Towards Equipping Transformer with the Ability of Systematic Compositionality

Chen Huang, Peixin Qin, Wenqiang Lei*, Jiancheng Lv

College of Computer Science, Sichuan University, Chengdu, China
Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu, China
{huangc.scu, qinpeixin.scu, wenqianglei}@gmail.com

Abstract

One of the key factors in language productivity and human cognition is the ability of *systematic compositionality*, which refers to understanding composed unseen examples of seen primitives. However, recent evidence reveals that the Transformers have difficulty generalizing the composed context based on the seen primitives. To this end, we take the first step to propose a compositionality-aware Transformer called CAT and two novel pre-training tasks to facilitate systematic compositionality. We tentatively provide a successful implementation of a multi-layer CAT on the basis of the especially popular BERT. The experimental results demonstrate that CAT outperforms baselines on compositionality-aware tasks with minimal impact on the effectiveness on standardized language understanding tasks.

Introduction

Linguistic research confirms the discreteness of linguistic symbols and their compositionality to construct larger linguistic expressions (Montague 1970; Frege 1948; Baroni 2020; Akyürek and Andreas 2022). These characteristics are known as *Systematic Compositionality* (Fodor and Pylyshyn 1988; Keysers et al. 2019; Lake et al. 2017). For instance, sentences are built from words and phrases. Such systematic compositionality fosters humans ability to understand and generalize to unseen combinations of seen primitives (Lake et al. 2017) and model complex phenomena (Liu et al. 2021). Therefore, it is widely recognized as an essential capability of human intelligence (Ma, Zhang, and Zhu 2023).

However, previous studies have shown that language models struggle with generalizing through composition (Cartuyvels, Spinks, and Moens 2021; Lake and Baroni 2018; Loula, Baroni, and Lake 2018). Even for large language models (LLMs), recent evidence suggests that they still struggle to establish systematic compositionality after fine-tuning on compositionality-aware datasets (Yu and Ettinger 2021) or prompting with in-context examples (An et al. 2023). The challenge lies in the fact that the semantics of a group of primitives vary depending on their meanings and how they are combined. Considering examples on word-level compositionality in Table 1 where words serve as the primitives of phrases, two phrases composed of different words may have similar semantics (e.g., *Safety officer* and *Security guard*), while phrases may exhibit different semantics through different combinations (e.g., *Water body* and *Body water*). Empirical evidence in Table 1 shows that even the 'omnipotent' ChatGPT still struggles to accurately capture the semantic changes among different word combinations, failing to align with human judgments regarding the similarity of phrase pairs. One possible explanation is that despite the ability to capture the meaning of words, current Transformer frameworks fail to develop systematic compositional skills (Dziri et al. 2023; Ma, Zhang, and Zhu 2023). In contrast, humans certainly do understand language by learning the meaning of words and composing more elaborate meanings (Cartuyvels, Spinks, and Moens 2021). Therefore, to achieve human-level language understanding, it is imperative to invest more effort in building a compositionality-aware model that promotes LLMs with stronger capabilities in systematic compositionality.

Phrase 1	Phrase 2	Similarity Score	
		LLM	**Human**
Safety officer	Security guard	0.6	0.881
Water body	Body water	0.7	0.381

Table 1: Compositionality example from Asaadi, Mohammad, and Kiritchenko (2019). ChatGPT struggles to achieve human-level language understanding as it fails to align with human judgments regarding the similarity of phrase pairs that are formed by different word composition.

Motivated by this, we take the first step to propose a Compositionality-Aware Transformer called **CAT** to facilitate the systematic compositionality, along with two novel pre-training tasks. 1) As depicted in Fig. 1, CAT introduces two modules: Multi-Primitive Composition and Representation Enhancement, to the vanilla Transformer encoder. These modules enable CAT to learn how to compose primitives and enhance the representation of the vanilla Transformer encoder, respectively. Specifically, the Multi-Primitive Composition module decomposes a contextual word representation[1] h_{cont} into multiple primitives, which

*Correspondence to Wenqiang Lei.

[1] Here, a word representation is the average of its corresponding

are represented as discrete latent space vectors. It then produces a compositional representation h_{comp}. The Representation Enhancement module integrates h_{comp} and h_{cont} to yield the final output h_{mix} with stronger systematic compositionality and without losing contextual information of h_{cont} required in downstream tasks. Notably, these modules can also be applied to the [CLS] token to achieve the semantic composition of a sentence. 2) Additionally, we propose two new pre-training tasks, i.e., Guided Decomposition and Semantics Composition, to further enhance the systematic compositionality. The former supervises the decomposition of h_{cont} using the OpenHowNet dataset (Qi et al. 2019b), which records the mappings from words to their corresponding primitives (i.e., Sememes[2]). The latter guides the composition of discrete primitives so that h_{comp} and h_{mix} are semantically informative without losing contextual information required in the downstream tasks. As such, by learning to compose primitives during pre-training, CAT could facilitate systematic compositionality of the vanilla Transformer.

