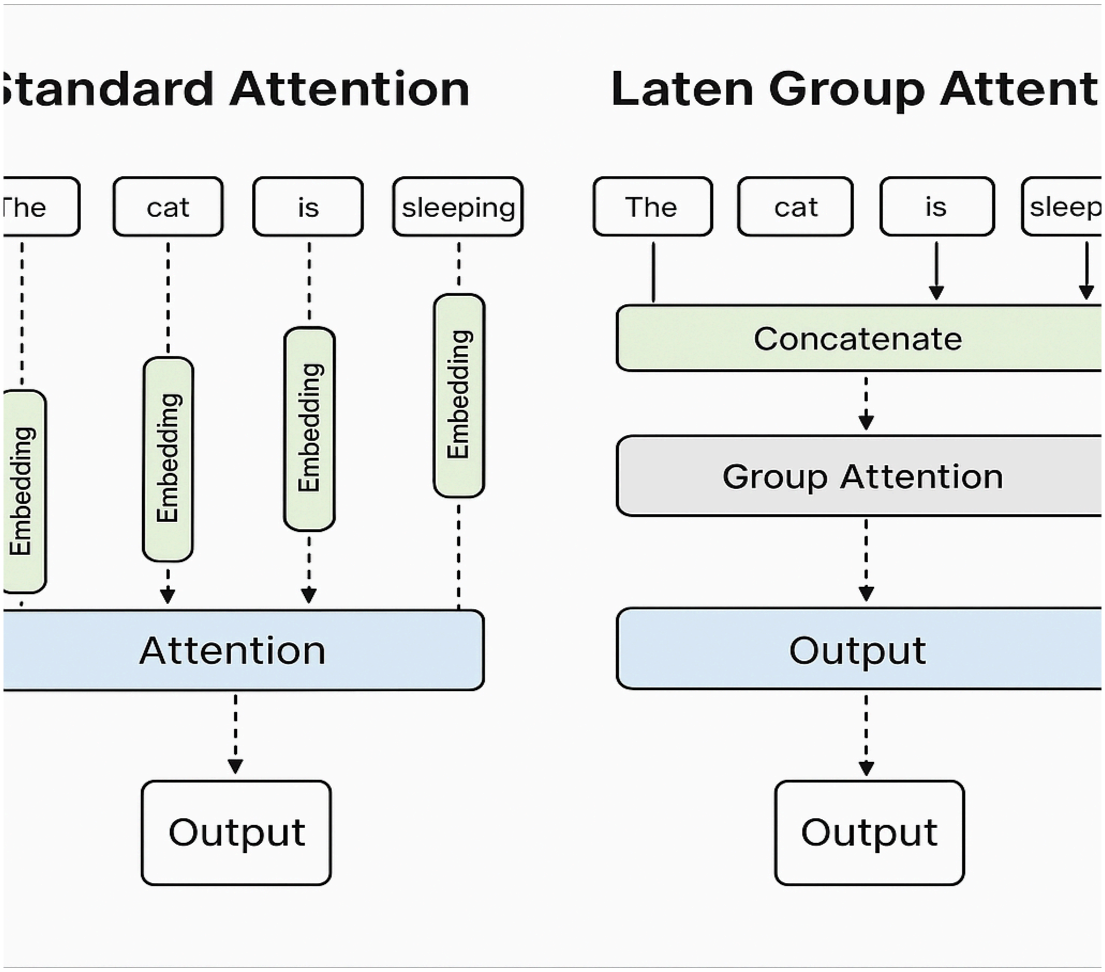
Title:

Dense Group Attention: A Hypothesis on Local Contextual Embedding through Structured Word Concatenation

Jubo Zhang
digbigAI, Independent Research Lab, Newark,CA
jerryzhang668@gmail.com

Abstract:

We hypothesize that *Dense Group Attention (DGA)*—a novel mechanism for forming local contextual embeddings in transformer-based language models—can improve the model's ability to capture short-range dependencies and phrase-level semantics. Unlike standard attention mechanisms that operate over isolated token embeddings, DGA constructs composite embeddings by concatenating a token's embedding with those of its immediate successors, forming dense group representations of size *g*. We propose that this structured embedding strategy enhances early-stage contextual awareness, potentially reducing the depth or complexity needed in downstream attention layers. This paper outlines the conceptual framework, motivations, and expected implications of Dense Group Attention, along with proposed methodologies for empirical validation.



1. Introduction

The Transformer architecture has revolutionized natural language processing (NLP), particularly through its use of self-attention to model long-range dependencies [1]. Yet, challenges remain in capturing rich, localized meaning early in the network pipeline. Attention heads often struggle to prioritize small, meaningful token clusters (e.g., verb phrases, idioms) without extensive training or stacking multiple layers [2].

