Mittuniversitetet
MID SWEDEN UNIVERSITY

Analysis of Inter-LLM and Intra-LLM Alignment in Image Description and Generation

Mid Sweden University
Quantitative Research and Development

Author: Saverio Napolitano

2025/2026

Overview

- Transformer
  - Attention
- Large Language Model (LLM)
  - Multimodal Large Language Model
    - Contrastive Image-Language Pre-training (CLIP)
  - Large Language Model for Image Captioning
  - Large Language Model for Image Generation

Transformer

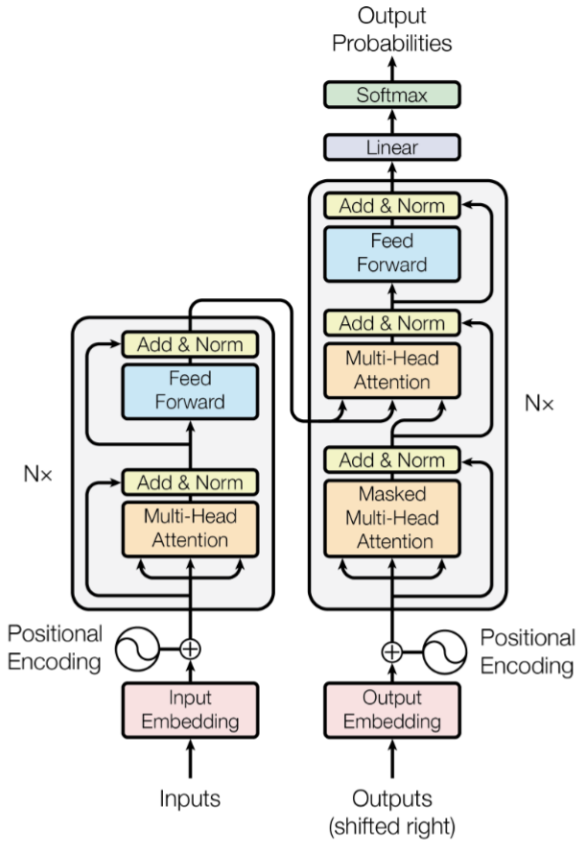

Figure: The Transformer - model architecture. Reprinted from Figure 1 in Vaswani et al. (2017) [1].