As suggested in previous studies (Yu and Ettinger 2021; Cartuyvels, Spinks, and Moens 2021; Hendrycks et al. 2020; Yu and Ettinger 2020), our evaluation closely revolves around the characteristics of the systematic compositionality to provide detailed insights. Due to the issue of computational resources, we restrict our analysis and comparison to the widely used BERT and tentatively implement and pretrain a multi-layer CAT from scratch following BERT. Our experimental findings demonstrate that our approach outperforms BERT and other baselines on compositionality-aware tasks while having minimal impact on the effectiveness of standardized language understanding tasks. In comparison to BERT, CAT exhibits superior performance in identifying semantic changes in both phrase-level and sentence-level compositionality-aware tasks without losing any advantages in standardized GLUE tasks. On average, our method experiences performance gains of +6.42 and +1.10, respectively. Furthermore, CAT shows promising improvements in compositional generalization (+1.83) and robustness to noisy context (+3.09). Given that compositionality has always been considered a major factor in language productivity and human cognition, we believe our work is an important proof-of-concept for promoting LLMs' capability in systematic compositionality. In conclusion, we claim the following contributions.

- For the first time, we propose a systematic compositionality aware Transformer called CAT, which explicitly equips the vanilla Transformer with the ability to learn to compose primitives.
- We also propose two novel pre-training tasks to further facilitate the systematic compositionality of CAT.
- We verify our effectiveness with extensive empirical studies and offer an in-depth analysis. We provide insight for future studies on LLM's systematic compositionality.

token embeddings.

[2]Minimum semantic units of our languages. A set of discrete sememes (such as *Family*, *Spouse*, *Female*) could compose the meanings of all the words (*Wife*, in this case) (Qi et al. 2019a).

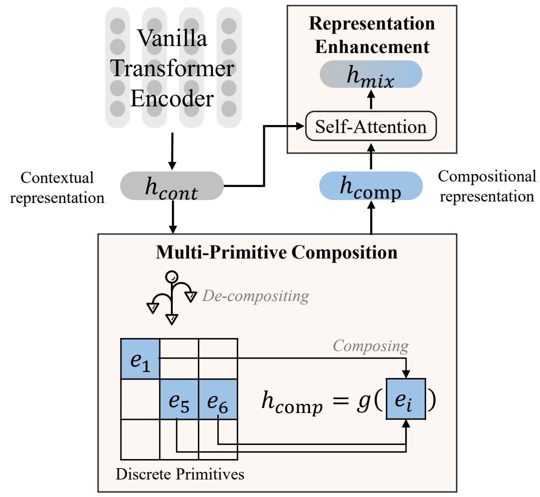

Figure 1: Illustration of CAT, which contains two modules. By learning to compose primitives, CAT facilitates systematic compositionality of the vanilla Transformer.

Related Work

Our focus is on the discreteness of linguistic symbols and their systematic compositionality. Therefore, we conduct a literature review on systematic compositionality and existing techniques for discretization & compositionality.

Systematic Compositionality. Systematic compositionality in the language is an attractive capability of combining discrete elementary units in systematic ways to create compound ones (Montague 1970; Frege 1948; Baroni 2020; Akyürek and Andreas 2022). It allows humans to make so-called "infinite use of finite means" (Chomsky 2014) and fosters the capacity of generalization (Lake et al. 2017). It has led to improved performance in various NLP tasks such as question answering (Bogin et al. 2021) and machine translation (Ataman and Federico 2018). In the era of LLMs, however, recent evidence shows that fine-tuning a given pre-trained model on a specific task may not improve its compositionality capability (Yu and Ettinger 2021). Over-parameterized LLMs, like ChatGPT, are still sensitive to the selection of in-context examples (An et al. 2023), making their compositionality ability fragile. More recently, researchers have found that current Transformers fail to develop systematic problem-solving skills (Dziri et al. 2023; Ma, Zhang, and Zhu 2023), which highlights the need for a compositionality-aware Transformer that may enhance LLMs' capabilities in systematic compositionality. To this end, we take the first step to propose a compositionality-aware Transformer and two novel pre-training tasks to facilitate systematic compositionality.

Techniques for Discretization & Compositionality. Exploring the compositional characteristics has drawn lots of attention, but they focus on composing continuous primitives. Notably, a continuous primitive is a continuous, real-valued variable that takes on values in connected regions of R^n. while a discrete primitive is a discretely valued variable that takes on either a limited or a countably infinite number of distinct values (Cartuyvels, Spinks, and Moens 2021). To compose continuous primitives, existing techniques include the group-equivariance theory (Higgins et al. 2018a; Gordon et al. 2019), syntactic attention (Russin et al. 2019; Li et al. 2019), disentangled representation (Burgess et al. 2018; Locatello et al. 2019), data argumentation (Akyürek and Andreas 2022; Andreas 2020), or neural modularization (Andreas et al. 2016). However, considering the discreteness of linguistic symbols, research on compositionality with discrete primitives is relatively rare. One possible reason for this is that optimizing discrete primitives during the backpropagation of a neural network in a stabilized and bias-free way is more challenging than optimizing contextual ones (Friede and Niepert 2021; Farinhas et al. 2021). Recent vector quantization techniques (van den Oord, Vinyals, and Kavukcuoglu 2017; Liu et al. 2021) propose using the Straight-Through-Estimator (Courbariaux, Bengio, and David 2015) to facilitate optimization on discrete codes/primitives. Nevertheless, the composition of discrete primitives is yet to be explored, let alone the pre-trained language model with systematic compositionality. As a result, it is unclear how to create a compositionality-aware language model and assess its performance on more general tasks.

Compositionality-Aware Transformer

To promote systematic compositionality, CAT introduces two modules, i.e., Multi-Primitive Composition and Representation Enhancement, to the vanilla Transformer encoder. These modules enable CAT to learn how to compose primitives and enhance the representation of the vanilla Transformer encoder, respectively.

