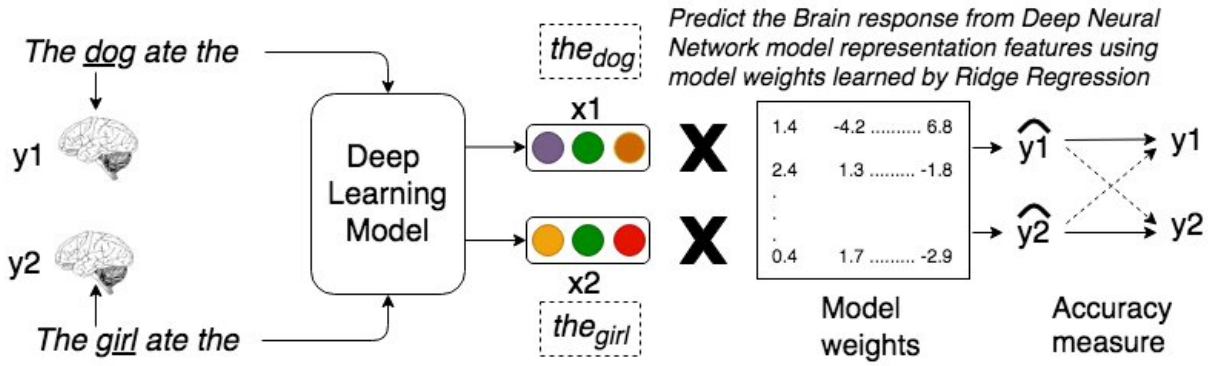
Relating sentence representations in deep neural networks with those encoded by the brain

July 15 2019, by Ingrid Fadelli



Experimental setup for micro-context tests. Given two sentences with similar words except one in the past (underlined), the test evaluates if the deep neural network model representation contains sufficient information to tell the two words apart. Credit: Jat et al.

Researchers at the Indian Institute of Science (IISc) and Carnegie Mellon University (CMU) have recently carried out a study exploring the relationship between sentence representations acquired by deep neural networks and those encoded by the brain. Their paper, pre-published on arXiv and set to be presented at this year's Association for Computational Linguistics (ACL) conference, unveiled correlations between activations in deep neural models and MEG brain data that could aid our current understanding of how the brain and deep learning

algorithms process language.

"The research is the product of a collaboration between Professor Tom Mitchell's brain research group at CMU and Professor Partha Talukdar's MALL lab at IISc," Sharmistha Jat, one of the researchers who carried out the study, told TechXplore. "Research in these groups focuses on questions such as how the brain understands language. One of the main objectives of the current study was to understand algorithms that process language in the brain and compare them to neural network language models."

Jat and her colleagues set out to investigate whether there is a correspondence between hidden layers of deep learning models and brain regions, particularly when processing language. In addition, they were curious to find out whether deep learning models can be used to synthesize brain data, which can then be used in other tasks.

"We investigated simple sentence representations learned by various neural network models ranging from simple sentence understanding neural networks to state-of-the-art language models and the brain," Jat said. "The main procedure to investigate the relationship was to predict brain representations from the model representations, the basic principle being, if both the representations carried similar information then they should be predictable from each other."

In their investigation, Jat and her colleagues considered several neural network architectures for word representation, including two recently proposed models called ELMo and BERT. They compared how these networks process particular sentences with data collected from human subjects using magnetoencephalography (MEG), a functional neuroimaging technique for mapping brain activity, as they read the same sentences. To begin with, they decided to use sentences with a simple syntax and basic semantics, such as "the bone was eaten by the

dog."

Interestingly, the researchers found that activations observed in the BERT model correlated the most with MEG brain data. Their findings also suggest that deep neural network representation can be used to generate synthetic brain data associated with new sentences, augmenting existing brain data.

"The language models predicted brain activity with good accuracy, which is very encouraging," Jat explained. "A recent transformer based representations (BERT) correspond best with brain encodings, which encourages further research in that direction. We also had multiple observations about what information (noun, verb, adjective, determiner) is better encoded in which type of network representations. Our hope is to uncover the 'secret sauce' to perfect language modeling algorithms."

According to Jat and her colleagues, this could be the first study to show that MEG recordings as humans are reading a sentence can be used to determine earlier words in that sentence. Moreover, they might have been the first to use deep neural network representations to successfully generate synthetic brain data.

The recent research by Jat and her colleagues provides new valuable insight into how sentences are represented in some state-of-the-art neural networks and how they are encoded by the human brain. Their observations could soon inform the development of new language processing algorithms, but also inspire further studies exploring the links between how the brain and deep learning models process information.

"In the future, we plan to continue studying core questions in language understanding (e.g., sentence understanding, word composition models, etc.) in the human brain using neural network models," Jat added.

More information: Sharmistha Jat et al. Relating simple sentence representations in deep neural networks and the brain. arXiv:1906.11861 [cs.CL]. arxiv.org/abs/1906.11861

Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2 [cs.CL]. arxiv.org/abs/1810.04805