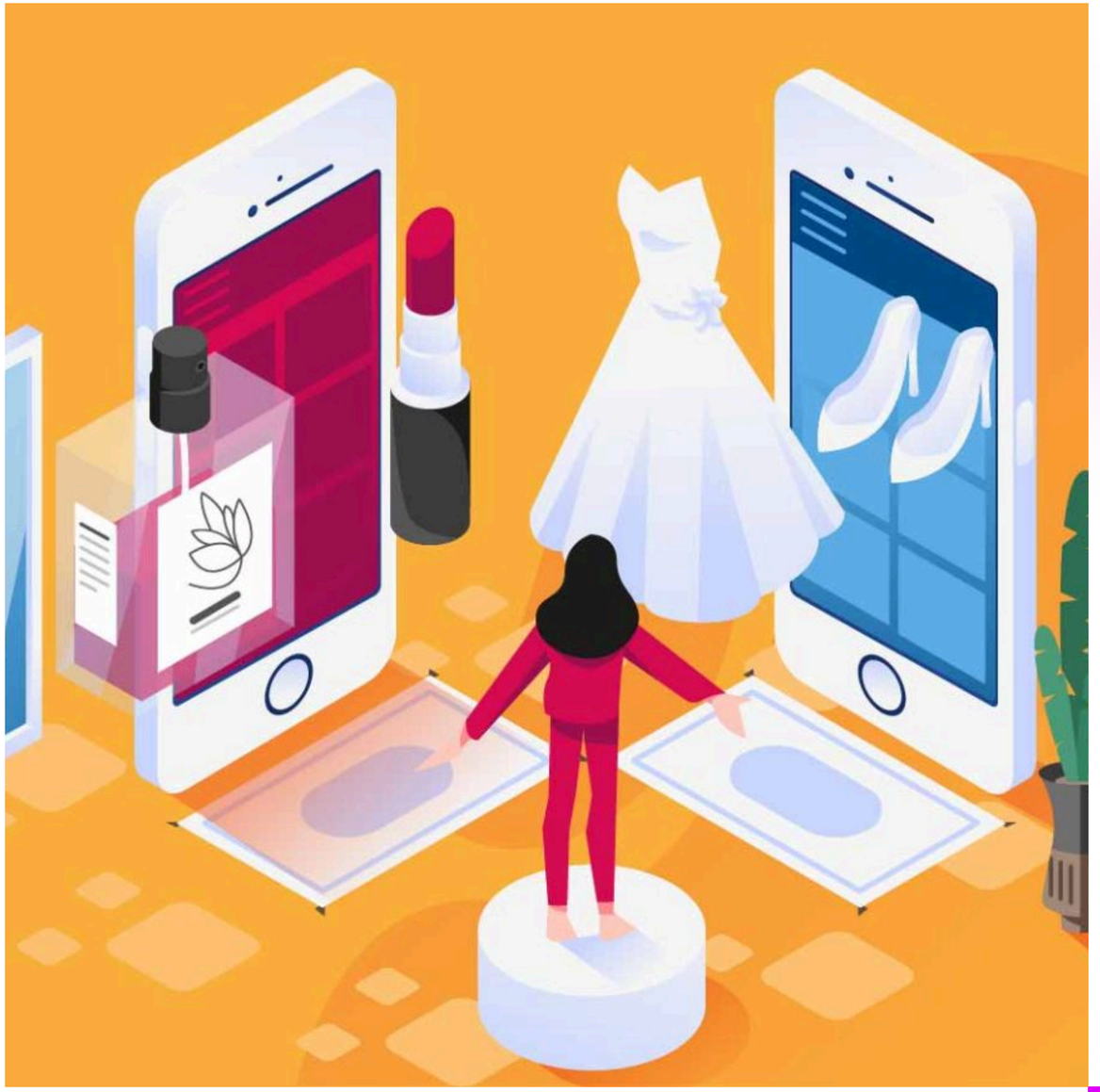
Instructor: Dr. Mohamed Ghetas

Introduction

- Problem Addressed: Explain the challenge of navigating vast product catalogs in online retail and the need for accurate product categorization and recommendations.
- Objective: To leverage multimodal learning to classify fashion products by integrating image and text data for a better user experience.
- Significance: Highlight how the approach improves shopping experiences and boosts customer satisfaction.