Multi-Primitive Composition Module

This module explicitly equips the vanilla Transformer encoder with an ability of learning to compose primitives. It decomposes a contextual word representation h_{cont} into multiple primitives that bear the strong similarities. It then produces a compositional representation h_{comp}.

Primitives Representation. For each h_{cont}, we assume that they are grounded in the same semantic space[3], spanned by a limited number of distinct primitives. This forces the model to decode the elementary units from the contextual word representation h_{cont}. Considering the discreteness of language symbols, we require the i-th primitive to be represented by a discrete latent space vector $e_i \in R^m$, where m is the dimension size. Denoting K as the size of the discrete latent space, all the semantic space in CAT is grounded in an L-way categorical variable, $e \in R^{K\times m}$, which we refer to as a *codebook*. For instance, Fig.1 contains a codebook with nine discrete primitives/codes. Note that the codebook is a trainable parameter.

Decompositing into Primitives. Given a contextual representation h_{cont}, one way to decompose it is the vector quantization (van den Oord, Vinyals, and Kavukcuoglu 2017; Razavi, van den Oord, and Vinyals 2019; Liu et al. 2021, 2022), which involves learning a discrete latent representation for an input vector. Given an input vector $h_{cont} \in R^m$, the vector quantization method maps h_{cont} to the nearest-neighbor quantized code in the codebook $e^{K\times m}$. More concretely, the discretization process for vector h_{cont} is described as follows.

$$e_{o_j} = \text{Discretize}(h_{cont}),$$
$$\text{where } o_i = \underset{j\in\{1,\ldots,K\}}{\arg\min} \|h_{cont} - e_j\|_2^2, \quad (1)$$

where e_j is j-th code in the codebook e. Finally, the contextual representation h_{cont} is quantified using 'hard K-Means clustering' and discretized into one code.

However, his approach fails to meet the requirements for decomposition as it only produces one code. Also, the fitness of one code is limited compared to the original semantics of h_{cont}. To address this, our *multi-primitive composition module* decomposes a contextual representation into multiple discrete codes. In particular, our method uses a soft version of K-means clustering, where the number of clusters is dynamically learned. We achieve this by replacing the $\arg\min$ operator with the sparse attention mechanism (Zhang, Titov, and Sennrich 2021) as follows.

$$O = \text{RMSNorm}(\text{ReLU}(f(Q, K))), \quad (2)$$

where $Q = h_{cont}W_Q$, $K = eW_K$, and f is a scoring function like cosine. The vector O represents the sparse attention weight, where each index corresponds to a code in the codebook e. The RMSNorm operator also is utilized to increase the optimization stabilization, as suggested in Zhang, Titov, and Sennrich (2021). By this means, our multi-primitive composition filters out irrelevant codes from the compositions of h_{comp} and allows us to decompose h_{cont} into multiple latent codes of dynamic size. It is worth mentioning that our multi-primitive composition has the additional benefit of addressing the differentiation challenge. Unlike raw vector quantization, which approximates the gradient of the $\arg\min$ operator using Straight-Through-Estimator (Courbariaux, Bengio, and David 2015) or Gumbel-Softmax (Jang, Gu, and Poole 2016), both of which can be unstable (Yin et al. 2018), our approach is entirely differentiable.

Composing Primitives. To capture the semantics of various combinations, it is important that different primitives have varying significance in the composition process, depending on the context. We accomplish this by utilizing the attention vector $O \in R^K$ to select multiple discrete primitives and combining them into a compositional representation h_{comp}. This is achieved through the function $g(O, e) = OV$, where $V = eW_V$. It is noteworthy that CAT takes into account the systematic compositionality of a sentence by breaking down the h_{cont} of the [CLS] token, which is known to capture the semantics of the entire sentence.

Representation Enhancement Module

The objective of the Representation Enhancement module is to integrate h_{comp} and h_{cont} in a way that produces the ultimate output h_{mix}, which possesses superior systematic compositionality while retaining the contextual information of h_{cont} necessary for downstream tasks. As a result,

[3]This is inspired by *mutual knowledge hypothesis* (Sperber and Wilson 1986), saying that knowledge required to interpret a message is grounded in the understanding of the message sender and receiver.

CAT enhances the representation of the vanilla Transformer h_{cont}. To accomplish this, a self-attention mechanism is employed to automatically adjust the importance weights of h_{cont} and h_{comp} during the integration.

Formally, given an input sentence x of length N, i-th word x_i with its representations h^i_{cont} and h^i_{comp}, we derive two sequence embeddings $H_{comp} = \{h^0_{comp}, h^1_{comp}, ..., h^N_{comp}\}$ and $H_{cont} = \{h^0_{cont}, h^1_{cont}, ..., h^N_{cont}\}$, h^0_{cont} and h^0_{comp} correspond to representations of [CLS]. To integrate them, we feed the concatenatation of H_{comp} and H_{cont} into a self-attention layer to obtain the final outputs $H = \{h^0_{mix}, h^1_{mix}, ..., h^N_{mix}\}$, where $h_m ix^i$ is the mixed representation of x_i or [CLS]. This process is described as follows, where the second step re-scales the dimension for multi-layer CAT when necessary.

$$\begin{aligned} H &= \text{self-attention}([h_{comp}, h_{cont}]) \\ H &= \frac{1}{2}(H[:, : N, :] + H[:, N :, :]) \end{aligned} \tag{3}$$

By this means, the CAT learns to automatically fuse the contextual and compositional representations of each word according to the input context. As a result, three types of representations are generated, namely the raw contextual representations h_{cont}, discrete representations h_{comp}, and mixed representations H. The learned mixed representation H can be used as the contextual input for the next CAT layer.

