Building a Robust Question Answering Model with MAML

Dylan Cunningham, Doug Klink

Department of Computer Science
Stanford University

Stanford Computer Science

Problem

- Question answering (QA) is widespread, impactful NLP task
- Current QA models have trouble with domain adaptation
- Holy grail: a model that can generalize with only small amount of out-of-domain training data

Background + Methods

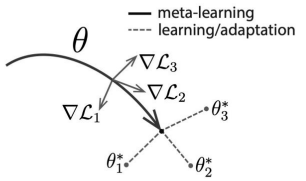

Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can quickly adapt to new tasks.[1]

Implemented Model-Agnostic Meta-Learning (MAML) training regime

- Each in-domain dataset treated as a separate training task
- Leveraged learn2learn library to streamline second-order backpropagation

- Tracked progress during training by periodically doing train/eval on out-domain data
- After training, finetuned the model with the three out-domain training sets (382 total question/answer pairs)

Algorithm 1 MAML Training Loop

Require: $p(\mathcal{T})$: distribution over tasks
Use pre-trained DistilBERT as initial θ
while not done **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
  for all $\mathcal{T}_i$ **do**
    Sample K examples from $\mathcal{T}_i$
    for n adaptation steps **do**
      Evaluate $\mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
      Compute adapted parameters: $\theta_i' = \theta - \alpha\nabla_\theta\mathcal{L}_{\mathcal{T}_i}(f_\theta)$
    end for
    Compute task loss: $\mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
  end for
  Update $\theta \leftarrow \theta - \beta\nabla_\theta \sum_i \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
end while

Experiments

- FS-Base ("Few-Shot Baseline"): finetune the baseline model with out-domain training data
- Meta-base: train new baseline model on in-domain data using Algorithm 1, then finetune on out-domain data

Idea: subdivide in-domain datasets

MAML shines when trained on many well-formed tasks

- Meta-NewsQA-Split: split NewsQA into two datasets to augment tasks, motivated by context length distribution (Figure 2)
- Meta-Even-Split: split all train datasets into small, medium, and large datasets to augment tasks

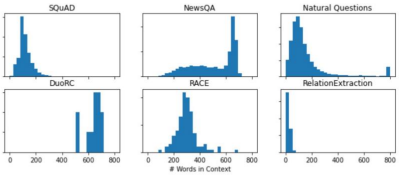

Figure 2: Context Length Across QA Datasets

Results

Model	EM	F1
Baseline	30.63	47.72
FS-base	**33.51**	**49.83**
Meta-base	**33.51**	47.37
Meta-NewsQA-Split	31.15	46.33
Meta-Even-Split	30.89	46.48

Table 3: Model scores on validation set

Analysis

- Challenging to beat FS-base, which uses classic training system
- **Likely need more tasks for meta-learning to show its strength**
- Potential to further optimize number of examples, meta learning rate, number of adaptation steps
- Task augmentation through dataset subdivision does not yield promising results

Conclusions

- **Meta-Learning seems promising for improving QA robustness**
- More well-formed tasks (e.g., all available QA datasets) and using the best hyperparameters could lead to increased performance over standard training regime (FS-base)

1. Figure from Chelsea Finn, Pieter Abbeel and Sergey Levine: *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*, 2017