Probing Black Box Models

Akhila Yerukola
Probing Task

- General approach to introspection techniques
  - Today, we’ll talk about probing sentence and word embeddings
- Informal Definition: “A probing task is a classification problem that focuses on simple linguistic properties of embeddings” [1]
- Probing methods are designed to evaluate the extent to which representations of language encode particular knowledge of interest

Probing Properties

- Probing tasks ask a simple question: to minimize interpretability problems
- Due to their simplicity, it is easier to control for biases in probing tasks than in downstream tasks
- Probing task methodology is agnostic wrt the encoder architecture
General Architecture

Idea:

- Given a pre-trained encoder trained on a given objective, obtain a single vector representation from it
- Probing model should be shallow: linear or 2-layer MLP

Example probing architecture [2]

Sentence Embedding Probes

**Sentence encoder:** BiLSTM-last/max, Gated CovNet

**Training task:**

- NMT
- SkipThought
- SNLI
- Untrained encoders with random weights

Sentence Embedding Probes

Probing Tasks:

1. Surface information: surface properties of the sentences
   a. SentLen
   b. Word Content

2. Syntactic information: Emb sensitive to syntactic properties of sentences they encode?
   a. Bigram Shift

3. Semantic information
   a. Tense
   b. Semantic Odd Man Out (SOMO)

Sentence Embedding Probes

Baselines:

1. Bag-of-Vectors: fastText
2. Arora style weighting [3]

Contextual Word Embeddings (CWR) Probes

- Probe contextual word embeddings like ELMO, BERT, OpenAI GPT
- **Idea:**
  - If a simple model can be trained to predict linguistic information about a word (e.g., its part-of-speech tag) or a pair of words (e.g., their semantic relation) from the CWR(s) alone, we can reasonably conclude that the CWR(s) encode this information.
Contextual Word Embeddings (CWR) Probes

Probing Tasks:

1. Syntactic information
   a. POS, NER
   b. Dependency Labeling, Constituency labeling

2. Semantic information (anything from WordNet)
   a. Semantic Role Labeling, Entailment, Concreteness
   b. Coreference resolution, Sentiment, Relation Classification

3. Local and Long range dependencies

Possible Applications

1. Word representations projects:
   a. What features of language do they capture, and what do they miss?
   b. CWR(s) capture more syntactic information than semantic information?

2. Language Modeling:
   a. If we encode sentences using a LM, what sort of properties do the sentence embeddings encode?
   b. How does the choice of pretraining task affect the vectors’ learned linguistic knowledge?

3. Interpretability and Transferability across various layers in contextualizers
   a. Lower layers encode syntax, higher levels encode semantics?
Moral of the Story

- Probe construction design: tradeoffs between probe complexity, probe task and hypotheses being tested
- Need for interpretability of models
- Keep probing!