Compositionality-aware Pre-training Tasks

This section introduces two new pre-training tasks that aim to improve the systematic compositionality ability. The first task is the guided decomposition task, which utilizes sememes to supervise the semantic decomposition of h_{cont}. The second task is the semantics composition task, which ensures the composition of discrete primitives so that h_{comp} and h_{mix} are semantically informative without losing contextual information required in the downstream tasks.

Guided Decomposition Task

To further enhance the interpretability and learning efficiency of the multi-primitive composition module, we resort to the semantic compositionality (Qi et al. 2019a; Wierzbicka 1996) and utilize sememes, which are defined as the minimum semantic units of human languages (Bloomfield 1926), to supervise the decomposition. Building upon the assumption that the meanings of all the words can be composed of a limited set of sememes (Qi et al. 2019a; Wierzbicka 1996), each code/primitive in the CAT's codebook is defined as a specific sememe. Thus, the compositional representation can be interpreted as the semantic composition of selected sememes. Consequently, the goal of this task is to learn the map from each word to its corresponding sememes.

Sememes Guided Decomposition. To achieve this, we resort to the OpenHowNet (Qi et al. 2019b), a widely acknowledged sememe knowledge base, as the supervision of semantic decomposition. We require the sparse attention mechanism to attend to the appropriate sememes in the codebook. Here, x_i represents the i-th word in an input sentence of length N, $e \in R^{K\times m}$ is the codebook in CAT (or the last layer of the multi-layer CAT), $O_i \in R^K$ is the attention weight to the codebook and $s_i \in R^K$ is a $\{0, 1\}$ vector containing the correct sememes for x_i, where $s_{ij} = 1$ if the j-th sememe belongs to x_i, otherwise $s_{ij} = 0$. To calculate the loss L_{GD} for incorrect sememe-matchings, we averaged the attention weights that are focused on the wrong sememes. Furthermore, we introduced the L1 norm on the attention weights from all layers to encourage sparsity.

$$L_{GD}(x_i) = \sum_{x_i \in x} \frac{sum((1 - s_i) \odot O_i)}{sum(1 - s_i)} + L_1(O_i) \tag{4}$$

Semantics Composition Task

The goal of the semantics composition task is to encourage h_{comp} and h_{mix} to be semantically informative without losing contextual information required in the downstream tasks. Here, we involve the following three learning objectives.

Reconstruction Consistency. It aims to encourage the compositional representations h_{comp} to bear strong similarity with the contextual ones h_{cont} in the vector space. We consider the following objectives, where the stop gradient operator sg and hyper-parameter $\beta \in [0, 1)$ are utilized to shift more optimization focus on updating discrete representation, rather than the contextual one.

$$\ell_{rc} = \|H_{comp} - sg(H_{cont})\|^2_2 + \beta\|sg(H_{comp}) - H_{cont}\|^2_2 \tag{5}$$

Semantic Sufficiency. It concerns that h_{cont}, h_{comp}, and h_{mix} contain sufficient semantic information. They should be effective on the pre-training task Ptask used in BERT.

$$\ell_{ss} = \text{Ptask}(H_{comp}) + \text{Ptask}(H_{cont}) + \text{Ptask}(H) \tag{6}$$

Nuance Minimization. It encourages the nuance between contextual and compositional representations $H_{nu} = H_{cont} - H_{comp}$ should be less informative, insignificant to the main semantics. Thus, we aim at maximizing the entropy of H_{nu} and minimizing its performance on the pre-training tasks used in the semantic sufficiency.

$$\ell_{nm} = exp(-\text{Entropy}(H_{nu})) + exp(-\text{Ptask}(H_{nu})) \tag{7}$$

Finally, we summarize the overall loss function for the semantics composition task as $L_{SC} = \ell_{rc} + \ell_{ss} + \ell_{nm}$. The overall loss function for CAT pre-training is $L = L_{task} + L_{GD} + L_{SC}$, where L_{task} is the loss for other pre-training tasks used in BERT.

Experiments

In this section, we concentrate on comparing our approach to the vanilla Transformer and conducting comprehensive evaluations that center on the features of systematic compositionality (as discussed in Section). Additionally, we explore other characteristics of CAT (as discussed in Section and), including robustness and combinatorial effectiveness. More details on implementation, datasets and pre-training can be found in the Appendix of our arXiv version.

Experimental Setup

Model Pre-training. We restrict our discussions to the vanilla Transformer framework, with a special focus on the popular BERT[4] to provide more detailed insights. To ensure the fairness of the experiment, we strictly follow the pre-training process of BERT (Kenton and Toutanova 2019) in our experiments. Specifically, we implement a multi-layer CAT (MCAT) and pre-train it from scratch. The BooksCorpus (800M words), English Wikipedia (2,500M words), and extra OpenHowNet dataset serve as our pre-training data for MCAT. We utilize the pre-training tasks of BERT, namely Masked LM and Next Sentence Prediction, as the Ptask in our Semantics Composition.