Dataset

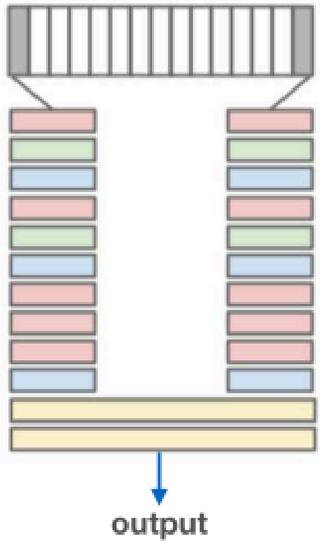
- Dataset Description:
  - Total samples in dataset: 44441
  - Images: High-quality (1080 x 1440 px).
  - Text: Titles and descriptions of products.
  - Labels: Product categories.
- Preprocessing Steps:
  - Images resized to 224x224 px.
  - Text tokenized and padded (max length: 100).
  - Labels encoded via one-hot encoding.

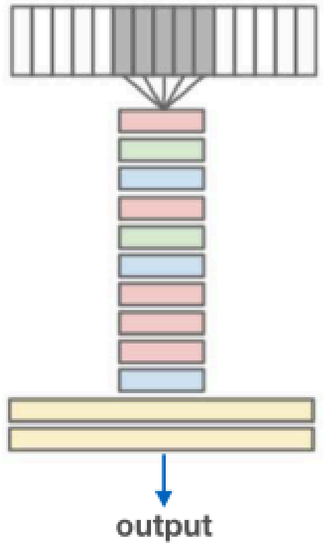


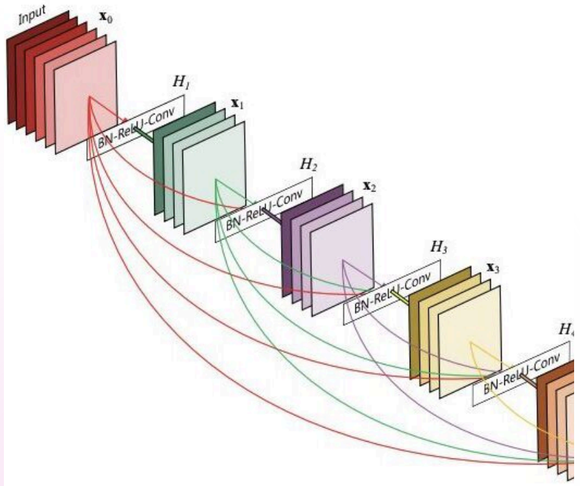
Model Design and Implementation

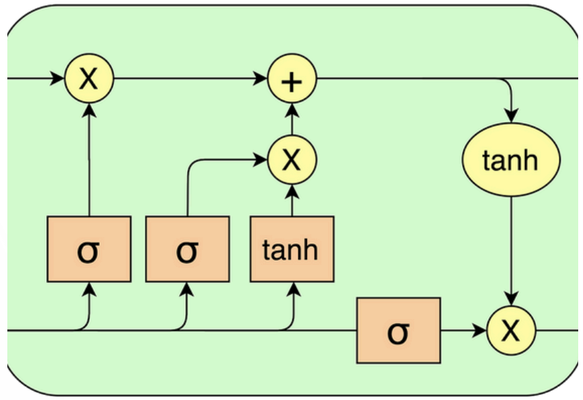
Multimodal Learning Techniques:

- Image: Features extracted using EfficientNetB0.
- Text: Processed with an LSTM-based model.
- Fusion: Early fusion by concatenating image and text features.









