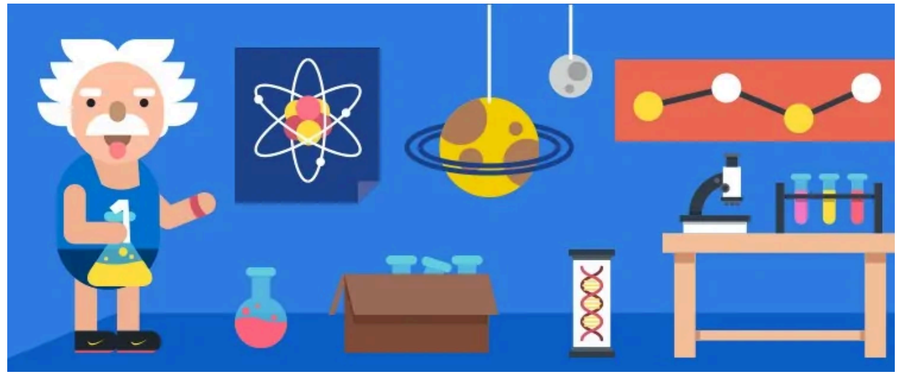
iNLP Project

TEAM 16 PRESENTS

CROSS DOMAIN QA USING DOMAIN ADAPTATION

Team 16 (Team Name) -
Swarang, Hari, Ishan

WHY?



- Cross-domain question answering refers to the task of developing systems or models that can effectively answer questions across different domains or topics.

- A ”domain” typically represents a specific subject area, and each domain may have its own set of linguistic nuances, terminologies, and contextual intricacies.

- This task poses a considerable challenge as it involves adapting language models to diverse and potentially unfamiliar domains. Our study aims to investigate different adaptation methods and assess their effectiveness in the context of question answering tasks.

EXPERIMENTAL SETUP

Our setup includes -

- Delineating the datasets utilized
- Detailing the models employed for analysis
- The methodology employed to evaluate the effectiveness of various adaptation techniques.

1 Datasets

- TriviaQA
- SQuAD
- NewsQA

2 Models

- T5
- BERT
- GEMMA-2B it

METHODOLOGY

Our focus centers on two distinct domain adaptation methodologies:

- Fine tuning and data augmentation coupled with fine-tuning.

- Our aim is to investigate the efficacy of these methods in enhancing the performance of models across varying domains.

FINE-TUNING

- Fine-tuning involves the process of taking a pre-trained model and further training it on domain-specific data.

- Fine-tuning is particularly effective in domain adaptation as it enables the model to learn domain-specific features, thereby enhancing its performance on tasks within the target domain.