Attention

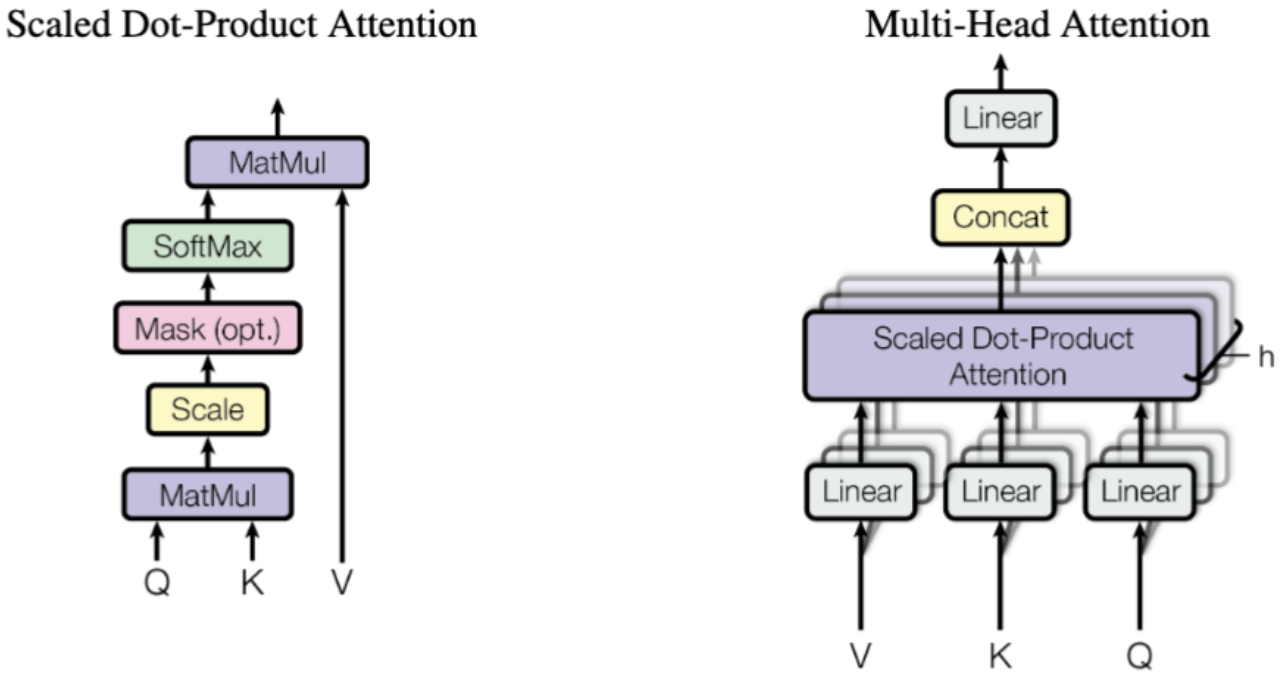

Figure: (left) Scaled Dot-Product Attention (right) Multi-Head Attention consists of several attention layers running in parallel. Reprinted from Figure 2 in Vaswani et al. (2017) [1].

Large Language Model (LLM)

