Domain Adversarial Training for Robustness in Question-Answering Models

Abhay Singhal[1], Navami Jain[1], Shayana Venukanthan[1]

[1]*Department of Computer Science, Stanford University*

Background

- While the Question Answering (QA) task is a promising application of NLP, its ability to generalize to new datasets remains a challenge.
- Models tend to overfit to specific datasets, or domains, they are trained on, decreasing their utility in real world applications.
- In the past, adversarial training has been applied to produce domain-agnostic question-answering. See figure 1 below.

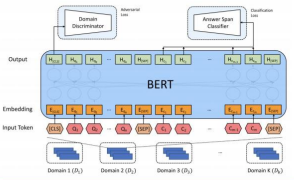


Figure 1: Training procedure for learning domain-invariant feature representations. The discriminator is trained to predict domain of the dataset based on the output [CLS] token. The model classifier predicts the appropriate answer while fooling the discriminator. Taken from Lee et. al 2019.

Task: Partial Domain Independence

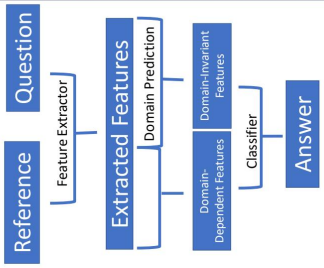


Figure 2: Partial Domain Independence

- We explore creating partially domain-invariant models that improves performance of the model while remaining generalizable
- Our final loss function for the QA model can be written as $\mathcal{L}_{QA} + \lambda\mathcal{L}_{adv}$
- The influence of the discriminator loss is set by the hyperparameter λ. L_{QA} represents classification loss while L_{adv} is discriminative loss

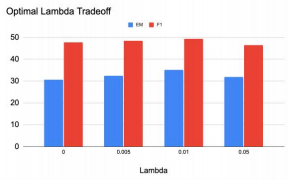
Adding an Adversarial Component to the Baseline

Model	Exact Match	F1 Score
Baseline w/o Adv. Training	30.63	47.72
Adv. Training with SGD	31.152	46.896
Adv. Training with Adam Optimizer	35.079	49.321

- Including the adversarial component improved both EM and F1 score.
- Using Adam optimizer led to further improvements.

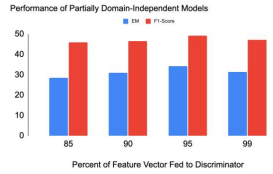
Approach

Optimal Lambda Tradeoff



- The adversarial network was implemented using different values of lambda in the loss function $\mathcal{L}_{QA} + \lambda\mathcal{L}_{adv}$
- A lambda value of 0.01 led to the best performance metrics, with an EM value of 31.94 and an F1-score of 49.321.
- This value was used in subsequent model trainings.

Partially Domain-Invariant Models

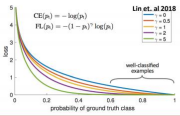


- Model performance was assessed when features were trained to be partially independent of the domain.
- In each case, a component of the feature vector (β) was trained on the discriminator while the remaining component was directly passed to the classifier.
- The model demonstrated optimal performance on the evaluation set when 5% of the features were withheld from the discriminator.
- This suggests some domain knowledge does not compromise generalizability and in fact improves model performance.

Model Refinement with Wasserstein Distance

- We explored model improvements by replacing the Kullback-Leiber (KL) divergence with a Wasserstein distance measure to adversarially train the discriminator function.
- At a high level, the Wasserstein distance is a distance metric between two probability distributions, defined as $W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$
- $\Pi(P_r, P_g)$ is the set of all joint distributions over **x** and **y** such that the marginal distributions are equal to P_r and P_g.
- In this case, the predicted domain from the discriminator is representative of the source domain and a uniform distribution is the target domain.

Handling Class Imbalance with Focal Loss



- Focal loss was implemented to handle imbalance in predictions caused by class imbalance in the training set.
- It adds a factor $(1-p)^\gamma$ to the standard cross entropy term, allowing the loss function to apply more focus on misclassified examples.

Final Results

Focal Loss

Gamma	Alpha	Exact Match	F1 Score
0.3	0	30.63	45.69
1	0	30.63	47.49
2	0	31.68	47.33
2	0.25	33.77	48.92
3	0	31.94	47.01

In concordance with the results from developers of focal loss (Lin et. al 2018), a gamma value of 2.0 and alpha value of 0.25 provided the best performance.

Wasserstein Distance

Lambda	Adversarial Loss Training	Sampler Type	Exact Match	F1 Score
0.01	Wasserstein	Weighted	31.94	48.49
0.05	Wasserstein	Random	32.72	49.24
0.01	KL-Divergence	Weighted	29.32	43.78
0.01	KL-Divergence	Random	35.079	49.321

- While KL-Divergence demonstrated optimal performance on the models tested, a lambda value of 0.05 improved performance on models implemented with Wasserstein distance.
- This suggests hyperparameters must be optimized specifically for application of Wasserstein. This is a potential future direction of this research.

Combining Focal Loss and Wasserstein Distance

- When both techniques are combined (with hyperparameters **λ = 0.01, α = 0.25, β = 0.95,** γ = **2.0**), we achieve our best performance, with F1=**51.16** and EM=**35.08** on the dev set and an F1=**60.069** and EM=**41.789** on the test set.

Summary

- Compared to our baseline model trained without an adversarial component, adding the discriminator improved performance in terms of F1-Score and Exact Match (EM). Developing features with partial domain independence also improved the model's performance on unseen data.
- While our dataset was heavily imbalanced, it remains unclear whether focal loss improved overall performance.
- While several combinations of hyperparameters were tested, a more extensive and organized hyperparameter search needs to be conducted to make conclusions on the utility of Wasserstein distance and focal loss.

References

Lee, S., Kim, D., & Park, J. (2019). Domain-agnostic Question-Answering with Adversarial Training. *ArXiv:1910.09342 [Cs]*. http://arxiv.org/abs/1910.09342.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018). Focal Loss for Dense Object Detection. *ArXiv:1708.02002 [Cs]*. http://arxiv.org/abs/1708.02002.

Shen, J., Qu, Y., Zhang, W., & Yu, Y. (2018). Wasserstein Distance Guided Representation Learning for Domain Adaptation. *ArXiv:1707.01217 [Cs, Stat]*. http://arxiv.org/abs/1707.01217