Explicitly Representing Syntax Improves Sentence-to-layout Prediction of Unexpected Situations

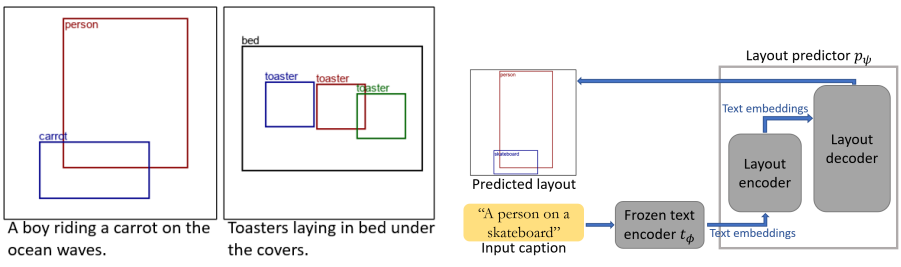
Ruben Cartuyvels[0000−0003−1063−4659], Wolf Nuyts[0000−0001−7254−4371], and Marie-Francine Moens[0000−0002−3732−9323]

ruben.cartuyvels@kuleuven.be
Department of Computer Science, KU Leuven, Belgium

Keywords: Natural Language Processing · Deep Learning.

1 Introduction

Current neural networks build powerful representations of content. However, unlike humans, they fail when confronted with unexpected situations and content which is out of context [4], which exposes a lack of compositional understanding. Compositionality enables humans to understand and generate a potentially infinite number of novel situations by viewing the situation as a novel composition of familiar simpler parts [8,2,3]. Since human language is characterized by recursive structures which correspond with recursion that humans perceive in the world [5,6], we hypothesize that representations that better encode the syntactical structure of a sentence are less sensitive to a decline in performance when confronted with unexpected situations. We test this hypothesis with the task of 2D visual object layout prediction given a natural language input sentence that describes an unexpected situation. Figure 1a shows example situations, and 1b gives an overview of our models.[1]



(a) (b)

Fig. 1: (a) Samples from the USCOCO dataset, (b) Model overview.

[1] This abstract discusses earlier work by the authors [9]. Code, trained models and the USCOCO data are available via https://github.com/rubencart/USCOCO.

2 Methods

1. We collect a test set of grammatically correct sentences and layouts, called Unexpected Situations of Common Objects in Context (USCOCO), describing compositions of entities and relations that unlikely have been seen during training (examples shown in figure 1a).
2. We train and evaluate transformer-based layout generation networks, that take as input text representations computed by a pretrained and frozen text encoder. We consider text encoders with various sizes (ranging from the smallest `GPT-2` to `LLaMA-33B`), in 2 categories.
   (a) **Implicit** syntax: transformers trained for next token prediction (NTP) such as `GPT-2`, `LLaMA` and a version of `GPT-2` pretrained on a smaller dataset [1,11,10,15]. These models are shown to have syntactic knowledge encoded in their hidden representations [7,14,16].
   (b) **Explicit** syntax: transformers trained for NTP that take a linearized version of the constituency trees as input, e.g. "(NP a dog) (VP catches (NP a frisbee))", including tags and brackets as tokens, and that apply attention mask constraints based on constituency structure [10,12].
3. We propose a novel structural loss function that better retains the syntactic structure of the sentence in the text representations by enforcing the alignment between syntax tree embeddings [13] and the output embeddings of the layout predictors. This loss function is evaluated both with models that explicitly integrate syntax and with models that implicitly encode syntax.

3 Results

- Scores drop drastically on USCOCO vs. on in-domain test data, for all models.
- Increasing model size or pretraining data size gives advantage on in-domain test data, but not on USCOCO, so this does **not** solve the issue.
- *Without* structural loss, implicit and explicit syntax models perform similarly on both test sets.
- *With* increasing structural loss weight, performance on USCOCO improves for explicit syntax models (fig. 2), but drops for implicit syntax models.
- A human evaluation experiment confirms our quantitative findings.