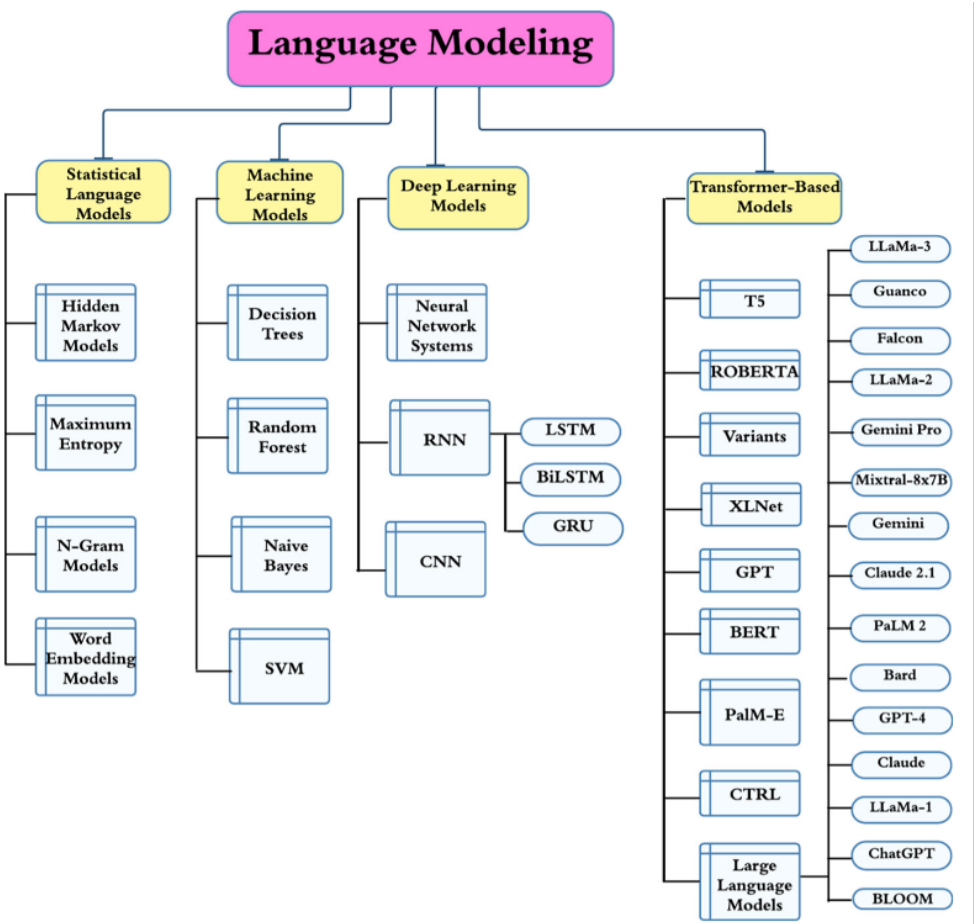

Figure: Types of language modeling. Reprinted from Figure 1 in Hadi et al. (2025) [2].

Large Language Model (LLM)

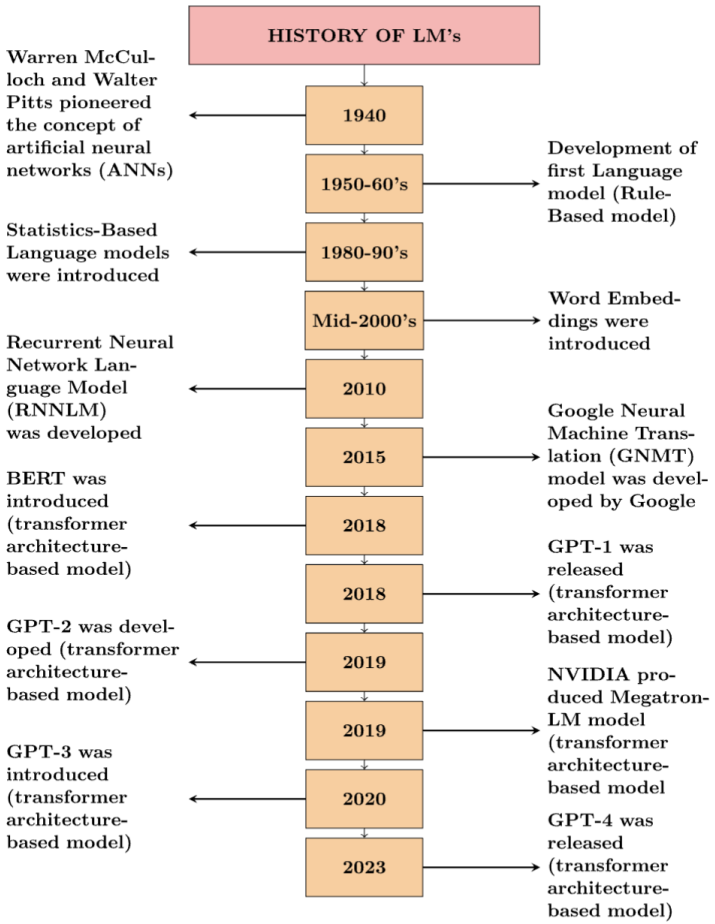

Figure: Brief history of language modeling. Reprinted from Figure 3 in Annepaka et al. (2025) [3].

Large Language Model (LLM)

