Generating Natural Language Proofs with Verifier-Guided Search

Kaiyu Yang, Jia Deng, and Danqi Chen
Department of Computer Science, Princeton University

Challenges in Generating Valid and Relevant Steps

• Input: a hypothesis $h$ and a set of supporting facts $C = \{s_1, s_2, ..., s_n\}$ in natural language
• Output: a proof tree $T$ for deriving $h$ from a subset of $C$
• The root node is $h$; the leaf nodes are sentences in $C$
• Others are intermediate conclusions generated by the model

Stepwise Methods

• Condition on premises + hypothesis:
• Generate relevant steps conditioning on the hypothesis
• Generate multiple candidate steps via beam search
• Prover: Generate relevant steps conditioning on the hypothesis

Single-Shot Methods

• Generate the entire proof altogether

Better leverage compositionality and generalize to longer proofs

Achieved limited success on challenging proofs authored by humans, e.g., EntailmentBank [Dalvi et al. 2021]

NLProofs (Natural Language Proof Search)

• A new method for stepwise proof generation
• Prover: Generate relevant steps conditioning on the hypothesis
• Verifier: Mitigate hallucination by training an independent network to check the proof steps
• Proof search: Use the prover/verifier to generate the final proof with the optimal validity score

Verifier

• Input: A proof step (multiple premises, one conclusion)
• Output: A score in $[0, 1]$ calculated by finetuning RoBERTa
• Step scores are aggregated to calculate proof scores

Our Method: NLProofs

• Generating Natural Language Proofs

Experiments

• State-of-the-art on two benchmarks
  RuleTaker [Tajford et al. 2021]: Simple, synthetic profs
  EntailmentBank [Dalvi et al. 2021]: Challenging, human-written proofs

Main Results on EntailmentBank

• Single-shot baselines: EntailmentWriter
  Stepwise baselines: IRGR [Ribeiro et al. 2022], MetGen [Hong et al. 2022]

• Ablations: Importance of Stepwise Verifier-Guided Search

Large Language Models w/ In-Context Learning

• Single-shot proof generation with 7 in-context examples
• GPT-3 and Codex cannot solve the task out of the box

Validation accuracies on Task 2 of EntailmentBank