Multi-Modal Data Retrieval

Deep Learning Divas

December 10, 2025

The Problem

- How do we search for images using text queries?
  - Or find relevant captions for a given image?
- Challenge: Images and text live in different spaces
  - Images: pixel intensities, visual features
  - Text: words, semantic meanings
- Need: A shared representation to bridge modalities

Why Does This Matter?

- Real-world applications:
  - Image search engines
  - Content-based retrieval systems
  - Accessibility tools for visually impaired users
- Traditional approach: Treat modalities separately
  - Limited cross-modal understanding
- Our opportunity: Modern deep learning enables shared representations
  - More accurate retrieval
  - Better generalization across domains

Our Solution: Modernizing Correspondence Autoencoders

- Original Corr-AE (Feng et al., 2014)
  - Used Restricted Boltzmann Machines for feature extraction
  - Shared latent space for image and text
- Our modernized approach:
  - Replace RBMs with pretrained models:
    - ResNet-50 for image features
    - BERT for text embeddings
  - Compare multiple autoencoder architectures
  - Evaluate different alignment loss functions
- Dataset: Flickr8k (8,000 images, 5 captions each)

Outline

1. Background & Related Work
2. Methodology
   - Feature extraction
   - Autoencoder architectures
   - Loss functions
3. Experimental Results
   - Performance metrics
   - Architecture comparison
4. Conclusions & Future Work