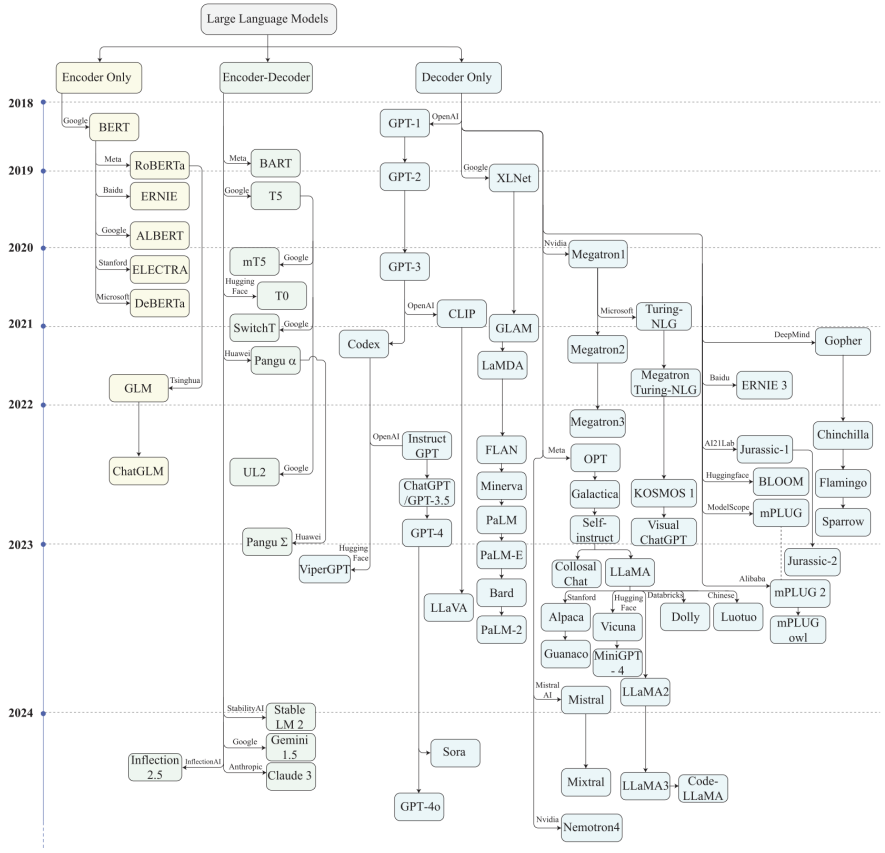

Figure: Evolutionary tree illustrating the progression of mainstream LLMs. Reprinted from Figure 12 in Shao et al. (2024) [4].

Large Language Model (LLM)

