With Such Friends Who Needs Enemies?

Using a "Friendly" Question Type Discriminator for Robust QA Training

224N Default Project | RobustQA Track
Author: Filippos Nakas | Advisor TA: Anglelica Sun

Problem: Robust QA

- State of the art transformer-based QA models fail to generalize well for doing QA on different domains that those they were trained on.
- To avoid solving this by computationally expensive and/or data inefficient ways we need to find training techniques that allow the model to attain domain-invariant representational power that will allow it to leverage the features that unite language across different domains of discourse.

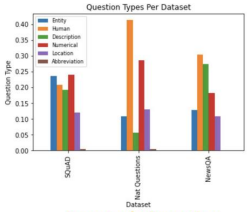
Past Approach: Domain-Adversarial Learning

- Adversarial learning was introduced in Goodfellow(2014) for generative models.
- Adapted for RobustQA in Lee(2019).
- Add a feed-forward adversarial discriminator network that is simultaneously trained to read BERT's hidden states to guess the domain of the encoded input data. In both the older paper and this one, it only reads the last hidden state of the "CLS" token.
- Add a term in the QA model's loss function that motivates it to "fool" the adversarial network and make it unable to guess better than random.
- A good candidate is the KL-divergence between a uniform distribution over domains based on their size and the SoftMax predictions of the adversary.
- The QA model is thus forced to learn hidden representations that are not recognizably linked to the properties of a particular domain, because that would provide information to the adversary that helps it guess correctly.

Potential Issue: Question-Type Blindness

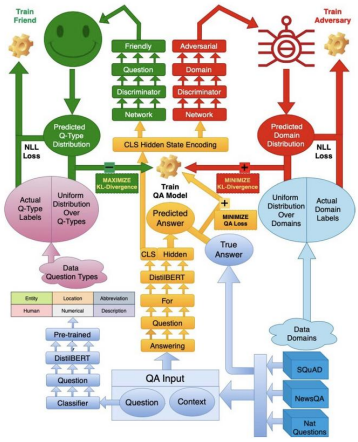
- Observation: question types are not similarly distributed for each dataset.
- Hypothesis 1: learning to leverage the data's question types is beneficial to the robustness of the model, because the relation of question type and answer type is to a great extent extent domain-invariant.
- Hypothesis 2: the observation suggests that if the model chooses representations based on the input's question type, the adversarial model could use their correlation to better predict the input's domain.
- Conclusion: domain adversarial training may be teaching the QA model to become question-type "blind" and thus indirectly damage its robustness

Question classification produced by script written by Danilo Croce (https://github.com/crux82/AILC-lectures2021-lab)



Research Question

Can we find a way to leverage question type information without compromising the adversarial objective?



New Method: Add "Friendly" Question Type Discriminator

- The data used for the model are tuples (c_i^l, q_i^l, y_i^l) denoting the i^{th} context, question, span data-point triplet in the l^{th} training domain out of total of K training domains.
- y_i^l is itself a tuple of $y_{i,s}^l$ and $y_{i,e}^l$ denoting respectively the start and end position of the gold standard answer span
- QA-specific loss function used in the paper is the average of the sum of negative log likelihood for the predicted start and end position over all data-points N split among the k in-domains

$$\mathbf{L_{QA}} = -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k}[logP_\theta(y_{i,s}^k|c_i^k, q_i^k) + logP_\theta(y_{i,e}^k|c_i^k, q_i^k)]$$

- QA model must also learn to minimize its adversary's success by minimizing the KL-divergence of the adversarial discriminator's softmax with the uniform distribution over domain frequency for the data.

$$\mathbf{L_{adv}} = -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k} KL(U(l)||logP_\phi(l_i^k|h_i^k))$$

- In our new version, the QA model must additionally learn to force the friendly QT network's to make non-random predictions by maximizing it's KL-divergence with the uniform distribution over Q-type frequencies, where M is the number of question types and q^m is the m^{th} question label:

$$\mathbf{L_{friend}} = -\frac{1}{N}\sum_{m=1}^{M}\sum_{i=1} KL(U(q)||logP_\omega(q_i^m|h_i^m))$$

- These three loss functions will now combined into one using the hyper-parameters λ_{adv} and λ_{friend}:

$$\mathbf{L} = \mathbf{L_{QA}} + \lambda_{adv}\mathbf{L_{adv}} - \lambda_{friend}\mathbf{L_{friend}}$$

Development Experiments

One Shot OOD

Run Type	F1	vs Baseline	EM	vs Baseline
Baseline	46.71	0.00	30.37	0.00
Adversarial	44.59	-2.12	28.27	-2.10
Adversarial Friendly	49.65	+2.94	33.25	+2.88

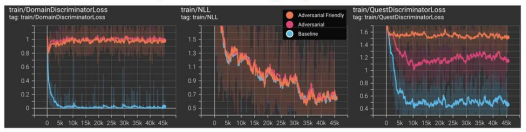
Fine-Tune OOD

Run Type	F1	vs Baseline	EM	vs Baseline
Baseline	47.4	0.00	32.2	0.00
Adversarial	50.34	+2.94	36.13	+3.93
Adversarial Friendly	51.99	+4.59	36.13	+3.93

Performance on Test Set:
EM: 41.032
F1: 58.364

- Adversarial training performs badly without extra fine-tuning on the OOD data (worse than baseline) but improves after finetuning.
- Might be because it obstructs the QA model from gaining specialized enough representations that are adequate to the task.
- At the same time may for the reason affording greater flexibility in acquiring Adding the friendly component to the training gives a jump to both for one-shot and fine-tuning possibly by allowing it to leverage useful but domain invariant relationships between types of question and answer.
- But how could it do this if the objectives of the two networks are contradictory? If Friend can predict the question can't Adversary increase its chances using the correlation between domain and question types?
- **Puzzle:** when we changed the friendly part of the QA net's objective from maximizing the Friend's KL-divergence from the uniform distribution to minimizing its actual NLL Loss **all performance gains bizarrely disappeared.**

Analysis: ....betraying our Friend?



- As hypothesized, the model naturally tends towards representations linked to question types and adversarial training disincentivizes it from doing so.
- But adding what was meant as a "friendly" objective to dampen this effect extenuates it further. The question classifier performs even worse!
- KL divergence objective forces the QA model to "show" Friend representations that lead to make a determinate choice: **but not necessarily the correct one.**
- Possible Explanation: QA Net has found a compromise: if helping the friendly network implicitly helps the adversary, fool the friend as well!
- Choose representations that point the question classifier to a particular wrong question type. This simultaneously maximizes the Friend's KL-divergence while giving no indirect hints to the adversary about the data's domain.
- **The increase in OOD performance suggests that the QA net might effectively be learning to leverage certain connections between different question types that allow it to leverage its knowledge in answering question of one type to answering questions of the other.**
- Sounds particularly promising for achieving robust performance in new domains with different question type distributions from the training data!

References

Donggyu Kim Seanie Lee and Jangwon Park. Domain-agnostic question-answering with adversarial training. In *MRQA@EMNLP*, 2019.

Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Yoshua Bengio Ian Goodfellow, Jean Pouget-Abadie. Generative adversarial nets. In *NIPS*, 2014.