Conceptual Parallels Between Capsule Networks and GPT:A Deep Technical-Conceptual Analysis

Mohammad Reza Besharati * , Mohammad Izadi , Nafiseh Jafari

Posted Date: 15 October 2025

doi: 10.20944/preprints202510.1210.v1

Keywords: capsule networks; generative pre-trained transformers; deep learning; conceptual analysis

 doi:10.20944/preprints202510.1210.v1

Article

Conceptual Parallels Between Capsule Networks and GPT: A Deep Technical-Conceptual Analysis

Mohammad Reza Besharati [1,*], Mohammad Izadi [1] and Nafiseh Jafari [2]

[1] Distributed and Multiagent Systems Lab (DiSysLab), Computer Engineering Department, Sharif University of Technology, Tehran, Iran

[2] University of Qom, Qom, Iran

* Correspondence: mohammad.besharati11@sharif.edu

Abstract

Capsule Networks (CapsNets) and Generative Pre-trained Transformers (GPT) represent landmark advancements in deep learning architectures, originating from distinct theoretical motivations and targeting different problem domains. Capsule Networks were devised to preserve hierarchical spatial relationships by grouping neurons into vector-outputting capsules that model instantiation parameters of entities. GPT models, built on the Transformer architecture, utilize powerful self-attention mechanisms applied to contextual token embeddings to model complex, long-range dependencies in sequential data such as language. This paper furnishes an in-depth comparative analysis of the conceptual and technical underpinnings of these architectures beyond mere terminology, unpacking their representation mechanisms, consensus-building dynamics, hierarchical information flow, and entity encoding. Through this synthesis, the paper elucidates shared foundational principles and key divergences, enriching theoretical insights and suggesting avenues for integrative architectures.

Keywords: capsule networks; generative pre-trained transformers; deep learning; conceptual analysis

1. Introduction

Neural networks have evolved from scalar-output perceptrons to rich vectorial and contextual embedding-based systems. Capsule Networks, introduced by Hinton et al. in 2017, mark an important conceptual shift emphasizing **vectorial representation of entities** encoding not only the presence but also spatial and pose parameters, pursued through an iterative routing-by-agreement mechanism which dynamically routes information to higher-level capsules based on consensus among lower-level predictions [Hinton et al., 2017]. Independently, the Transformer architecture, which forms the backbone of GPT models developed by OpenAI starting in 2018, revolutionized sequence modeling by eliminating recurrent constraints in favor of powerful multi-head self-attention layers that selectively attend to contextual tokens across long distances [Vaswani et al., 2017; Radford et al., 2018-2023].

Despite their independent origins and domain-specific design goals—CapsNets focusing on computer vision and hierarchical part-whole relationships, and GPT on natural language understanding and generation—both architectures share profound architectural philosophies centered on **vector-based entity encoding**, **dynamic context-dependent routing/attention**, and **hierarchical representation learning**. This paper provides a technical exposition and synthesis of these conceptual parallels and operational differentiators.

 doi:10.20944/preprints202510.1210.v1

2. Historical and Theoretical Background

2.1. Early Developments in Neural Representation

The journey from early artificial neural networks (ANNs) started in the 1940s and 1950s with McCulloch-Pitts neurons and perceptrons, which computed scalar outputs reflecting simple pattern recognition [Rosenblatt, 1958]. While capable of limited tasks, this scalar paradigm struggled with complex hierarchical object representation. Hinton's vision shifted towards using **multi-dimensional vectors**, recognizing that objects and their parts are more richly characterized by multiple instantiation parameters (pose, orientation, deformation) rather than binary presence.

Transformers emerged from the broader evolution of sequence models away from recurrent units (LSTMs, GRUs) towards architectures leveraging **pure attention** for parallelizable and scalable long-range dependency modeling [Vaswani et al., 2017]. The conceptual leap was to treat tokens as embedded vectors representing semantic and syntactic attributes contextualized by dynamically computed weighted averages over all tokens.

2.2. Capsule Networks Development

Hinton introduced Capsule Networks in 2017 as a response to limitations of convolutional neural networks (CNNs), particularly their pooling operations which discard important spatial hierarchical relations [Hinton et al., 2017]. Capsules are neuron groups outputting vectors representing entity instantiations. The **dynamic routing algorithm** routes output from lower-level capsules to higher-level capsules that agree, operationalized by the scalar product (dot product) between predicted and actual capsule outputs, refining these assignments iteratively.

CapsNets demonstrate state-of-the-art results on benchmarks like MNIST and show robustness in recognizing overlapping and spatially transformed digits, affirming the importance of vectorial pose-aware hierarchical representations.

2.3. Transformer and GPT Architecture

Transformers introduced **self-attention**, calculating pairwise similarity between token embeddings via scaled dot-product attention, thereby reweighting information flow adaptively across tokens for each layer [Vaswani et al., 2017]. GPT models build upon this by stacking multiple Transformer decoder blocks with pre-training on massive corpora to learn generalized language representations capable of zero-shot and few-shot tasks [Radford et al., 2018-2023].

The **token embeddings** represent discrete words/subwords as continuous vectors encoding rich semantic and syntactic properties without explicitly modeling spatial or pose features.

3. Representation Mechanisms: Vectorial Entity Encoding

3.1. Capsules: Activity Vectors as Instantiation Parameters

Capsules output vectors whose **length** encodes the likelihood of the entity's existence and **orientation** encodes attributes like pose and deformation. This vectorial output is fundamentally different from scalar activations in CNN neurons and focuses on preserving detailed instantiation information [Hinton et al., 2017].