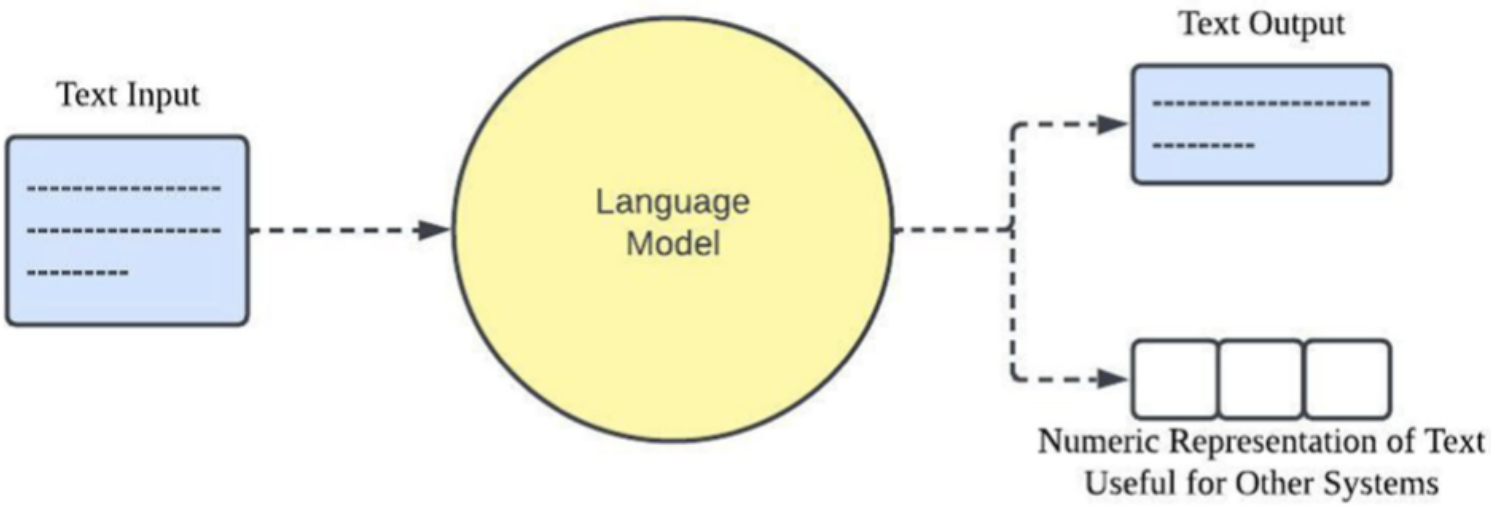

Figure: Workflow of a language model. Reprinted from Figure 2 in Annepaka et al. (2025) [3].

Multimodal Large Language Model

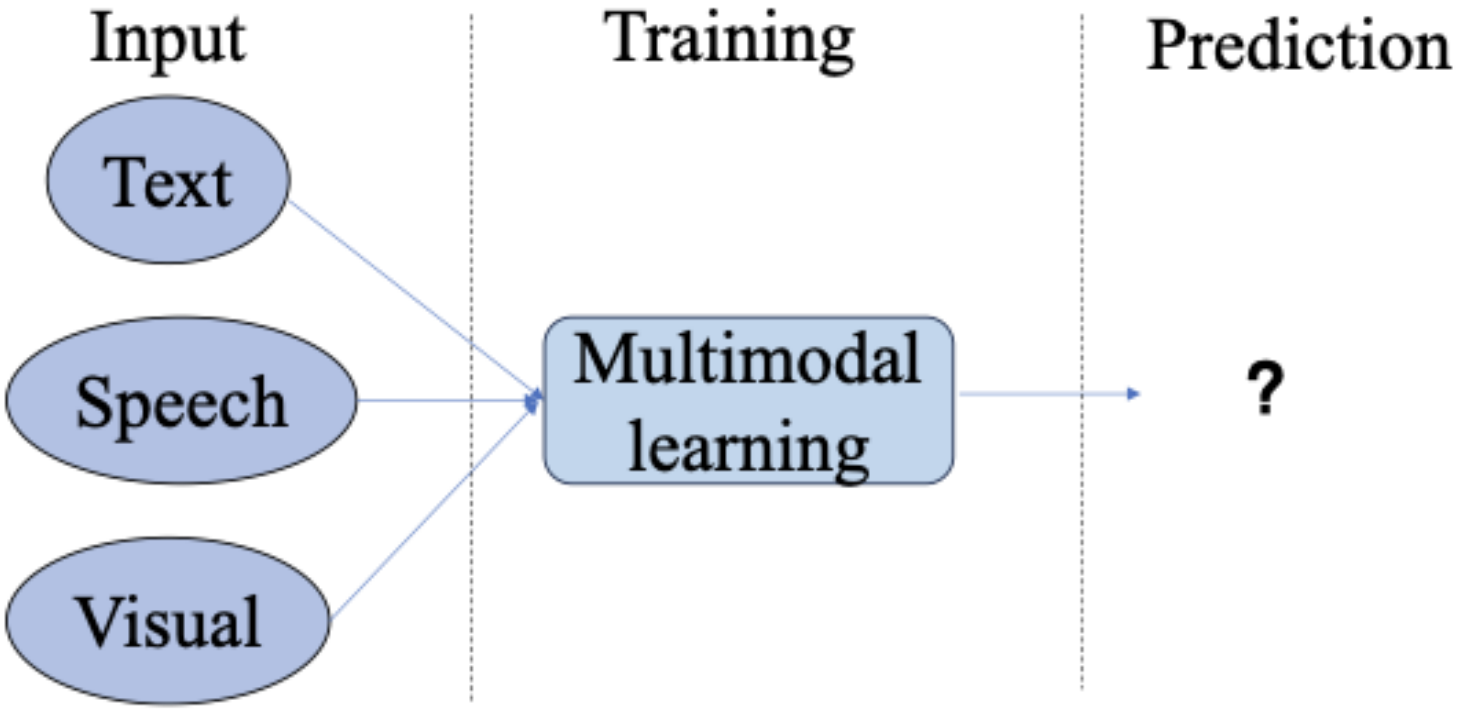

Figure: Definition of multimodal. Reprinted from Figure 1 in Wu et al. (2023) [5].

