Title:
Word Compounding Layers: A Hypothesis on Efficient Local Semantic Grouping for Language Models

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

Abstract:
This paper introduces the hypothesis that *Word Compounding Layers (WCL)*, a technique for selectively merging semantically coherent word groups using a lightweight auxiliary model, can improve the computational efficiency and contextual awareness of large language models. We propose replacing Dense Group Attention — a method that concatenates fixed local token embeddings — with a more targeted approach that identifies and merges true linguistic compounds (e.g., verb groups, idiomatic phrases) while preserving fine-grained details (e.g., adjectives). This is achieved by training a separate, small compounding model to detect meaningful token groupings and then integrating its learned behavior into the early layers of a larger transformer model. We hypothesize that this technique reduces redundancy, preserves semantic precision, and improves training and inference efficiency without sacrificing performance.

---

1. Introduction

Modern language models such as BERT, GPT, and T5 rely on token-level embeddings and self-attention mechanisms to capture context and meaning. Techniques like Dense Group Attention (DGA) have been proposed to enrich local semantics by concatenating multiple adjacent token embeddings. While effective at capturing short-range dependencies, DGA increases input dimensionality linearly with group size, raising memory and compute costs.

In this paper, we propose an alternative hypothesis: Word Compounding Layers (WCL) can provide the benefits of DGA while significantly improving efficiency. WCL uses a trained auxiliary model to identify linguistically motivated word compounds—e.g., phrasal verbs, multi-word expressions, or noun phrases—and merges them at the embedding level. These compounds are treated as atomic units in the larger model, reducing sequence length and computational load while preserving critical semantic distinctions [1,2].

---

2. Hypothesis

We hypothesize that Word Compounding Layers improve language model efficiency and semantic resolution by selectively grouping only those tokens that form coherent linguistic units.

Specifically:

- **H1**: Word Compounding Layers can reduce token sequence length without loss of semantic granularity by merging only contextually appropriate word groups.

- **H2**: Integrating WCL into the early embedding layers of a large model results in more efficient training and inference compared to Dense Group Attention.

- **H3**: Semantic distinctions—such as differences between “red apple” and “green apple”—are better preserved by WCL than by fixed-window grouping strategies like DGA [3].

---

3. Methodology

3.1 Compounding Model Design

We propose training a small model (e.g., BiLSTM-CRF, lightweight Transformer, or rule-enhanced parser) on either supervised compound annotations or unsupervised objectives such as compositional similarity and masked compound recovery [4].

The model learns to segment input sequences into compound units based on linguistic features such as:

- Grammatical function (e.g., auxiliary verbs)

- Semantic coherence (e.g., idioms, collocations)

- Syntactic dependency (e.g., verb-object)

3.2 Integration with Large Language Models

Once trained, the compounding model is used to identify compound units in tokenized input. These units are then:

- **Merged at the embedding level** (via mean, weighted sum, or learned transformation) [5]

- **Encoded into the vocabulary**: Frequently occurring compounds may be assigned unique token IDs, similar to byte-pair encoding [6]

This segmentation process is integrated into the preprocessing and early embedding layers of a Transformer model.

3.3 Preservation of Semantic Detail

WCL's selective grouping strategy ensures that:
✅ "has been running" → **has_been_running**
❌ "red apple" remains **separate** from "green apple"

This preserves semantic contrast between adjacent phrases, addressing one of the key limitations of DGA [3].

---

4. Comparison with Dense Group Attention

Feature	Dense Group Attention	Word Compounding Layers
Embedding Dimensionality	Increases with group size	Constant per compound unit
Grouping Strategy	Fixed-size window (g tokens)	Learned, variable-length compounds
Token Reduction	None	Sequence length may be reduced
Linguistic Awareness	None (position-based)	High (data-driven and/or rule-based)
Semantic Precision	May blur modifier meaning	Preserves modifiers like adjectives

Efficiency	Lower	Higher

WCL addresses DGA's main limitations by introducing content-aware grouping that can reduce overhead while improving representation quality [2].

---

5. Evaluation Plan

5.1 Benchmarks

- GLUE and SuperGLUE: general language understanding
- SQuAD: reading comprehension
- CoNLL-2003: named entity recognition
- OpenWebText: language modeling

5.2 Metrics

- Accuracy, F1, BLEU, Perplexity
- Training speed (epochs to convergence)
- Memory usage, FLOPs per sequence

Comparison groups:

1. Baseline transformers
2. DGA-enhanced models
3. WCL-enhanced models

---

6. Related Work

WCL draws on prior work in:

- **Multi-word expression detection** [4]
- **Chunk-based and phrase embeddings** [5]
- **Efficient Transformer models**, including sparse and adaptive mechanisms [7]

However, WCL's key novelty lies in externalizing compound detection to a separate module and incorporating this during early model stages.

---

7. Conclusion

We propose Word Compounding Layers as a hypothesis-driven mechanism to improve the efficiency and precision of large language models. By using a compact model to detect and merge meaningful token groups, WCL promises a leaner architecture that retains semantic clarity. Future research will evaluate this approach empirically and explore its broader utility across language domains.

References

[1] S. Swayamdipta et al. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *EMNLP*, 2020.

[2] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *ACL*, 2019.

[3] I. Tenney et al. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR*, 2019.

[4] C. Ramisch. Multiword Expressions Acquisition: A Generic and Open Framework. *Springer*, 2015.

[5] W. Yin and H. Schütze. Learning Word Meta-Embeddings. In *ACL*, 2016.

[6] R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. In *ACL*, 2016.

[7] A. Fan et al. Reducing Transformer Depth on Demand with Structured Dropout. In *ICLR*, 2020.