Baselines & Implementation. To simplify, we use the notations CAT_{cont}, CAT_{comp}, and CAT_{mix} to represent the contextual, compositional, and mixed representations of our multi-layer CAT, respectively. As for the baselines, BERT is of particular importance to us due to its relevance to our network architecture and pre-training data. In addition to BERT, we also consider two other variants, namely RoBERTa[5] (Liu et al. 2019) and DistilBERT (Sanh et al. 2019), following the approach of Yu and Ettinger (2021, 2020). For all experiments, we fine-tune each model with the corresponding backbone frozen and an additional MLP layer on top of the backbone for learning, which takes the $[CLS]$ embeddings as inputs. We selected the best model based on its performance on the validation set for downstream task testing.

Evaluation Tasks. It's important to note that most existing evaluation methods and datasets for systematic compositionality are only suitable for generative-based language models like SCAN (Higgins et al. 2018b) and in-context prompting (An et al. 2023). Instead, our evaluations follow recent works (Yu and Ettinger 2021; Hendrycks et al. 2020; Yu and Ettinger 2020), which are designed for discriminative-based language models like BERT and ours. Since each evaluation task differs, we provide the details in the corresponding subsections.

Systematic Compositionality Evaluation

After pre-training on corpora with diverse words and their combinations, we assess the systematic compositionality of CAT on phrases and sentences. We also evaluate the compositional generalization to the out-of-distribution dataset.

Evaluation on Phrases & Sentences In this section, we show the evaluation on both phrase-level and sentence-level compositionality-aware tasks.

Task & Dataset. Our goal is to assess the capability of recognizing the semantic correlation that results from different word compositionality. In this section, we consider the two evaluation sub-tasks, i.e., phrase correlation and adversarial paraphrase sentence classification, to assess the semantics correlation of different phrases and different sentences, respectively. In our experiments, inspired by previous studies (Yu and Ettinger 2021, 2020), we fine-tune each model on data that are good candidates for requiring composition and test the fine-tuned model on these sub-tasks.

Methods	Adversary	Correlation	
	PAWS (SO)	BiRD	BiRD-ABBA
BERT	88.22	19.50	1.64
RoBERTa	89.06	20.32	3.21
DistilBERT	87.44	17.49	1.03
$\text{CAT}_{cont.}$	89.05	20.19	1.25
$\text{CAT}_{comp.}$	**90.33**	**34.44**	**7.19**
CAT_{mix}	<u>89.32</u>	<u>27.72</u>	<u>6.25</u>

Table 2: Systematic compositionality evaluation (%) on phrases and sentences. 'SO' means 'swap-only'.

- <u>Phrase Correlation</u>. It aims to evaluate whether CAT and baselines capture compositional phrase information and identify the semantic correlation between two phrases. Following Yu and Ettinger (2021, 2020), we fine-tune each model on the PAWS (Zhang, Baldridge, and He 2019), which consists of sentence pairs with high lexical overlap. This fine-tuning task is formulated as a binary classification of whether two sentences are paraphrases or not. We then assess the fine-tuned model on BiRD (Asaadi, Mohammad, and Kiritchenko 2019), which is a bigram-relatedness dataset designed to evaluate composition (e.g., *Safety officer* and *Security guard*). In addition to testing on the full BiRD dataset, we conduct a controlled experiment to remove the effects of word overlap by filtering the BiRD dataset to pairs in which the two phrases consist of the same words (e.g., *Water body* and *Body water*). We refer to the filtered dataset as BiRD-ABBA. For both BiRD and BiRD-ABBA, the model performance is measured by the alignment with human judgments of phrase meaning correlation. We report the Pearson correlation between the cosine of phrases and human-rated score[6]. In this case, we could determine the systematic compositionality of each model by measuring the ability in capturing compositional phrase information beyond lexical content.
- <u>Adversarial Paraphrase Sentences Classification</u>. It determines if the semantics hold when partial words in the input sentence are swapped. To achieve this, we fine-tune each model on PAWS and test them on an adversarial PAWS dataset, called PAWS (swap-only) (Zhang, Baldridge, and He 2019). PAWS (swap-only) simulates the changing of word compositions (orderings). It contains both paraphrase and non-paraphrase pairs with high bag-of-words overlap and word swapping. In this study, we report the accuracy of the successful identification of non-paraphrase pairs to assess our effectiveness.

Main Results. As shown in Table 2, the results suggest that *CAT_{mix} and CAT_{comp} show superiority on systematic compositionality on both tasks*. On average, CAT_{mix} improves the performance on the adversarial paraphrase sen-

[4]We restrict our discussions to BERT without loss of generality. Due to computational resource limitations, we plan to implement a model with more significant parameter quantities in the future.

[5]Different from BERT and us, RoBERTa use more pre-training datasets and customized framework.

[6]Human scores are available in the datasets

Method	Movie		STSB		MNLI		AMAZON	
	IMDB	SST2	Images	MSRvid	Telephone	Letters	Music	Video
	IID	OOD	IID	OOD	IID	OOD	IID	OOD
BERT	79.82	76.61	<u>80.94</u>	88.00	<u>74.87</u>	72.79	73.32	65.16
RoBERTa	**82.09**	75.47	**81.31**	88.04	**76.29**	72.94	**74.30**	66.06
DistilBERT	78.64	75.07	80.43	88.38	73.14	70.03	73.04	64.58
CAT$_{cont.}$	79.34	76.15	80.05	88.81	74.31	72.85	<u>73.98</u>	65.22
CAT$_{comp.}$	77.81	<u>78.56</u>	78.31	**89.67**	73.36	<u>73.08</u>	72.41	<u>67.32</u>
CAT$_{mix}$	<u>80.23</u>	**79.36**	79.97	<u>89.13</u>	<u>74.87</u>	**73.96**	73.88	**67.44**

Table 3: Evaluation on out-of-distribution datasets (%). We report the Pearson's coefficient for STSB, and accuracy for others.

tences classification by 1.10 and phrase correlation by 6.42 compared to our primary baseline BERT. Additionally, it enhances the performance on the adversarial paraphrase sentences classification by 0.26 and phrase correlation by 5.22 compared to the best baseline (i.e., RoBERTa). These observations imply that after being pre-trained to compose primitives, CAT improves the ability in identifying the semantics correlation formed by word compositionalities. Although CAT$_{comp}$ is notably superior to CAT$_{mix}$, we would demonstrate in Sections that CAT$_{mix}$ strikes a good balance between systematic compositionality and effectiveness on standardized language understanding task.