Contrastive Language Image Pre-training (CLIP)

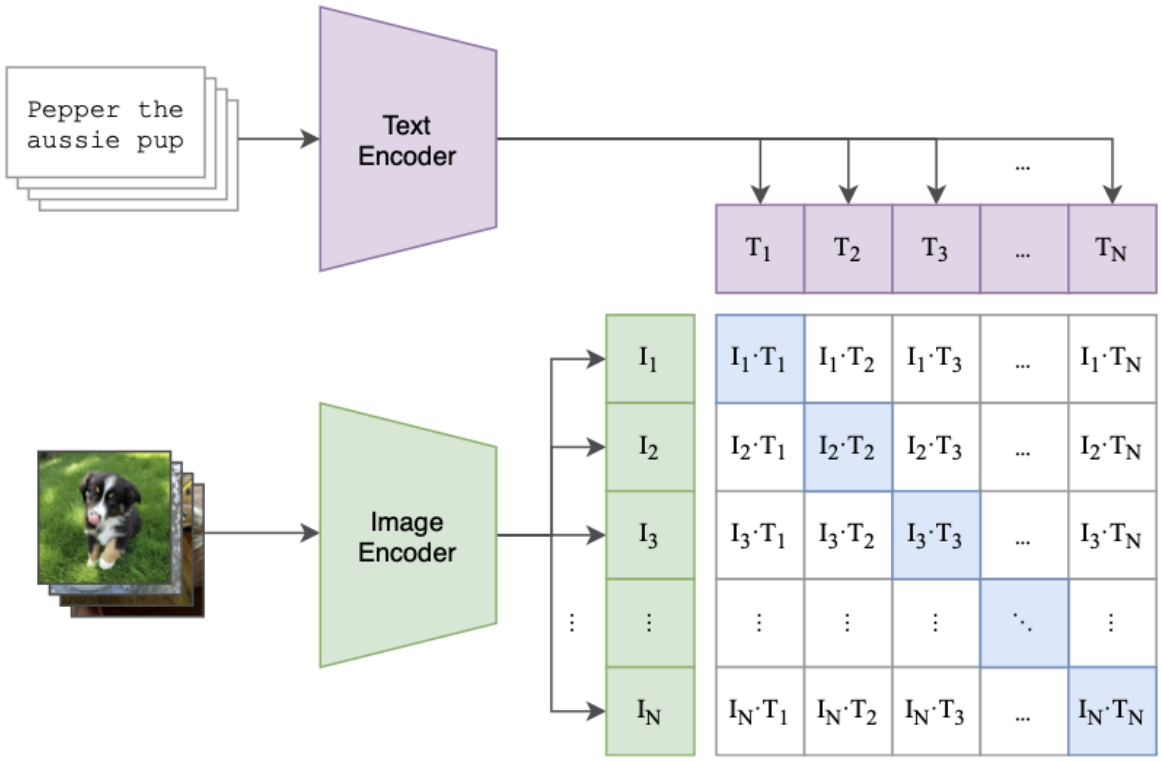

Figure: CLIP architecture and training. Reprinted from Figure 1 in Radford et al. (2021) [6].

