The emergent algebraic structure of RNNs and embeddings in NLP (Cantrell, [2018])

This work is more of an application of abstract algebra than an application of category theory.

Overview.

In this paper, the authors hypothesize that the inherent algebraic structure of language may manifest in learned word vectors, and they perform a series of tests to assess this.

They begin by training word vectors with an RNN over a Twitter account prediction task. Since an RNN updates its hidden state each time it reads a word vector, they define the action of words on hidden state vectors to be the function $RNN(w_1, h_1) = h_2$ where h_2 is the result of the RNN update to hidden state h_1 after reading word w_1. They interpret this function to be a representation of the algebraic structure of words.

To assess the degree to which this algebraic structure satisfies some property P, they randomly sample words w from their vocabulary V and check whether they satisfy P. For example, in order to assess the degree to which the space satisfies closure under multiplication, they compute the vector $RNN(w_2, RNN(w_1, h))$ for a wide variety of random word pairs w_1, w_2 and random hidden states h, and then evaluate a loss function of the following form for each w_1, w_2, h:

$$\min_{\forall w_3 \in V} dist(RNN(w_2, RNN(w_1, h)), RNN(w_3, h))$$

They define an arbitrary small value such that if the average loss is smaller than that value, they consider the algebraic property to be "satisfied."

Given the properties that the trained vectors "satisfy", the authors conclude that words can be considered as elements of a Lie group. They then propose a technique to embed words directly in a Lie group by representing words with constrained matrices and postulate that this strategy could take advantage of the hierarchical structure of language.

Comments.

While other authors have explored algebraic and categorical formulations of models [4] and training procedures [3], there has been comparatively little research on the structures present in learned embeddings and weight vectors. Part of this is due to many sources of randomness and error in model fitting, compounded with the difficulty of efficiently representing the composition of multiple sources of approximation.

The authors of this paper sidestep this issue by treating the model fitting process as a black box and evaluating learned structure empirically. While their strategy is far from rigorous, it provides useful insight that may inform more formal analyses in the future.

The most unique component of this paper is the concept of an "axiomatic loss" that assesses the extent to which learned embeddings satisfy some algebraic quantity. The notion of a property being "approximately satisfied" is hand-wavy at best, but we may be able to extend it to a more rigorous formulation that identifies and separately handles different sources of approximation and randomness (dataset, initialization, etc).

One interesting aspect of this paper's construction is the choice to examine the structure of the word embedding space in terms of the actions of a word-parameterized RNN on the space of hidden vectors. This is in contrast to [2], which assumes a natural linear structure in the word embedding space and uses a tensor product to represent the composition of meaning. This paper's construction is more efficient in terms of the number of parameters, but is also more opaque because of the complexity of the RNN.

Unfortunately, the authors' empirical evidence is shaky as best. The dataset they use is extremely small, and word embeddings trained on Twitter data are both unstable and materially different from word embeddings trained on other corpora [5]. A more comprehensive approach might involve experiments over a wider range of corpora and embedding methods.

REFERENCES

[1] Cantrell, S. A. (2018). The emergent algebraic structure of rnns and embeddings in NLP. *CoRR*, abs/1803.02839.

[2] Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

[3] Fong, B., Spivak, D. I., and Tuyéras, R. (2017). Backprop as functor: A compositional perspective on supervised learning. *arXiv preprint arXiv:1711.10455*.

[4] McCullagh, P. (2002). What is a statistical model? *Annals of statistics*, pages 1225–1267.

[5] Shiebler, D., Belli, L., Baxter, J., Xiong, H., and Tayal, A. (2018). Fighting redundancy and model decay with embeddings. *arXiv preprint arXiv:1809.07703*.