Minimal-Pair Edits for Counterfactual Generation

He He, Nitish Joshi, Johnny Ma

New York University CILVR

11-4-20

Introduction

- Large-scale LMs for NLU have known, systematic gaps in data understanding (SNLI never-contradiction).
- Minimal Edit expert contrast sets can "fill" these gaps near local decision boundaries at **evaluation** time.
- Not clear that crowd sourcing counterfactuals can improve model generalizability at **training** time.

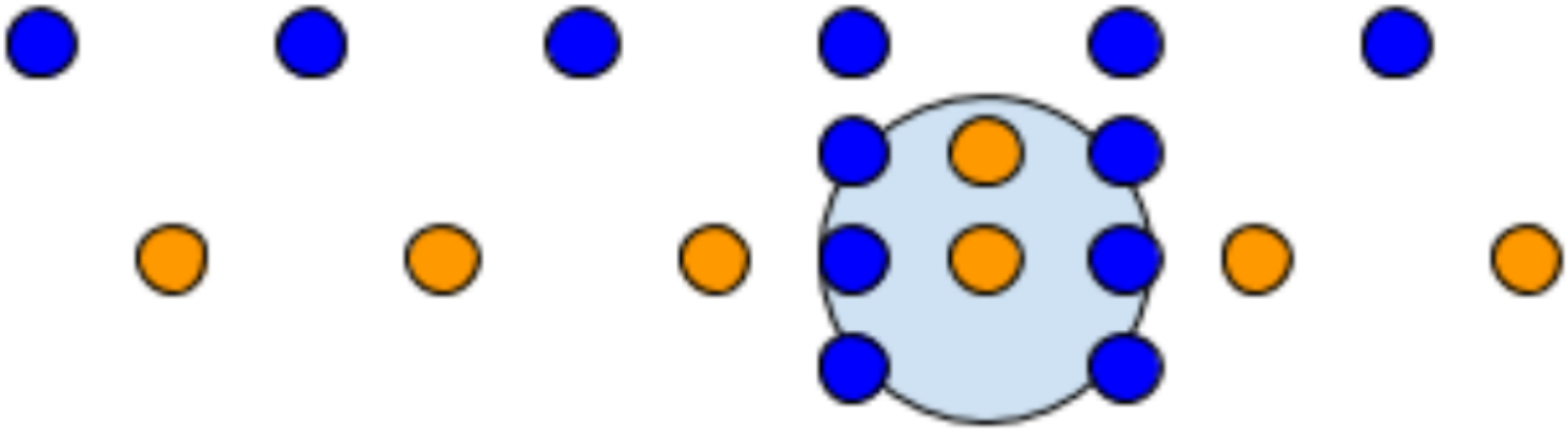

Figure 1: Local ball around test instance.

Automatically Generating Contrast Sets

- CLARE (Li *et al.* 2020) generates MP candidates with USE, then evaluates success of label attack.
  - Replacing and merging nouns are most successful attack types.
- LIT (Liu *et al.* 2020) use ERG/ACE parser as backbone to transform syntax and semantics.
  - Transforms 20% of SNLI instances, English only.

Research Goal

Can we generate minimal pair, expert quality counterfactual edits by masking and filling with BART?

MP Generation

- Idea is to learn a masking and filling policy to generate MPs that target local decision boundaries.
- With a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$, learn a generator for $x \rightarrow x'$ such that $f(x') \neq (\hat{y} = f(x))$.
- RL policy gradient for masking policy: $\ell(f(x'), f(x)) + d(x, x')$, $\ell =$ KL-divergence, $d =$ Bert-score?
- Reading Comprehension index based masking model?

Prelim BART on SNLI and IMDB

- Test on IMDB Sentiment Classification (Base BERT) and SNLI entailment (Base RoBERTA).
- For each data instance (sentence), mask N-Gram around content word (NOUN, ADJ, VERB), fill masked span with BART (fairseq base).
  - BART: Fill span with any set of tokens, beam = 10. Toss out stop words and long sequences. Not fine tuned.
- Get label prediction from fine-tuned model. Deem as "counterfactual" if model label changes from gold label.
  - 7% of IMDB MPs, 4% of SNLI MPs.

Example Data

Review Sentence	Token Change	Gold Label	New Label
but for the most part this is real amateur film-making that quickly becomes a pleasure to watch.	('painful', 'a pleasure')	Negative	Positive
This movie is a good choice for someone who likes to watch some awful deaths and practically torture.	('film', 'movie')	Positive	Negative

Premise	Hypothesis	Premise Token Change	Gold Label	New Label
A man in a suit is smoking a cigarette and talking on the phone.	A man makes a bet with his bookie.	('his cellphone.', 'the phone.')	Neutral	Entailment
Two men enjoying a night out together.	The two men are drunk.	('beer', 'night out')	Entailment	Neutral

Questions

- How to determine a high quality, true counterfactual?
- Which reward/learning scheme to use for counterfactual generation?
- Problems of model bias, is this generative approach feasible ?
- What to do with generated examples ?