Discovery of Natural Language Concepts in Individual Units of CNNs

Abstract

We show that individual units of CNNs learned on NLP tasks could act as natural language concept detectors. We define concept as building blocks of natural language sentences.

Approach: Alignment Score between Units and Concepts

1. Train CNN (e.g., ByteNet) on language task (e.g., Translation)
2. For each unit $u$, find top $k$ sentences which highly activate $u$.
   
   
   
   $s_i$: That would not be democratic.
   
   $s_j$: That would be cheap and it would not be right.
   
   $s_k$: This is not how it should be in a democracy.

3. Obtain candidate concepts from constituency parse tree of top $k$ sentences $s_k$

   
   Candidate Concepts
   
   $C_u = \{s_1, s_2, \ldots, s_k\}$

4. Compute Alignment Score between Units and Concepts

   
   $\text{alignment_score}(\text{concept } C_u, \text{unit } u)$

5. Rank Concepts and Units

   
   $\text{Score}(C_u, u) = \text{alignment_score}(C_u, u)$

   
   
   

Which Concepts are Sensitive to Each Unit?

Layer 1, Unit 1: Should

- $\mu_{s_1} - \mu_{s_2}$: how much and when?
- $\mu_{s_2}$: basis, when and how?
- $\mu_{s_3}$: we need to ask these, do we start?
- $\mu_{s_4}$: should we do this at this point?
- $\mu_{s_5}$: we are wondering now is it true or are they?

Layer 4, Unit 56:

- $\mu_{s_1} - \mu_{s_2}$: goals of representation
- $\mu_{s_3}$: how much and when?
- $\mu_{s_4}$: basis, when and how?
- $\mu_{s_5}$: we need to ask these, do we start?
- $\mu_{s_6}$: should we do this at this point?
- $\mu_{s_7}$: we are wondering now is it true or are they?

How Selectively does Each Unit Respond to Aligned Concepts?

Selectivity (unit $u$)

$\mu_{s_1} = \max_{a_i > \mu_{s_2}} a_i - \min_{a_i < \mu_{s_2}} a_i$

- Units are selectively responsive to specific concepts.
- Our method successfully aligns such concepts to units.

Which Concepts Appear More Often?

- These units can serve as detectors for specific natural language concepts.
- There are units capturing syntactically or semantically related concepts.

Concept Granularity Evolves with Layer

- As we move up the network, the concepts appear more coarse-grained.
- Concepts that (1) appear more often in training data & (2) have more influence on loss value are detected in more units.