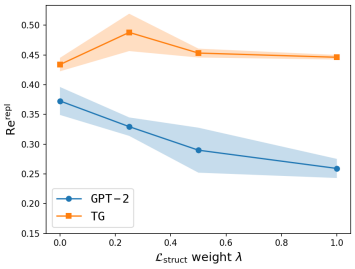


Fig. 2: Recall vs. structural loss weight for implicit (`GPT-2`) and explicit (`TG`) syntax models.

Conclusion. We proposed a contrastive loss that enforces the encoding of syntax in the representation of a visual scene and show that it increases generalization to unexpected compositions if used with text encoders that explicitly integrate syntax. The loss has the potential to be used in other generation tasks that condition on structured input.

Acknowledgments. This work is part of the CALCULUS project, which is funded by the ERC Advanced Grant H2020-ERC-2017 ADG 788506.[2] It also received funding from the Research Foundation – Flanders (FWO) under Grant Agreement No. G078618N. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by FWO and the Flemish Government.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., Johnson, M.: Bllip 1987-89 wsj corpus release 1 ldc2000t43 (2000). https://doi.org/10.35111/fwew-da58, https://catalog.ldc.upenn.edu/LDC2000T43
2. Chomsky, N.: Aspects of the Theory of Syntax. MIT press (1965)
3. Frankland, S.M., Greene, J.D.: Concepts and compositionality: In search of the brain's language of thought. Annual Review of Psychology **71**(1), 273–303 (2020). https://doi.org/10.1146/annurev-psych-122216-011829, https://doi.org/10.1146/annurev-psych-122216-011829, pMID: 31550985
4. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
5. Hauser, M.D., Chomsky, N., Fitch, W.T.: The faculty of language: What is it, who has it, and how did it evolve? Science **298**(5598), 1569–1579 (2002). https://doi.org/10.1126/science.298.5598.1569, https://www.science.org/doi/abs/10.1126/science.298.5598.1569
6. Hawkings, J.: A Thousand Brains: A New Theory of Intelligence. Basic Books (2021)
7. Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 4129–4138. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1419, https://doi.org/10.18653/v1/n19-1419
8. Humboldt, W.: On language: On the diversity of human language construction and its influence on the mental development of the human species. Cambridge University Press (1999)
9. Nuyts, W., Cartuyvels, R., Moens, M.: Explicitly representing syntax improves sentence-to-layout prediction of unexpected situations. Trans. Assoc. Comput. Linguistics **12**, 264–282 (2024). https://doi.org/10.1162/TACL_A_00643, https://doi.org/10.1162/tacl_a_00643
10. Qian, P., Naseem, T., Levy, R., Astudillo, R.F.: Structural guidance for transformer language models. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP. pp. 3735–3745. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.289, https://doi.org/10.18653/v1/2021.acl-long.289

[2] https://calculus-project.eu/

11. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
12. Sartran, L., Barrett, S., Kuncoro, A., Stanojevic, M., Blunsom, P., Dyer, C.: Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. Trans. Assoc. Comput. Linguistics **10**, 1423–1439 (2022), https://transacl.org/ojs/index.php/tacl/article/view/3951
13. Shiv, V.L., Quirk, C.: Novel positional encodings to enable tree-based transformers. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS. pp. 12058–12068 (2019), https://proceedings.neurips.cc/paper/2019/hash/6e0917469214d8fbd8c517dcdc6b8dcf-Abstract.html
14. Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S.R., Das, D., Pavlick, E.: What do you learn from context? probing for sentence structure in contextualized word representations. In: 7th International Conference on Learning Representations, ICLR. OpenReview.net (2019), https://openreview.net/forum?id=SJzSgnRcKX
15. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023). https://doi.org/10.48550/arXiv.2302.13971, https://doi.org/10.48550/arXiv.2302.13971
16. Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S., Bowman, S.R.: Blimp: The benchmark of linguistic minimal pairs for english. Trans. Assoc. Comput. Linguistics **8**, 377–392 (2020). https://doi.org/10.1162/tacl_a_00321, https://doi.org/10.1162/tacl_a_00321