A Multimodal Siamese Network for Learning Similarity in the WikiDiverse Dataset

AICS Project

Dawit J
University of Gothenburg

January 15, 2026

Contents

Motivation & Problem

- **The Challenge:** Human cognition seamlessly integrates text and images. Aligning images and text that refer to the same entity (e.g., a person, place, or event). Can AI learn to understand when they describe the same thing?
- **Goal:** Learning cross-modal similarity—a crucial component for retrieval, question answering, and recommendation systems, but not entity linking.
- **Key Question:** Does explicitly modeling interactions between text and image features improve similarity learning compared to processing them separately?

Approach:Two Attention Mechanisms

- **Model Architecture:** A Siamese Network with twin encoders (ResNet50 for images, BERT for text) [Bromley et al., 1993, Koch et al., 2015, Nag, 2022].
- **Experiment 1 – Base Attention:** Independent self-attention within each modality. No cross-modal interaction.
- **Experiment 2 – Multimodal Attention:** A cross-attention layer allows image and text features to interact before the final similarity score.

Model Architecture

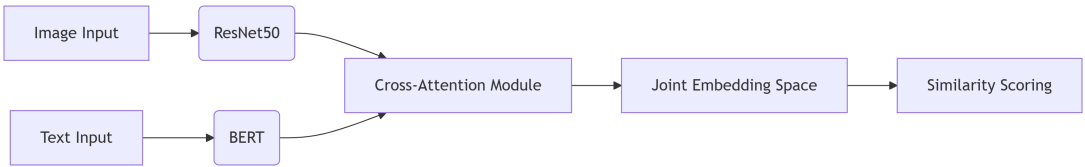

Figure: Model Architecture: Siamese Network

Dataset & Training

- **Dataset:** WikiDiverse (85,000 image-caption pairs of People, Places, Concepts, Events) [Liu et al., 2021].
- **Key Feature:** Contains natural ambiguity (e.g., "Turkey" = country or bird).
- **Training:** Contrastive loss, 20 epochs, early stopping. Limited to 100 pairs/entity to prevent bias.

Dataset statistics

Split	Inst (Sent)	Inst (Ment)	Ment/Inst	Recall@10	F1 Score
Training	6312	13205	2.09	88.62%	–
Validation	755	1552	2.06	89.17%	74.19%
Testing	757	1570	2.07	88.01%	73.34%

Table: Dataset statistics and model performance across training, validation, and test splits

Main Quantitative Results

- **Table 2:** Performance comparison (Accuracy, Precision, Recall, F1, AUC).
- **Key Finding:** Multimodal attention outperforms base attention across all metrics.
- **Biggest Gain:** $+4.05\%$ in Precision – meaning far fewer false positives.

Models Performance

Metric	Base attention(%)	Multimodal attention(%)	Difference(%pt)
Accuracy	83.49	86.22	+2.73
Precision	82.80	86.85	+4.05
Recall	85.54	86.31	+0.77
F1 Score	84.15	86.58	+2.43
AUC	90.52	91.61	+1.09

Table: Performance Metrics Comparison

Understanding the Gain - Ambiguity Resolution

- **The Test:** The dataset contains ambiguous surface forms (e.g., "Turkey" = country or bird? "Prime Minister" = which one?).
- **Qualitative Finding:** Multimodal attention better disambiguates these cases.
- **Example:** "Turkey"
  - **Base Model:** Confused images of the bird and the country for a "Thanksgiving" caption.
  - **Multimodal Model:** Used visual (food) + textual ("Thanksgiving") context to correctly choose the bird.
- **Why it matters:** Shows the model is learning deeper contextual similarity, not just keyword matching—a key ability for downstream tasks like entity linking.

Discussion & Implications

- **Answering the Research Questions:**
  1. **Yes,** joint multimodal attention improves similarity learning over independent base attention.
  2. **Yes,** the model shows promise in separating ambiguous entity referents through context, acting as a powerful similarity component for larger systems.
- **Theoretical Implication:** Aligns with ideas from Word Sense Disambiguation—context (from both modalities) is key to meaning.
- **The model Practical Implication:** This model isn't an entity linker, but its rich, attention-informed similarity scores can significantly improve the front-end of retrieval and entity linking pipelines.

Limitations & Future Work

- **Limitations:**
  - Uses frozen, pretrained encoders (ResNet/BERT). Not fine-tuned end-to-end for this specific task.
  - Performance lower on abstract "Concepts" vs. concrete "People/Places".
- **Future Directions:**
  1. Fine-tune the encoders jointly with the attention layer.
  2. Explore more complex, hierarchical attention over image regions.
  3. Incorporate external knowledge (e.g., entity definitions) to tackle abstract concepts.
  4. Test generalization on other, noisier datasets.

Conclusion

- **Main Contribution:** Demonstrated that a Siamese network with explicit multimodal (cross) attention learns a superior similarity space compared to one with independent modality attention.
- **Key Outcome:** The model effectively aligns images and captions by letting them interact, leading to better performance and improved disambiguation of ambiguous entities.
- **Final Remark:** Carefully designed cross-modal interaction is a powerful tool for learning semantic similarity, providing a robust component for next-generation multimodal AI systems.

References

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1993).
Signature verification using a "siamese" time delay neural network.
In *Advances in Neural Information Processing Systems*, volume 6, pages 737–744.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015).
Siamese neural networks for one-shot image recognition.
In *Proceedings of the ICML Deep Learning Workshop*.

Liu, X. et al. (2021).
Wikidiverse: A multimodal dataset for knowledge-base entity linking.
arXiv preprint arXiv:2104.05127.

Nag, R. (2022).
A comprehensive guide to siamese neural networks.
Accessed: 2025-06-22.

Thank you!