Large Language Model for Image Captioning

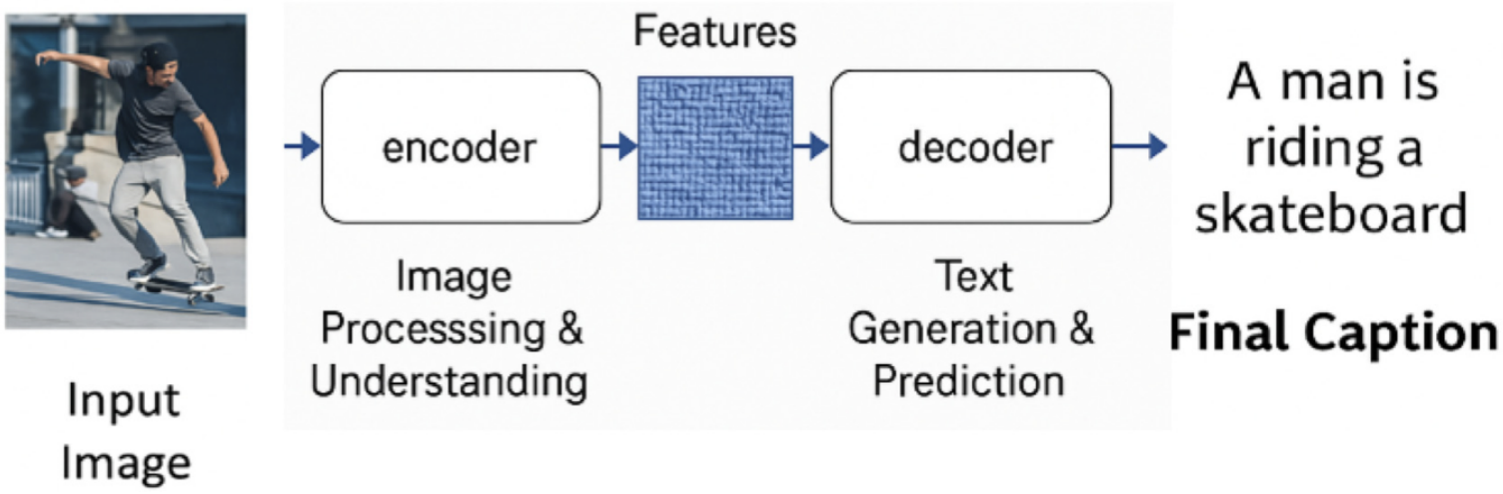

Figure: The general structure framework for image captioning. Reprinted from Figure 1 in Abdulgalil et al. (2025) [7].

