"A Structural Probe for Finding Syntax in Word Representations" by Hewitt and Manning (2019)

Hewitt and Manning (2019) introduce a novel approach, the **structural probe**, to investigate whether syntactic information is embedded within word representations generated by language models like BERT and ELMo. The structural probe is a supervised linear transformation that projects contextual word embeddings into a lower-dimensional space, where squared Euclidean distances (L2) between word vectors approximate syntactic distances in a parse tree. In this transformed space, syntax is encoded if the squared L2 distance between two words corresponds to the number of edges between them in the tree, with each word's squared L2 norm representing its depth. To reconstruct these distances and hierarchical relationships, the probe uses supervised learning to find the optimal transformations that approximate these properties for each model.

The probe measures (1) how well syntactic relationships (word connections) are captured using metrics like Undirected Unlabeled Attachment Score (UUAS) and distance Spearman correlation (DSpr), and (2) the extent to which word hierarchy or depth is represented using norm Spearman correlation (NSpr) and root identification accuracy (root%). Each test sentence's predicted parse tree distances are used to compute a Minimum Spanning Tree (MST), aligning projected structures with the gold parse trees and helping assess the probe's ability to capture syntax.

Experiments show that **syntax trees are indeed encoded** within ELMo and BERT embeddings, supported by high performance on parsing-related metrics. Additionally, it was found these models can represent syntax effectively even in low-dimensional subspaces, with BERTLARGE consistently outperforming BERTBASE and ELMo, highlighting the greater syntactic encoding capacity of larger models.

Strengths: One key strength of this paper is its innovative approach to probing syntax through geometric distances, offering clear, quantifiable metrics (e.g., UUAS, DSpr) for understanding syntactic structures beyond simple word-to-word links. Additionally, the probe's distance metric is guaranteed to be nonnegative and symmetric, which simplifies the probing process and ensures consistent, logical results. This distance measure doesn't just identify pairwise word connections; it also captures the overall syntactic structure of sentences. The authors also rigorously test their method across different layers of BERT and on various treebank datasets and baselines, offering comprehensive insights into where syntactic information is most prominently encoded within the model.

Limitations: The probe's reliance on supervised learning using gold-standard parse trees may make it more reflective of specific annotated data than of inherent properties in embeddings, potentially limiting its generalizability to other languages or domains without high-quality syntactic resources. Additionally, the probe's linearity might oversimplify the complex, non-linear relationships within neural networks. The focus on global syntactic structure (distances between all word pairs) may miss local syntactic details, such as headedness, and while the probe reveals syntactic structures in the vector space, it remains uncertain to what extent models utilize this information in actual predictions.

Supervised vs. Unsupervised Probing Approach: The proposed probing approach is **supervised**. It requires gold-standard syntactic parse trees (e.g. from the Penn Treebank) to train the structural probe, aligning the projected embedding distances with actual tree distances. **Advantages** include precise alignment with specific linguistic structures and the ability to quantitatively measure the encoding of syntax in embeddings. It allows researchers to target particular aspects of syntax and obtain measurable results. **Disadvantages** involve dependency on annotated data, which may not be available for all languages or domains. Supervised probes may also overfit to the training data, capturing artifacts of the annotations rather than genuine properties of the embeddings. In contrast, unsupervised approaches may be more versatile and reveal inherent structures in the data, but they might lack the precision and interpretability of supervised methods with less accuracy in identifying 'specific' linguistic phenomena.

Importance of Probing Models for Syntactic Knowledge: Probing models for syntactic knowledge is crucial for both practical and theoretical reasons. Practically, understanding whether and how models encode syntax can contribute to improvements in NLP tasks like parsing, translation, and question answering. It helps in diagnosing model weaknesses and guiding architectural enhancements to better capture linguistic structures and context. Probing enhances our theoretical grasp of how neural networks process language, offering insights into the similarities between computational models and human language learning. It tackles essential questions regarding how language is represented within artificial systems, pushing forward the field of computational linguistics by connecting neural computation with linguistic theory.

Unclear points: The mathematical formalization of the structural probe, particularly the derivation of the transformation that aligns embedding distances with tree distances, was intricate and could benefit from a more intuitive explanation. Some aspects of the linear transformation (e.g., why the squared L2 norm and distance work better than their regular forms) were challenging to follow.