In detail, regarding the adversarial paraphrase sentences classification, CAT$_{comp}$ achieves the best results thanks to the explicit systematic compositionality modeling, which improves its performance by 2.11 compared to BERT, 1.27 compared to RoBERTa, and 2.89 compared to DistilBERT. Moreover, CAT$_{cont}$ performs similarly to BERT, while CAT$_{mix}$ outperforms BERT by 1.10. Interestingly, despite RoBERTa using more pre-training data than BERT and ours, it is still weaker than CAT$_{comp}$ and CAT$_{mix}$. These findings highlight the significance of systematic compositionality modeling in enhancing the performance of pre-trained models on compositionality-aware phrasal tasks.

Regarding the phrase correlation, CAT$_{mix}$ performs significantly better than BERT and even outperforms the best baseline, RoBERTa, by 7.4 on BiRD and 3.04 on BiRD-ABBA. However, we also observe that the overall results on BiRD-ABBA are much lower than BiRD, as identifying the relatedness of bigram pairs with high word overlap is more challenging. In fact, it requires assessing semantic changes in different compositions of the same group of words. We further tested ChatGPT (refer to Appendix), which has a superabundance of training data and several hundred times more parameters than BERT. We found that its score on BiRD-ABBA is very unsatisfying. In particular, the result of ChatGPT scoring on BiRD data is 45.65, while the result on BiRD-ABBA plummets to 23.28. This indicates poor alignment with human judgments of phrase meaning similarity. These results suggest that, at least on this compositionality-aware task, ChatGPT, which may seem "omnipotent", may be far from the general intelligence demonstrated in humans.

Evaluation on Out-of-distribution Datasets Here, we present the assessment of compositional generalization.

Task & Dataset. After pre-training on corpora with diverse words and their combinations, we aim to evaluate the compositional generalization of each model to out-of-distribution (OOD) datasets. To achieve this, our evaluation follows Hendrycks et al. (2020). Given a dataset pair (A, B), we fine-tune each model on A (i.e., IID), and test it on B (i.e., OOD) that contains realistic distribution shifts to A. Following Hendrycks et al. (2020), we consider the following dataset pairs: IMBD (Maas et al. 2011) and SST2 (Socher et al. 2013), STSB-image and STSB-MSRvid (Cer et al. 2017), MNLI-telephone and MNLI-letters (Williams, Nangia, and Bowman 2018), AMAZON-music and AMAZON-video (He and McAuley 2016). For STSB, we report Pearson's correlation coefficient, while for the other datasets, we report accuracy.

Main Results. As shown in Table 3, *CAT$_{mix}$ is better at generalizing to out-of-distribution composed semantics*. Due to pre-training on a larger corpus, RoBERTa demonstrates significant performance in general language understanding tasks (i.e., IID) compared to others. However, despite the IID effectiveness of RoBERTa, it suffers a significant decrease in effectiveness on OOD datasets. On the contrary, CAT$_{mix}$ exhibits advantages in OOD data, with an average improvement of 1.85 compared to RoBERTA (3.89 on SST2, 1.09 on STSB-MSRvid, 1.02 on MNLI-Latter, and 1.38 on AMAZON-Video) and 1.83 compared to BERT (2.75 on SST2, 1.13 on STSB-MSRvid, 1.17 on MNLI-Latter, and 2.28 on AMAZON-Video).

Compared to CAT with different representations, CAT$_{cont}$ is better on IID datasets than CAT$_{comp}$, with an average improvement of 1.45. This empirical evidence is consistent with the previous conclusion of theoretical analysis (Liu et al. 2021), stating that the combinatorial expressiveness of discrete codes is able to model complex language phenomena, but it is still weaker than the contextual representations. However, after mixing CAT$_{cont}$ and CAT$_{comp}$ together, CAT$_{mix}$ strikes a good balance.

Robustness of Systematic Compositionality

CAT benefits from the discreteness of primitives and their composition. Research conducted earlier has demonstrated that discrete variables possess the attribute of being robust to noise (Liu et al. 2021). This finding has inspired us to conduct a thorough examination and evaluate the robustness of CAT in the process of composing discrete primitives. Specifically, we are interested in examining the multi-primitive composition module in CAT, which filters out irrelevant codes/primitives from the compositions of h_{comp} and enables us to break down h_{cont} into several dynamic-sized

	BERT	**RoBERTa**	**DistilBERT**	**CAT**$_{cont}$	**CAT**$_{disc}$	**CAT**$_{mix}$
CoLA	52.10	61.60	47.60	60.18	50.09	60.28
MNLI_m	78.12	80.12	73.62	75.28	72.85	74.28
MNLI_mm	78.41	79.41	74.61	75.97	74.40	75.19
MRPC	74.40	75.70	73.20	74.95	74.03	74.93
QNLI	85.45	87.75	82.85	84.21	81.19	83.21
QQP	71.20	77.90	70.10	76.19	71.78	76.23
RTE	64.62	65.92	55.22	62.93	60.68	62.76
SST2	94.50	95.80	93.10	93.98	93.44	94.14
STSB	85.80	87.29	83.70	85.71	82.85	84.42
Avg.	76.07	**79.05**	72.67	<u>76.60</u>	73.48	76.16

Table 4: Effectiveness on standardized test. CAT has minimal impact on the effectiveness of the GLUE task.

latent codes. In this section, our objective is to determine whether CAT is capable of filtering out irrelevant primitives during composition.