This paper proposes Dense Group Attention (DGA) as a pre-attention embedding strategy designed to inject localized compositional structure into the model from the outset. Our hypothesis is that by representing tokens as group embeddings — formed by concatenating a token's embedding with those of a fixed number of following tokens — we can enrich local semantic context and reduce redundancy in attention computations [3].

2. Hypothesis

We hypothesize that Dense Group Attention improves language modeling performance by explicitly embedding local token groups, thus reducing the burden on attention layers to infer local structure.

Specifically:

- **H1**: Dense group embeddings encode richer local context, enabling more accurate predictions in tasks involving short-range semantic dependencies (e.g., sentiment analysis, NER, syntactic parsing) [4].

- **H2**: DGA improves sample efficiency, requiring fewer training iterations to reach a given performance level on sequence-based NLP tasks.

- **H3**: DGA can reduce the required attention depth or number of heads, as some local dependencies are pre-encoded in the embeddings [5].

---

3. Methodology

3.1 Dense Group Embedding Construction

Given a sequence of token embeddings `[E1, E2, ..., En]`, for a group size *g*, each position *i* is represented as:

$$\mathbf{G_i = [\ E_i\ ;\ E_{i+1}\ ;\ \dots\ ;\ E_{i+g-1}\ ]}$$

where “;” denotes vector concatenation.
Padding is used for final positions to ensure consistent dimensionality [6].

3.2 Model Integration

These group embeddings replace or augment standard token embeddings in transformer-based models. The attention mechanism remains unchanged but now operates on higher-dimensional input vectors of size *g* × *d*, where *d* is the original embedding size.

Two configurations are proposed:

- **Full Replacement**: Replace all input embeddings with group embeddings.

- **Hybrid Embedding**: Concatenate group and original embeddings, then project.

3.3 Experimental Setup

We propose testing the DGA hypothesis across multiple benchmarks:

- **GLUE**: For general language understanding
- **SQuAD**: For span-based question answering
- **Penn Treebank**: For syntactic structure and language modeling [7]

Baseline models will be standard Transformers, with comparable parameter budgets.

Metrics to evaluate include:

- Accuracy / F1 Score
- Training speed (convergence rate)
- Model size vs. performance trade-off

---

4. Expected Benefits

- **Stronger Local Semantics**: Phrase- and clause-level meaning captured directly in embeddings.
- **Shallower Networks**: Reduced reliance on multi-head, deep-layer modeling of local structure [5].
- **Data Efficiency**: Faster convergence due to enriched early-layer input representations [8].

---

5. Limitations and Challenges

- **Dimensional Explosion**: For large *g*, embeddings grow linearly, increasing memory requirements.
- **Overfitting to Local Patterns**: May bias model toward short-range dependencies at the expense of long-range reasoning.
- **Padding Effects**: Terminal sequence tokens require handling to avoid padding artifacts [6].

6. Related Work

DGA shares conceptual lineage with:

- **Convolutional Neural Networks (CNNs)**: which use local receptive fields to capture spatial structure [9].
- **Relative Positional Embeddings**: which encode local proximity explicitly [4].
- **Chunk-based Models**: which group tokens by syntactic or semantic units (e.g., phrases, subwords).

However, DGA is unique in its use of *dense, non-overlapping forward concatenation* to form each token's contextual representation, operating purely at the embedding level.

Comparison with Related Techniques

Technique	Context Method	Local Pattern Focus	Memory Cost	Positional Bias
Standard Attention	Individual token attention	Moderate	Moderate	Positional encodings
Convolution + Attention	Sliding window filters	High	Low	Implicit
Sparse Attention	Limited context windows	Varies	Low	Structured bias
Dense Group Attention	Concatenated local embeddings	Very High	High	Explicit grouping

7. Conclusion and Future Work

We present **Dense Group Attention** as a structured embedding technique that may enhance the efficiency and expressiveness of language models by encoding local context early in the pipeline. We hypothesize that this method improves local pattern recognition, sample efficiency, and may reduce downstream model complexity. Future work will involve extensive empirical validation, dynamic group sizing mechanisms, and potential integration with attention sparsification techniques for scalability.

References

[1] A. Vaswani et al. Attention Is All You Need. In *NeurIPS*, 2017.

[2] J. Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.

[3] Y. Tay et al. Efficient Transformers: A Survey. In *ACM CSUR*, 2022.

[4] Z. Dai et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. In *ACL*, 2019.

[5] I. Tenney et al. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. In *ICLR*, 2019.

[6] A. Baevski and M. Auli. Adaptive Input Representations for Neural Language Modeling. In *ICLR*, 2019.

[7] A. Wang et al. GLUE: A Multi-Task Benchmark and Analysis Platform for NLP. In *ICLR*, 2019.

[8] K. Clark et al. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*, 2020.

[9] Y. Kim. Convolutional Neural Networks for Sentence Classification. In *EMNLP*, 2014.