Exploring Domain Adversarial Training & Data Augmentation for Out-of-Domain Question Answering

Deveshi Buch, Caroline Choi, Melinda Zhu | Mentor: Kamil Ali

CS224N Project
RobustQA
Winter 2022

Background

- Current question-answering (QA) models such as Internet search engines have not matched human-level generalization [1]
  - Model performance on limited-resource domains can benefit from techniques encouraging generalization
- **Goal:** Develop robust QA system that can generalize to "out-of-domain" data
  - Explore multiple domain adversarial + data augmentation techniques
  - Baseline: pretrained DistilBERT-based QA model [2]

Data

- Format: **(question, context, answer)**
  - Answer span generated via (question, context) inputs
- **Indomain:** Natural Questions, NewsQA, SQuAD
  - Search log, Wikipedia, news article, crowdsource
- **Oodomain:** RelationExtraction, DuoRC, RACE
  - Film review, exam, synthetic, crowdsource, Wikipedia
- Training (>50,000 ex) & validation (>27,000 ex) sets comprised primarily of indomain data
  - Test set (>4,000 ex) entirely oodomain
- Where possible, model trained on indomain data and finetuned on limited oodomain examples to boost performance

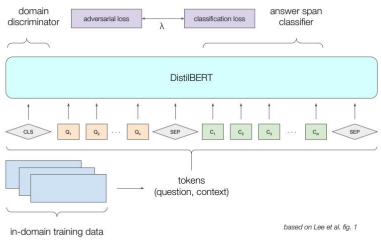
References

[1] CS224N Staff. Cs224n default final project: Building a qa system (robust qa track), 2022.
[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv, abs/1910.01108*, 2019.
[3] Seanie Lee, Donggyu Kim, and Jangwon Park. Domain-agnostic question-answering with adversarial training, 2019.
[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org, 2015.
[5] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. Advances in neural information processing systems, 29, 2016.
[6] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.
[7] Amane Sugiyama and Naoki Yoshinaga. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44. ACL, 2019.

Methods

Domain Adversarial Training

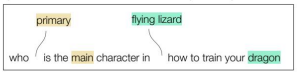
- **QA model & discriminator, compete to learn domain-invariant features**
- DAT based on [3]; gradient reversal layer (GRL) based on [4]; label smoothing implemented [5]
- QA model initialized w/ pretrained DistilBERT
- Discriminator architecture:
  [Linear, ReLU, Dropout] x 3 , [GRL]
- QA model loss: conventional QA loss + adversarial loss
- QA & discriminator trained alternatively
- In-domain training, oo-domain finetuning



Data Augmentation

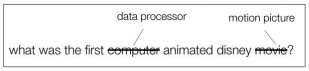
Random Insertion (RI)

Add synonyms in randomly-chosen locations within questions for robustness to varied phrasing.



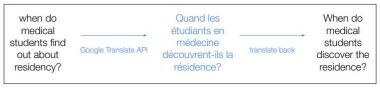
Synonym Replacement (SR)

Replace random words in questions with synonym for **robustness to varied wording and meaning**.



Back Translation (BT)

Capture meaning beyond language-specific features to preserve original semantics without significant word alterations.



real examples from our work shown; based on [6,7]; BT for oodomain finetune only due to space constraints

Experiments

Our DAT model with step learning rate scheduling & finetuned on out-of-domain data, improved upon the baseline.

Model	Indomain val (EM / F1)	Oodomain val (EM / F1)
Baseline	54.54 / 70.31	34.55 / 49.88
Augmented (synonym replacement) baseline	54.40 / 69.91	34.50 / 49.78
DAT finetuned on `oodomain_train_aug` (SR)	47.88 / 63.93	34.82 / 49.07
DAT (w/ LRS) no finetuning	**55.29 / 71.39**	31.41 / 47.90
DAT (w/ LRS) finetuned on `oodomain_train`*	49.29 / 65.59	**35.86 / 50.19**
DAT (w/ LRS) finetuned on `oodomain_train_aug`	48.37 / 64.24	33.25 / 47.12

*test performance: 40.09 / 57.79

LRS = learning rate scheduling; subset of results shown; see report for full experiments & analysis

Analysis

- DAT w/ learning rate scheduling & `oodomain_train` finetuning improved oodomain val performance
- Finetuning for longer on `oodomain_train` boosted performance
- Freezing domain discriminator before finetuning improved oodomain val performance
- GRL appears to discourage domain-specific feature learning
- SR was most effective out of all data augmentation techniques because it is likely to preserve meaning

dev results by dataset

DAT w/ LRS, finetuned on `oodomain_train`	DuoRC	RACE	RelationExtraction
F1	36.20	30.35	77.40
median context length (char)	3839	1632	129

Question: on which instrument s was introduction and rondo capriccioso created to be played on
Context: The Introduction and Rondo Capriccioso in A minor (French: Introduction et Rondo capriccioso en la mineur), Op. 28, is a composition for violin and orchestra written in 1863 by Camille Saint-Saëns for the virtuoso violinist Pablo de Sarasate.
Expected: violin **Answer:** violin and orchestra

Question: Who plays Jasper and Horace? **w/ SR:** Who recreates Jasper and Horace?
Context: Cruella dismisses Anita and vows revenge against her and Roger. She has her henchmen, Jasper and Horace (Hugh Laurie and Mark Williams) break into their home and steal the puppies while Roger and Anita are gone for a walk.
Expected: Hugh Laurie and Mark Williams **Answer:** Hugh Laurie and Mark Williams

Top: Erroneous example output from our best model. Able to identify keywords & associated words, but doesn't actually understand what question is asking.
Bottom: Example of robust augmentation result. SR retains original question meaning such that the system is able to deduce the correct answer with both prompts.

Conclusions

Improvements with DAT + GRL + finetune on `oodomain_train`. Data augmentation not always effective: can confuse the model with grammatically incorrect phrasings and unfitting synonyms.

Future work:

- DAT: different discriminator architectures, LR schedulers
- Data augmentation techniques for `oodomain_train`
  - Different translating mechanisms for back translation
  - Selective interpolation (LISA)

We would like to acknowledge our mentor and course staff for support and guidance + Azure for computing resources.