CS224N : Lecture 18 - Future of NLP + Deep Learning

Extremely large models and GPT3

- GPT-1 : Improving Language Understanding by Generative Pre-Training
- GPT-2 : Language Models are Unsupervised Multitask Learners
- GPT-3 : Language Models are Few Shot Learners
    - 175 billion parameters
    - Trained on 500 billion tokens
    - Same architecture as GPT-2 (EXCEPT, locally banded sparse attention patterns
    - ▼ Meta-learning

      The model develops a broad set of skills and pattern recognition abilities at training time

    - Pros
        - Language Modeling
            - Penn Tree Bank
            - Story Completion
        - Knowledge Intensive Tasks
            - ex. Reading Comprehension
    - Cons
        - Structured problems that require multiple steps of reasoning
            - RTE, Arithmetic, Word problems, Analogy making

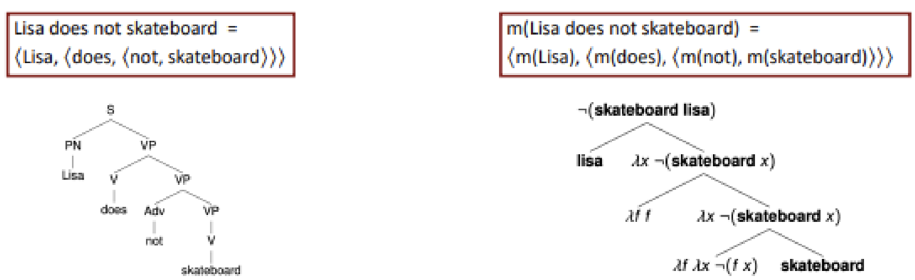
    - Limitations and Open Questions

- Seems to do poorly on more structured problems that involve decomposing into atomic / primitive skills:
  - RTE / arithmetic / word problems / analogy making
- Performing permanent knowledge updates interactively is not well studied.
- Doesn't seem to exhibit human like generalization (systematicity).
- Language is situated and GPT-3 is merely learning from text without being exposed to other modalities.

Compositional Representations and Systematic Generalization

Are neural representations compositional?

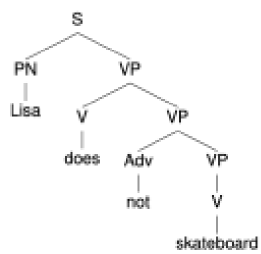
- According to Montague, Compositionality is about the existence of a homomorphism from syntax to semantics:



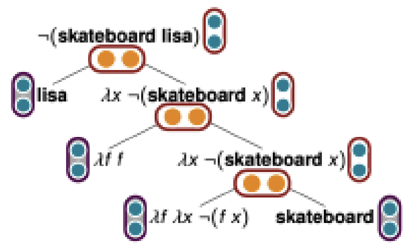
- Tree Reconstruction Error (TRE) [Andreas 2019]: Compositionality of **representations** is about how well the representation approximates an explicitly homomorphic function in *a learnt representation space*

- TRE [Andreas 2019]: Compositionality of representations is about how well the representation approximates an explicitly homomorphic function in a learnt representation space

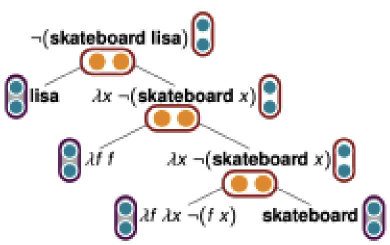
Lisa does not skateboard =
⟨Lisa, ⟨does, ⟨not, skateboard⟩⟩⟩



NN(Lisa does not skateboard) ≈
f(v(Lisa), f(v(does), f(v(not), v(skateboard))))



NN(Lisa does not skateboard) ≈
f(v(Lisa), f(v(does), f(v(not), v(skateboard))))



leaf vectors as well as the composition operator are *learnt by TRE*

Tree Reconstruction Error (TRE)

First choose :

- a distance function $\delta : \Theta \times \Theta \to [0, \infty)$ satisfying $\delta(\theta, \theta') = 0 \Leftrightarrow \theta = \theta'$
- a composition function $* : \Theta \times \Theta \to \Theta$

Define $\hat{f}_\eta(d)$, a *compositional approximation to* f with parameters η, as:

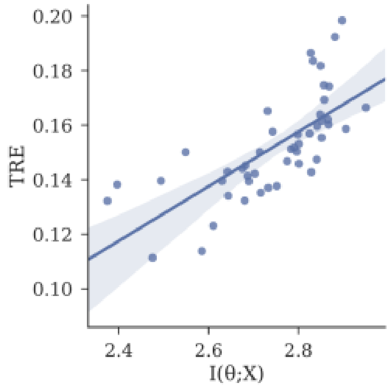
$$\hat{f}_\eta(d_i) = \eta_i \qquad \text{for } d_i \in \mathcal{D}_0$$
$$\hat{f}_\eta(\langle d, d' \rangle) = \hat{f}_\eta(d) * \hat{f}_\eta(d') \qquad \text{for all other } d$$