Large Language Model for Image Captioning

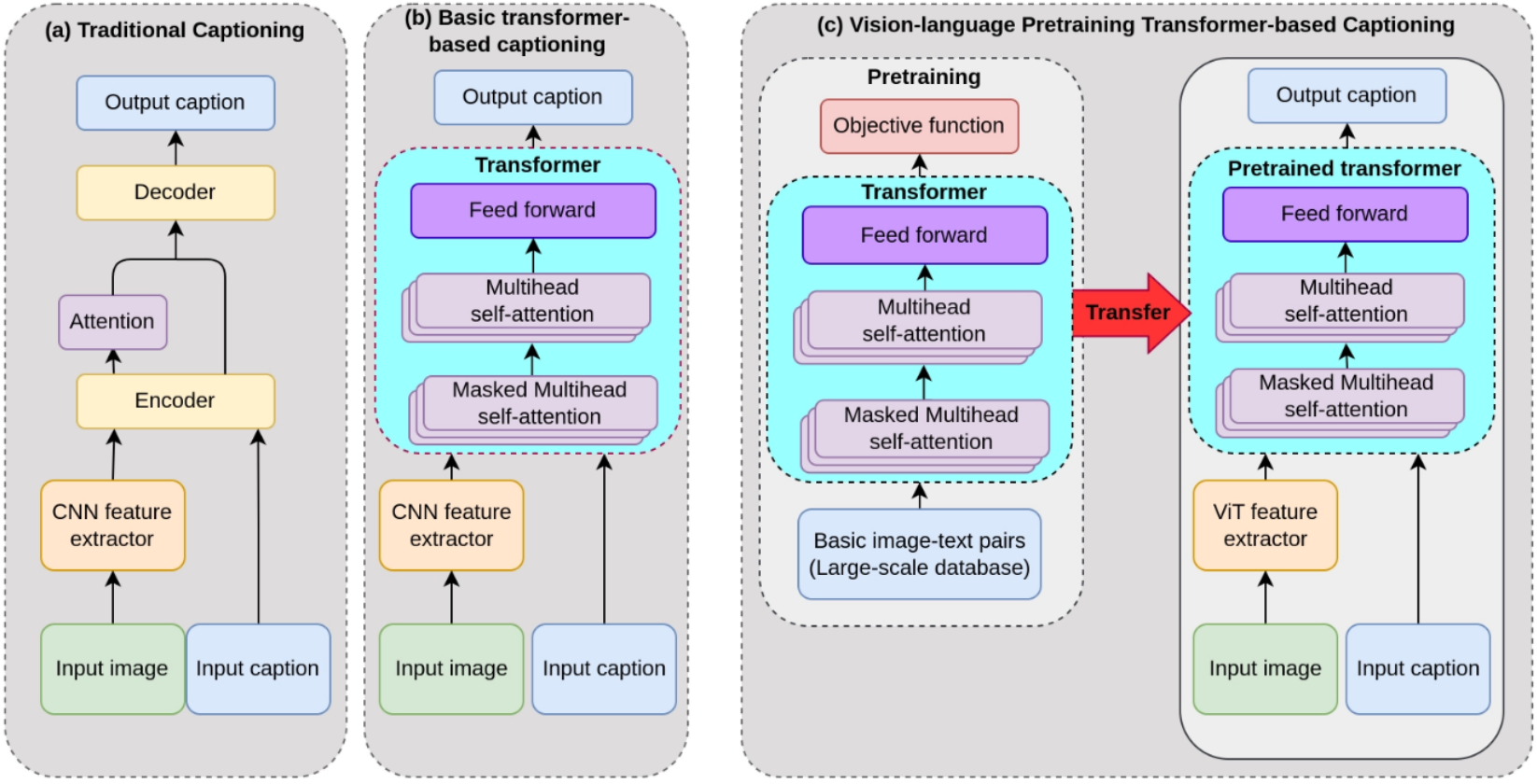

Figure: Development of captioning structures and systems. Reprinted from Figure 1 in Ondeng et al. (2023) [8].

Large Language Model for Image Generation

Figure: Visual comparison of images generated with different text encoders. Reprinted from Figure 1 in Wang et al. (2025) [9].

Bibliography


Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In *Advances in Neural Information Processing Systems*, volume 30, 2017.


Muhammad Usman Hadi, Qasem Al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Syed Zohaib Hassan, Maged Shoman, Jia Wu, Seyedali Mirjalili, and Mubarak Shah.
Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects.
February 2025.


Yadagiri Annepaka and Partha Pakray.
Large language models: a survey of their development, capabilities, and applications.
Knowledge and Information Systems, 67(3):2967–3022, March 2025.


Minghao Shao, Abdul Basit, Ramesh Karri, and Muhammad Shafique.
Survey of different large language model architectures: Trends, benchmarks, and challenges.
IEEE Access, 12:188664–188706, 2024.

Bibliography


Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu.
Multimodal large language models: A survey.
In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256, 2023.


Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
Learning transferable visual models from natural language supervision.
In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.


Huda Diab Abdulgalil and Otman A. Basir.
Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to multimodal large language models.
Natural Language Processing Journal, 12:100159, 2025.


Oscar Ondeng, Heywood Ouma, and Peter Akuon.
A review of transformer-based approaches for image captioning.
Applied Sciences, 13(19), 2023.

Bibliography

Andrew Z. Wang, Songwei Ge, Tero Karras, Ming-Yu Liu, and Yogesh Balaji.
A comprehensive study of decoder-only llms for text-to-image generation.
In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28575–28585, 2025.