Model Architecture:

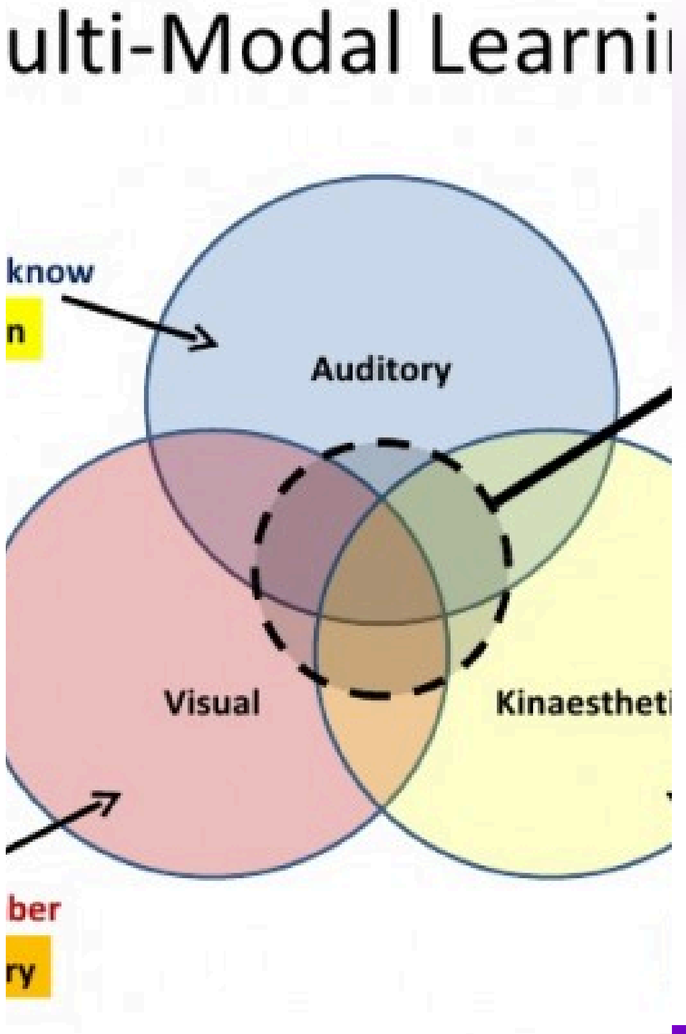
- Image Input: EfficientNetB0 → Global Average Pooling.
- Text Input: Embedding layer → LSTM.
- Output: Fully connected layer with softmax for classification.

Implementation Details:

- Framework: TensorFlow/Keras.
- Optimizer: Adam (learning rate: 0.001).
- Loss Function: Categorical Crossentropy.

Multimodal Learning Comparison

Paper	Model Architecture	Image Representation	Text Representation	Fusion Technique	Dataset	Accuracy
This Project	EfficientNetB0 + LSTM	EfficientNetB0	LSTM	Early Fusion (Concatenation)	Curated Fashion Product Images Dataset	91%
"Multi-modal Fashion Retrieval with Cross-modal Attention"	CNN + Transformer	CNN	Transformer	Cross-modal attention	UT-Zappos dataset	89.5%
"Learning Deep Representations of Fine-grained Visual Descriptions"	CNN + LSTM	CNN	LSTM	Attention-based fusion	DeepFashion dataset	88.3%



Demonstrate that the model (91% accuracy) outperformed existing models.

Challenges

- Balancing image and text feature contributions during fusion.
- Handling missing or inconsistent text descriptions.

Improvements

- Data Augmentation: Enhance generalizability with augmented datasets (images and text).
- Hybrid Fusion: Explore advanced fusion techniques for better performance.
- Transfer Learning: Fine-tune pre-trained models for domain-specific feature extraction.

Conclusion

- Multimodal learning proved effective for fashion product categorization.
- Future Directions:
  - Incorporate additional data like user reviews and brand attributes.
  - Explore audio or video data for richer insights.
- Emphasize the groundwork laid for AI-based recommendation systems to enhance personalization in e-commerce.

THANK YOU

Any questions ??