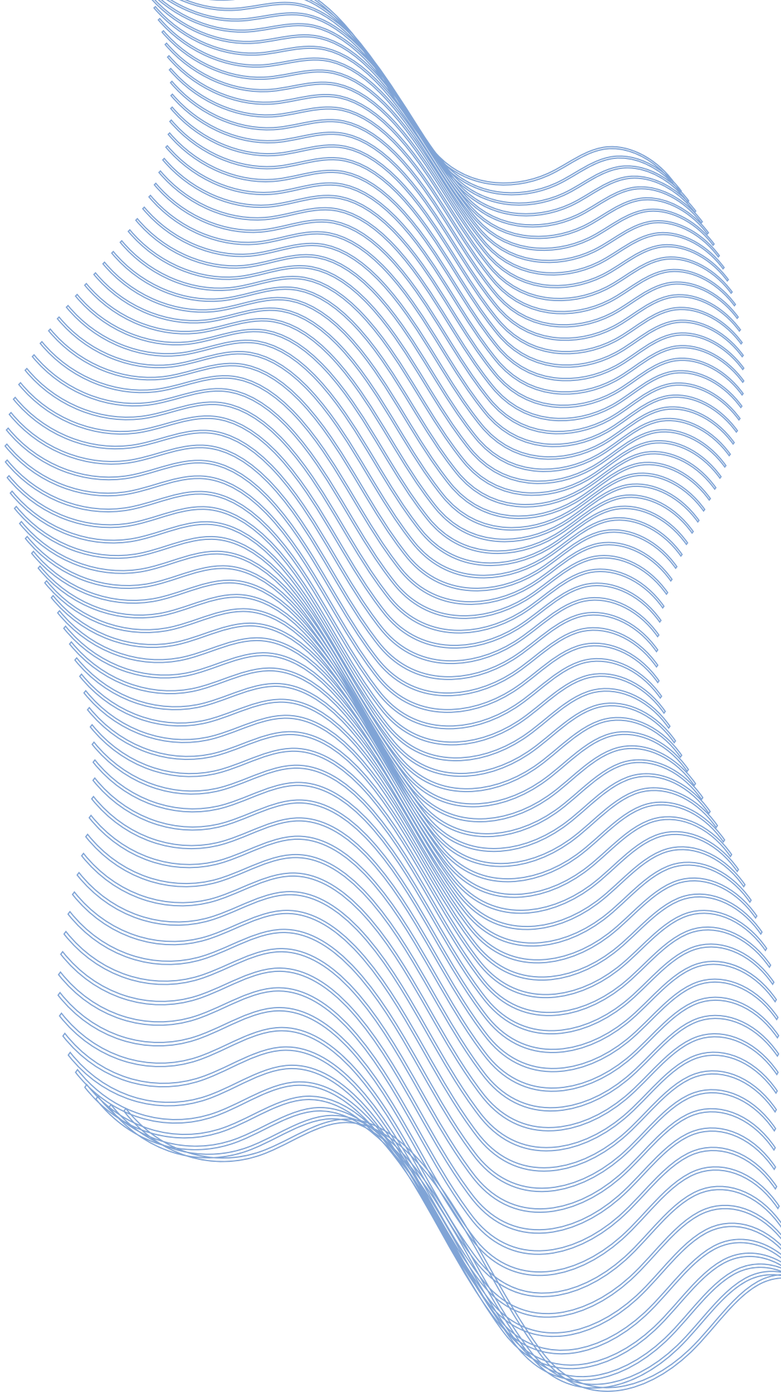
- We select 10,000 data points from the training set of both the base domain and each target domain for fine-tuning the models.

DATA AUGMENTATION

- Data augmentation is a technique used to artificially increase the size and diversity of the training data by applying various transformations to the existing examples. eg. Synonym replacement, Paraphrasing, etc.

- Augmenting the training data with synthetic examples generated from the existing data, the model can learn more effectively from the available information, leading to improved performance when applied to the target domain.

- To facilitate the paraphrasing process, we leverage the PARROT library.



CHALLENGES FACED

“Challenges are what make life interesting and overcoming them is what makes life meaningful.”

- Joshua Marine

1. Summarization because of varying context sizes across the models. Solved through using sumy.

2. Finetuning experiments were computationally intensive, especially for models with large parameter sizes, requiring significant compute power to complete the process effectively

3. Dissimilarities between these datasets can affect the evaluation due to varied structure of datasets, not trivial to evaluate the results.

RESULTS

- Increased performance on the f1 score when fine-tuned on a domain and tested on the same domain.
- Dip in performance for domain on which it has not been fine-tuned
- Gemma-2B instruct performs well but has an exact match score of zero

	Test Dataset	Exact match	F1 score
Finetuned on SQUAD	SQUAD	0.627	0.985
	NewsQA	0.407	0.915
Finetuned on NewsQA	SQUAD	0.422	0.894
	NewsQA	0.593	0.963

Table 4: Results from fine-tuning BERT on target domains

	Test Dataset	Exact match	F1 score
Finetuned on SQUAD	SQUAD	0.592	0.793
	NewsQA	0.417	0.644
Finetuned on NewsQA	SQUAD	0.433	0.683
	NewsQA	0.624	0.813