Given a dataset $\mathcal{X}$ of inputs x_i with derivations $d_i = D(x_i)$, compute:

$$\eta^* = \arg\min_\eta \sum_i \delta(f(x_i), \hat{f}_\eta(d_i))$$

Then we can define datum- and dataset-level evaluation metrics:

$$\text{TRE}(x) = \delta(f(x), \hat{f}_{\eta^*}(d))$$
$$\text{TRE}(\mathcal{X}) = \frac{1}{n} \sum_i \text{TRE}(x_i)$$



- This graph plots the mutual information between the input and the representation I(θ; X) against TRE.
- As the model learns (characterized by decreasing mutual information), we notice that the representations become more compositional!
- Overall, we observe that learning is correlated with increased compositionality as measured by TRE!

Do neural NLP models generalize systematically?

- Maximize *compound divergence* to create challenging train / test splits!
    - **Atoms**: primitive elements (entity words, predicates)
    - **Compounds**: compositions of primitive elements.

Train:
  Did Christopher Nolan produce Goldfinger?
  Who directed inception?
Test:
  Did Christopher Nolan direct Goldfinger?
  Who produced Goldfinger?

Atoms:
  produce
  direct
  inception
  goldfinger
  Christopher Nolan
  Who [predicate] [y]?
  Did [x] [predicate] [y]?

Compounds:
  Did Christopher Nolan [predicate] Goldfinger?
  Who directed [entity]?

- Basic Machinery for producing compositionally challenging splits:

Let $\mathscr{F}_A(\text{data}) \equiv$ normalized frequency distribution of atoms

Let $\mathscr{F}_C(\text{data}) \equiv$ normalized frequency distribution of compounds

Define atom and compound divergence as:

$$\mathscr{D}_A(\text{train}\,||\,\text{test}) = 1 - C_{0.5}(\mathscr{F}_A(\text{train})\,||\,\mathscr{F}_A(\text{test}))$$
$$\mathscr{D}_C(\text{train}\,||\,\text{test}) = 1 - C_{0.1}(\mathscr{F}_C(\text{train})\,||\,\mathscr{F}_C(\text{test}))$$

where,

$$C_\alpha(P\,||\,Q) = \sum_k p_k^\alpha q_k^{1-\alpha}$$

is the chernoff coefficient between two categorical distributions that measures similarity.

Goal:
Split data into train / test such that compound divergence is maximized and atom divergence is minimized!

- So do neural networks generalize systematically?
- Furrer 2020: "Pre-training helps for compositional generalization, but doesn't solve it"

Model	*CFQ (Maximum Compound divergence)*
T5-small (no pretraining)	21.4
T5-small	28.0
T5-base	31.2
T5-large	34.8
T5-3B	40.2
T5-11B	40.9
T5-11B-mod	42.1

Increasing #parameters

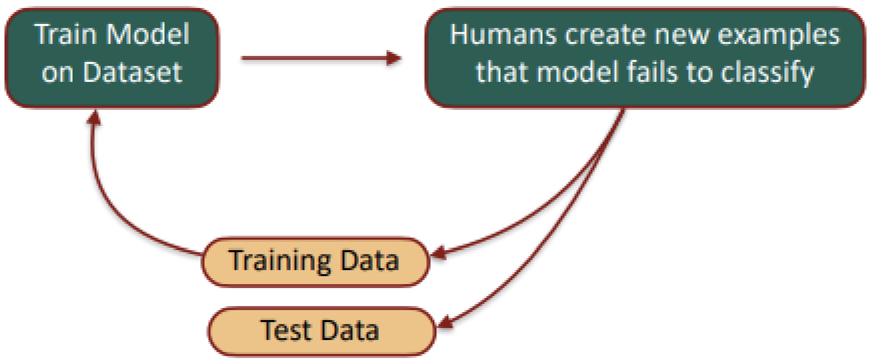
58

Source: Results from Furrer 2020 "Compositional Generalization in Semantic Parsing: Pre-training vs. Specialized Architectures"

Improving how we evaluate models in NLP

- Instead of testing models on static benchmarks, evaluate on an ever changing dynamic benchmark.

- Recent Examples:
  - Adversarial NLI by Nie et al. 2020
  - DynaSent by Potts et al. 2020
  - other related examples: “Build It, Break It” Workshop at EMNLP 17



Overview of dynamic benchmarks

1. Start with a pre-trained model and fine-tune it on the original train / test datasets
2. Humans attempt to create new examples that fool the model but not other humans
3. These examples are then added into the train / test sets and the model is retrained on the augmented dataset

- Main Challenges: Ensuring that humans are able to come up with hard examples and we are not limited by creativity.
- Current approaches use examples from other datasets for the same task as prompts

Grounding language to other modalities

- Many have articulated the need for using modalities other than text
- Bender and Koller [2020]: Impossible to acquire “meaning” (communicative intent of the speaker) from form (text / speech signal) alone
- Bisk et al [2020]: Training on only web-scale data limits the world scope of models.