DSPy: Programming – not prompting – foundation models

Omar Khattab, Matei Zaharia, Christopher Potts, and many others

DSPy

Stanford NLP

Problem: You wouldn't dream of setting classifier weights by hand, but you're endlessly fiddling with prompt strings!

DSPy: Specify *what* you want your system to do and let optimizers figure out *how* best to achieve this.

HotPotQA

Program	Optimizer	GPT 3.5	Llama2-13b-Chat
dspy.RAG (with CoT)	BootstrapFewshot	42.3	38.3
dspy.ReAct	+ human reasoning	33.0	28.3
	BootstrapFewshot	31.0	24.7
	BootstrapFewshot×2	39.0	40.0
BasicMultiHop	BootstrapFewshot	48.7	42.0
	Ensemble	**54.7**	**50.0**

PyTorch design principles

```python
class BasicMultiHop(dspy.Module):
  def __init__(self, passages_per_hop):
      self.retrieve = dspy.Retrieve(k=passages_per_hop)
      self.generate_query = dspy.ChainOfThought("context, question -> search_query")
      self.generate_answer = dspy.ChainOfThought("context, question -> answer")

  def forward(self, question):
      context = []
      for hop in range(2):
          query = self.generate_query(context=context, question=question).search_query
          context += self.retrieve(query).passages
      return self.generate_answer(context=context, question=question)

multihop = BasicMultiHop(passages_per_hop=3)

# Optimize the bootstrapped demonstrations:
bootstrap = dspy.BootstrapFewShot(metric=exact_match).compile(
  multihop, trainset=qa_trainset, valset=devset)

# Fine-tune T5-large (770M) for near-SoTA:
multihop_t5 = dspy.BootstrapFinetune(metric=exact_match).compile(
  multihop, teacher=bootstrap, trainset=qa_trainset, target='t5-large')

# Jointly optimize instructions and demonstrations:
prompt_model = dspy.OpenAI("gpt-3.5-turbo")
multihop_mipro = dspy.MIPRO(prompt_model=prompt_model, metric=exact_match).compile(
  multihop, trainset=qa_trainset)
```

Haize Labs vicuna-7b-v1.5 attack success rates

Architecture	ASR
None (Raw Input)	10%
Architecture (5 Layer)	26%
Architecture (5 Layer) + dspy.MIPRO Optimization	44%

Table 1: ASR with raw harmful inputs, un-optimized architecture, and architecture post DSPy compilation.

Breakaway MEDIQA-CORR 2024 Winners

Rank	Team	Error Sentence Detection Accuracy
1	WangLab	83.6%
2	EM_Mixers	64.0%
3	knowlab_AIMed	61.9%
4	hyeonhwang	61.5%
5	Edinburgh Clinical NLP	61.1%
6	IryoNLP	61.0%
7	PromptMind	60.9%
8	MediFact	60.0%
9	IKIM	59.0%
10	HSE NLP	52.0%

dspy.MIPRO

Lively open-source project: http://dspy.ai