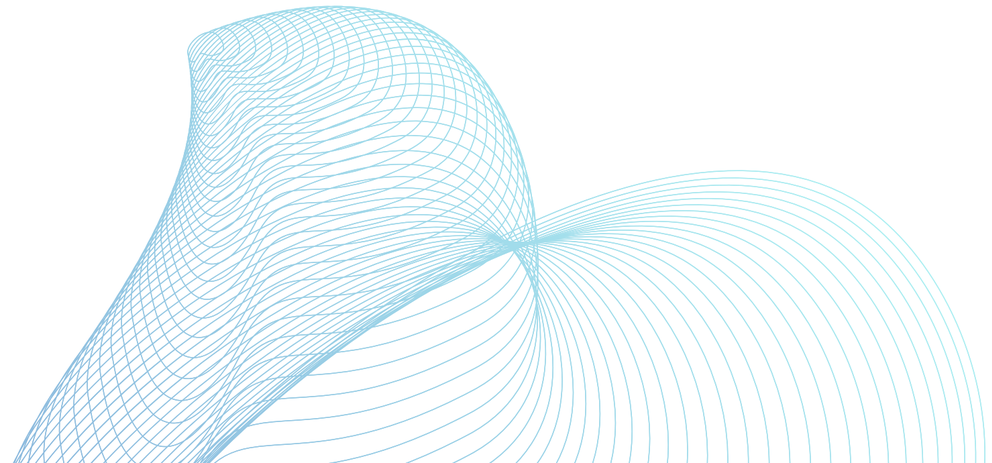
Table 5: Results from fine-tuning T5 on target domains

	Test Dataset	Exact match	F1 score
Finetuned on SQUAD	SQUAD	0.0	0.837
	NewsQA	0.0	0.734
Finetuned on NewsQA	SQUAD	0.0	0.705
	NewsQA	0.0	0.818

Table 6: Results from fine-tuning Gemma-2B on target domains

RESULTS

- Slight increase in performance of models due to data augmentation coupled with finetuning
- BERT performing the best among all the three models followed by Gemma-2B instruct and lastly T5
- Again, Gemma has an exact match scores of zero



	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.622	0.988
SQUAD	NewsQA	0.413	0.925
Finetuned on	SQUAD	0.430	0.882
NewsQA	NewsQA	0.618	0.971

Table 7: Results from fine-tuning and using data augmentation BERT on target domains

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.599	0.817
SQUAD	NewsQA	0.412	0.747
Finetuned on	SQUAD	0.448	0.743
NewsQA	NewsQA	0.641	0.822

Table 8: Results from fine-tuning and using data augmentation T5 on target domains

	Test Dataset	Exact match	F1 score
Finetuned on	SQUAD	0.0	0.845
SQUAD	NewsQA	0.0	0.734
Finetuned on	SQUAD	0.0	0.711
NewsQA	NewsQA	0.0	0.839

Table 9: Results from fine-tuning and using data augmentation Gemma-2B on target domains

ANALYSIS

In analyzing the observed results, we scrutinize the experimental setup, methodological approach, and potential external factors that could influence the outcomes. Factors such as dataset characteristics, model architecture, and the intricacies of adaptation techniques are considered to gain deeper insights into the observed performance trends. By examining these aspects comprehensively, we aim to elucidate the underlying reasons behind the observed outcomes and inform future research directions.

1

Improved Performance with Data Augmentation and Fine-Tuning

- Increased diversity and volume of training samples
- Robust learning and better generalization to unseen data
- Performance gains may not meet expectations

2

Impact of Summarization on Model Performance

- Prevents abrupt context truncation
- May impact quality and relevance of generated answers
- Potential limitations on overall performance improvement

3

Unique Behavior of Gemma-2B

- Generative nature leads to significant differences from ground truth
- Lower exact match scores due to divergence in responses
- Long or irrelevant responses contribute to lower exact match scores

Thank You

ANY Questions?