3.2. GPT Token Embeddings

In GPT, tokens are embedded into fixed-dimensional continuous vector spaces which carry latent semantic and syntactic features. These embeddings serve as the starting point for multi-layer transformations wherein attention mechanisms modulate inter-token relationships [Vaswani et al., 2017].

Though not explicitly geometric, these embeddings can represent nuanced linguistic properties such as word sense, syntactic roles, and contextual nuances.

 doi:10.20944/preprints202510.1210.v1

3.3. Summary

Both architectures emphasize the importance of **multi-dimensional vector representations** as the fundamental computational unit, enabling richer encoding of complex entity states—whether that be an object's pose or a token's contextual meaning.

4. Information Routing and Consensus Mechanisms

4.1. Dynamic Routing-by-Agreement in Capsule Networks

Capsules at one level predict outputs for capsules at the next level using learned transformation matrices. A higher-level capsule activates based on agreement among these predictions measured by scalar products (dot products) [Hinton et al., 2017]. This **routing-by-agreement** is an iterative, bottom-up process enforcing part-whole consistency.

4.2. Self-Attention in GPT

In contrast, GPT implements a continuous **self-attention mechanism** where each token forms queries (Q), keys (K), and values (V). Attention weights are computed via softmax normalized scaled dot-products [Vaswani et al., 2017]:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

This computes a weighted average of values based on pairwise similarity, dynamically selecting contextually pertinent information.

4.3. Conceptual Parallels

Both processes use **vector similarity (scalar product)** as a measure of agreement to decide how information is propagated or aggregated. Capsules employ discrete routing refinement emphasizing strict consensus; Transformers perform soft weighting suitable for distributed contextual blending.

5. Hierarchical Information Flow and Abstraction

5.1. Capsule Network Hierarchies

CapsNets hierarchically build from low-level part capsules detecting edges or simple features, to higher-level capsules representing aggregated objects with computed poses [Hinton et al., 2017]. Each layer refines and abstracts the representation, enforcing spatial consistency.

5.2. Multi-Layer Transformer Representations

GPT's multiple stacked Transformer blocks progressively re-encode token embeddings into more abstract semantic representations, capturing linguistic structures and world knowledge [Rachtibat, 2023]. Each layer reweighs token contribution via self-attention, iteratively refining context representation.

5.3. Comparative Insight

Despite target data differences—images for Capsules and sequences for GPT—both architectures realize information processing as **multi-layer hierarchical refinement** producing increasingly global and abstract entity representations.

 doi:10.20944/preprints202510.1210.v1

6. Limitations and Challenges

6.1. Capsule Network Limitations

CapsNets face significant computational overhead due to iterative routing, limiting scalability to large datasets or complex inputs [Sabour et al., 2017; LaLonde & Bagci, 2018]. Additionally, explicit pose encoding limits generalization to tasks not involving geometric variability.

6.2. GPT Limitations

While scalable and versatile, GPT models require enormous training data and compute resources. Attention is a powerful but sometimes opaque mechanism, leading to challenges in interpretability and reasoning consistency [Bender et al., 2021].

7. Discussion: Towards Hybrid Models

Recent research explores combining CapsNet's geometry-aware routing with attention's scalability, such as transformer-based capsule routing systems. These hybrids aim to integrate **pose-aware hierarchical encodings** with **contextualized adaptive attention** [Khodadadzadeh et al., 2021].

8. Conclusion

Capsule Networks and GPT exemplify two complementary paradigms in modern deep learning architecture design. Both move beyond scalar neuron activations to sophisticated **vectorial representations** modulated by **agreement-based dynamic routing or attention**. Understanding these architectures as part of a unifying conceptual framework offers fertile ground for future integrative advances in representation learning.

References

1. Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic routing between capsules." Advances in neural information processing systems 30 (2017).
2. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).
3. Radford, Alec, Rafal Jozefowicz, and Ilya Sutskever. "Learning to generate reviews and discovering sentiment." (2018).
4. Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." OpenAI blog 1, no. 8 (2019): 9.
5. Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018): 3.
6. Radford, A., et al. (2018–2023). Various publications on GPT architectures. OpenAI.
7. Hinton, Geoffrey, Yoshua Bengio, Demis Hassabis, Sam Altman, D. Amodei, D. Song, T. Lieu, B. Gates, Y. Q. Zhang, and I. Sutskever. "Statement on AI risk." Center for AI safety 31, no. 5 (2023): 2023.
8. Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?." In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 610-623. 2021.
9. Khodadadzadeh, Massoud, Xuemei Ding, Priyanka Chaurasia, and Damien Coyle. "A hybrid capsule network for hyperspectral image classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2021): 11824-11839.
10. Huang, Zhongzhan, Mingfu Liang, Jinghui Qin, Shanshan Zhong, and Liang Lin. "Understanding self-attention mechanism via dynamical system perspective." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1412-1422. 2023.

 doi:10.20944/preprints202510.1210.v1

11. Achtibat, Reduan, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. "Attnlrp: attention-aware layer-wise relevance propagation for transformers." arXiv preprint arXiv:2402.05602 (2024).
12. Jafari, Nafiseh, Mohammad Reza Besharati, and Maryam Hourali. "SELM: Software Engineering of Machine Learning Models." In New Trends in Intelligent Software Methodologies, Tools and Techniques, pp. 48-54. IOS Press, 2021.