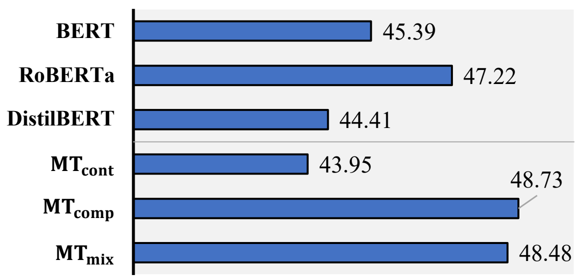

Figure 2: Illustration of robustness evaluation.

Task & Dataset. Our goal is to determine if CAT can filter out irrelevant primitives during composition. For this purpose, following Jia and Liang (2017), we test the ability of CAT to comprehend contexts that include adversarially inserted irrelevant sentences and answer questions about the given contexts. We utilize reading comprehension datasets, i.e., SQuAD (Rajpurkar et al. 2016) and its variant with adversarial noises. These noises are automatically generated to distract models without altering the correct answer or misleading humans. Essentially, we fine-tune each model on the SQuAD and test it on the SQuAD-adversarial.

Main Results. As shown in Figure 2, CAT, which employs mixed representations, exhibits remarkable robustness to adversarial samples. It outperforms BERT by 3.09 and the best baseline by 1.26. This performance gain may be attributed to the fact that semantic decomposition based on a discrete codebook is beneficial for filtering out irrelevant information in the [CLS] embedding. Such semantic decomposition improves the effectiveness of the compositional representation CAT$_{comp}$, which achieves even better performance than CAT$_{mix}$. One possible explanation is that irrelevant information contained in the contextual representation CAT$_{cont}$ may be fused into the CAT$_{mix}$ during our mixing procedure. However, this does not affect our advantage.

Effectiveness on Standardized Test

Our previous study suggests the combinatorial effectiveness of discrete primitives (Liu et al. 2021). After demonstrating CAT's efficacy in compositionality-aware tasks, in this section, we proceed to assess the effectiveness of composed primitives in standardized tests. Our objective is to investigate whether the contextual information required by downstream tasks would be affected when the vanilla Transformer is equipped with systematic compositionality capabilities.

Task & Dataset. In this section, we assess the language understanding capability on GLUE (Wang et al. 2018) following BERT, and report scores on each task after finetuning. In line with BERT, we report the F1 score for MRPC and QQP datasets, the Spearman correlation score for STSB, and the accuracy score for the remaining tasks.

Main Results. According to Table 4, on average, CAT$_{cont}$ and CAT$_{mix}$ exhibit slightly better performance than BERT (+0.53, +0.09, respectively), while CAT$_{comp}$ is weaker. As previously discussed, the compositional representation CAT$_{comp}$ is better tailored for compositionality-aware and discretization-oriented tasks, while the contextual representation CAT$_{cont}$ is more appropriate for standardized tasks. More important, CAT$_{mix}$ strikes a good balance between compositionality-aware and standardized tasks. Our findings highlight that the proposed CAT is adept at capturing semantic compositionality and significantly outperforms baselines on compositionality-aware tasks, with a minimal impact on the effectiveness of standardized tasks.

Conclusion

Our research delves into the characteristics of systematic compositionality in human languages. For the first time, we propose a compositionality-aware Transformer (CAT) and two new pre-training tasks to facilitate systematic compositionality. We tentatively provide a successful implementation for multi-layer CAT and empirically verify its effectiveness. Our approach captures semantic compositionality better and significantly outperforms baselines on compositionality-aware tasks, with minimal impact on the effectiveness of standardized language understanding tasks.

Systematic compositionality is widely believed to be characteristic of human intelligence. Our study provides a primary exploration of this challenge, and our findings may provide an important proof-of-concept for producing better LLMs, which crystallizes the past experiences and generalizes them to the new composed contexts.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62272330); in part by the Fundamental Research Funds for the Central Universities (No. YJ202219).

References

Akyürek, E.; and Andreas, J. 2022. Compositionality as Lexical Symmetry. *arXiv preprint arXiv:2201.12926*.

An, S.; Lin, Z.; Fu, Q.; Chen, B.; Zheng, N.; Lou, J.-G.; and Zhang, D. 2023. How Do In-Context Examples Affect Compositional Generalization? *arXiv preprint arXiv:2305.04835*.

Andreas, J. 2020. Good-Enough Compositional Data Augmentation. In *ACL*, 7556–7566.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 39–48.

Asaadi, S.; Mohammad, S.; and Kiritchenko, S. 2019. Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 505–516. Minneapolis, Minnesota: Association for Computational Linguistics.

Ataman, D.; and Federico, M. 2018. Compositional Representation of Morphologically-Rich Input for Neural Machine Translation. In *ACL*, 305–311.

