Robust QA using Adversarial Learning

{Aasavari Kakne (adkakne), Samarpreet Singh Pandher (samar89), Vivek Kumar (vivkumar)} @stanford.edu

INTRODUCTION

Question Answering (QA) is a critical task for NLP applications such as conversational agents and search engines in which generalization to new domains is highly desirable

BACKGROUND

SOTA QA models often fail to generalize to new domains without significant fine-tuning. We aim to build a robust QA model using **adversarial learning approach.**

Lee et.al. achieved improved performance in terms of EM and F1 using Adversarial approach on MRQA Shared Task 2019.

DATASET

3 In-Domain Datasets : SQuAD, NewsQA, Natural Questions

3 Out-of-Domain datasets : DuoRC, RACE, RelationExtraction

METRICS

F1 score : the harmonic mean of precision and recall

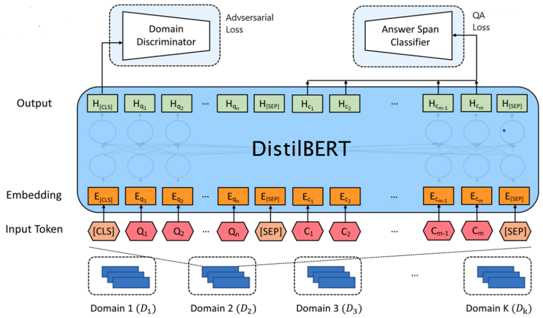
Exact Match : a binary measure (i.e. true/false) of whether the system output matches the ground truth answer exactly

METHODS

Our Adversarial Training approach consists of :

Generator Model : pre-trained DistilBERT

Discriminator Model : 3-layer MLP


Domain Discriminator
Advsersarial Loss
Answer Span Classifier
QA Loss
Output
$H_{[CLS]}$ H_{q_1} H_{q_2} ... H_{q_n} $H_{[SEP]}$ H_{c_1} H_{c_2} ... $H_{c_{m-1}}$ H_{c_m} $H_{[SEP]}$
DistilBERT
Embedding
$E_{[CLS]}$ E_{q_1} E_{q_2} ... E_{q_n} $E_{[SEP]}$ E_{c_1} E_{c_2} ... $E_{c_{m-1}}$ E_{c_m} $E_{[SEP]}$
Input Token
[CLS] Q_1 Q_2 ... Q_n [SEP] C_1 C_2 ... C_{m-1} C_m [SEP]
Domain 1 (D_1)
Domain 2 (D_2)
Domain 3 (D_3)
...
Domain K (D_k)

$$\mathcal{L}_G = \mathcal{L}_{QA} + \lambda\mathcal{L}_{adv}$$

EXPERIMENTS

For our adversarial experiments, we tuned

- Lambda (i.e. weight of adversarial loss)
- Dropout
- Hidden size of Discriminator

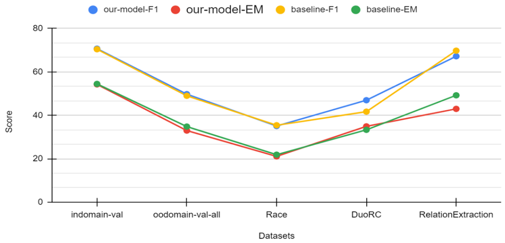
BEST MODEL

Best performance on Out-of-Domain Validation Set for

- Lambda = 0.01
- Dropout = 0.2
- Hidden Size of Discriminator = 768

RESULTS

Model Datasets	Our best model		baseline	
	F1	**EM**	**F1**	**EM**
indomain-val	**70.68**	54.25	70.43	**54.46**
oodomain-val-all	**49.75**	**32.98**	49.0	**34.82**



ANALYSIS

- Large Dropout and small Lambda boosts discriminator and forces generator to learn domain in-variant features.
- Score improvement on in-domain dataset doesn't improve score for all oo-domain datasets in general.

CONCLUSIONS

Adversarial Training helps the QA model generalise to out-of-domain datasets, and shows improved performance over the baseline on oo-domain dataset for F1 score by 0.75.

References :

[1] Seanie Lee, Donggyu Kim, and Jangwon Park. Domain-agnostic question-answering with adversarial training.CoRR abs/1910.09342, 2019