Baroni, M. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791): 20190307.

Bloomfield, L. 1926. A set of postulates for the science of language. *Language*, 2(3): 153–164.

Bogin, B.; Subramanian, S.; Gardner, M.; and Berant, J. 2021. Latent compositional representations improve systematic generalization in grounded question answering. *Transactions of the Association for Computational Linguistics*, 9: 195–210.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Cartuyvels, R.; Spinks, G.; and Moens, M.-F. 2021. Discrete and continuous representations and processing in deep learning: Looking forward. *AI Open*, 2: 143–159.

Cer, D.; Diab, M.; Agirre, E. E.; Lopez-Gazpio, I.; and Specia, L. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In *The 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 1–14.

Chomsky, N. 2014. *Aspects of the Theory of Syntax*, volume 11. MIT press.

Courbariaux, M.; Bengio, Y.; and David, J.-P. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, 3123–3131.

Dziri, N.; Lu, X.; Sclar, M.; Li, X. L.; Jian, L.; Lin, B. Y.; West, P.; Bhagavatula, C.; Bras, R. L.; Hwang, J. D.; et al. 2023. Faith and Fate: Limits of Transformers on Compositionality. *arXiv preprint arXiv:2305.18654*.

Farinhas, A.; Aziz, W.; Niculae, V.; and Martins, A. 2021. Sparse Communication via Mixed Distributions. In *ICLR*.

Fodor, J. A.; and Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2): 3–71.

Frege, G. 1948. Sense and reference. *The philosophical review*, 57(3): 209–230.

Friede, D.; and Niepert, M. 2021. Efficient Learning of Discrete-Continuous Computation Graphs. *Advances in Neural Information Processing Systems*, 34: 6720–6732.

Gordon, J.; Lopez-Paz, D.; Baroni, M.; and Bouchacourt, D. 2019. Permutation equivariant models for compositional generalization in language. In *ICLR*.

He, R.; and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, 507–517.

Hendrycks, D.; Liu, X.; Wallace, E.; Krishnan, A. D. R.; and Song, D. 2020. Pretrained Transformers Improve Out-of-Distribution Robustness. In *Association for Computational Linguistics*.

Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; and Lerchner, A. 2018a. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Higgins, I.; Sonnerat, N.; Matthey, L.; Pal, A.; Burgess, C.; Bošnjak, M.; Shanahan, M.; Botvinick, M.; Hassabis, D.; and Lerchner, A. 2018b. SCAN: Learning Hierarchical Compositional Visual Concepts. In *ICLR*, volume 6. ICLR.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical Reparameterization with Gumbel-Softmax. In *ICLR*.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031. Copenhagen, Denmark: Association for Computational Linguistics.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Keysers, D.; Schärli, N.; Scales, N.; Buisman, H.; Furrer, D.; Kashubin, S.; Momchev, N.; Sinopalnikov, D.; Stafiniak, L.; Tihon, T.; et al. 2019. Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. In *ICLR*.

Lake, B.; and Baroni, M. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*, 2873–2882. PMLR.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Li, Y.; Zhao, L.; Church, K.; and Elhoseiny, M. 2019. Compositional Language Continual Learning. In *ICLR*.

Liu, D.; Lamb, A.; Ji, X.; Notsawo, P.; Mozer, M.; Bengio, Y.; and Kawaguchi, K. 2022. Adaptive Discrete Communication Bottlenecks with Dynamic Vector Quantization. *arXiv preprint arXiv:2202.01334*.

Liu, D.; Lamb, A. M.; Kawaguchi, K.; ALIAS PARTH GOYAL, A. G.; Sun, C.; Mozer, M. C.; and Bengio, Y. 2021. Discrete-Valued Neural Communication. *Advances in Neural Information Processing Systems*, 34.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 4114–4124. PMLR.

Loula, J.; Baroni, M.; and Lake, B. 2018. Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 108–114.

Ma, Y.; Zhang, C.; and Zhu, S.-C. 2023. Brain in a Vat: On Missing Pieces Towards Artificial General Intelligence in Large Language Models. *arXiv preprint arXiv:2307.03762*.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.

Montague, R. 1970. Universal grammar. *Theoria*, 36(3): 373–398.

Qi, F.; Huang, J.; Yang, C.; Liu, Z.; Chen, X.; Liu, Q.; and Sun, M. 2019a. Modeling Semantic Compositionality with Sememe Knowledge. In *ACL*, 5706–5715.

Qi, F.; Yang, C.; Liu, Z.; Dong, Q.; Sun, M.; and Dong, Z. 2019b. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, 14866–14876.

Russin, J.; Jo, J.; O'Reilly, R. C.; and Bengio, Y. 2019. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Sperber, D.; and Wilson, D. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6309–6318.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Wierzbicka, A. 1996. *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122.

Yin, P.; Lyu, J.; Zhang, S.; Osher, S.; Qi, Y.; and Xin, J. 2018. Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets. In *ICLR*.

Yu, L.; and Ettinger, A. 2020. Assessing Phrasal Representation and Composition in Transformers. In *EMNLP*, 4896–4907.

Yu, L.; and Ettinger, A. 2021. On the Interplay Between Fine-tuning and Composition in Transformers. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Zhang, B.; Titov, I.; and Sennrich, R. 2021. Sparse Attention with Linear Units. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6507–6520.

Zhang, Y.; Baldridge, J.; and He, L. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1298–1308.