CS 772 – FINAL PROJECT EVALUATION

Team ID	25
Member Name	Roll No
Aditya Pande	22M2108
Balbir Singh	22M0747
Tarun Bisht	23D0386
Vivek Kumar Trivedi	22N0457
Date	5 May 2024

Problem Statement : Parameter Efficiency

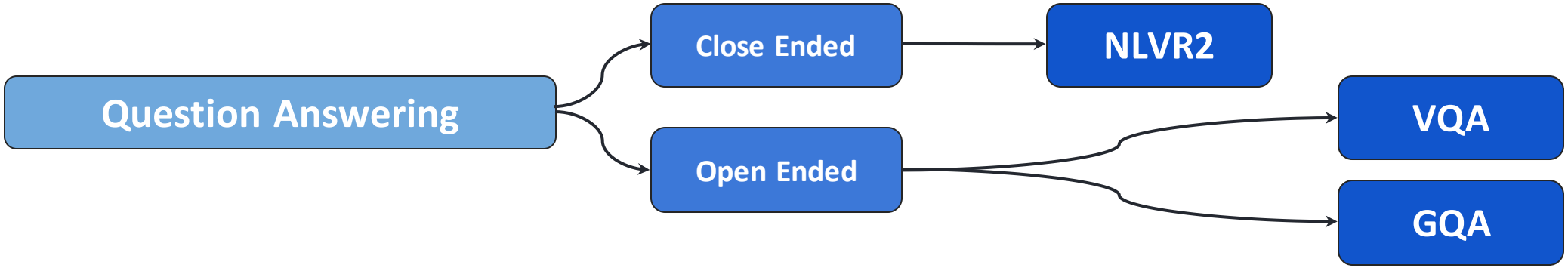
- **Title : Parameter Efficient Training for Multimodal Tasks**
- **Input : An image (or pair for NLVR)**

+

Question regarding the image/s

- **Output : Answer to the question**

Problem Statement : QnA



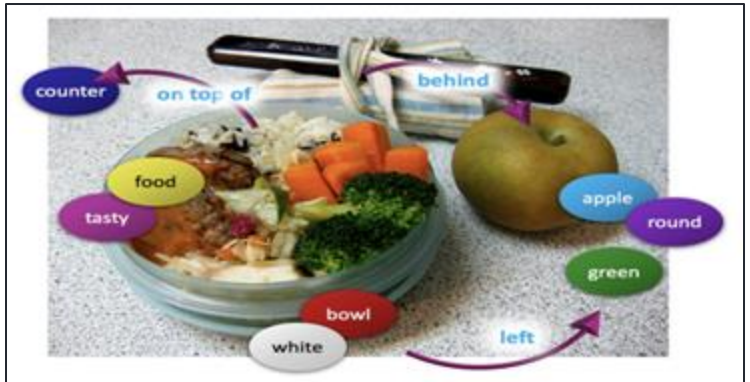
NLVR2 (NL for Visual Reasoning)

Simple True/False questions on a pair of images

VQA (Visual QnA)

Open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer

GQA (General QnA)



More fine tuned question

- Is the bowl to the right of the green apple?
- What type of fruit in the image is round?

Motivation for the problem

- **Multimodality:** Visual Semantic Mapping, integrating visual and textual modalities
- **Parameter Reduction:** without compromising model performance
- **Single Model (Hyperformer):** for multiple tasks
- **Language Phenomenon:** Semantic Understanding, Common sense reasoning
- **Semantic Understanding:** Understanding meaning of question being asked.
- **Commonsense Reasoning:** Answering Open Ended Questions.
- **QnA with Retrieval:** answering using open knowledge domain.

Literature Survey

Topic	Reference
Adapter	Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. **Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning**, pages 2790–2799. PMLR, 2019
Hyperformer	Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. **Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks.** In Annual Meeting of the Association for Computational Linguistics, 2021.
Compacter	Rabeeh Karimi mahabadi, James Henderson, and Sebastian Ruder. **Compacter: Efficient low-rank hypercomplex adapter layers.** In Thirty-Fifth Conference on Neural Information Processing Systems, 2021.
VL-Adapter	Sung, Y.L., Cho, J. and Bansal, M., 2022. **VL-adapter: Parameter-efficient transfer learning for vision-and-language tasks.** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5227-5237)
BART	Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L., 2019. **Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.** arXiv preprint arXiv:1910.13461.
ResNet	He, K., Zhang, X., Ren, S. and Sun, J., 2016. **Deep residual learning for image recognition.** In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
CLIP	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. **Learning transferable visual models from natural language supervision.** In ICML, 2021
LORA	Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan AllenZhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. **Lora: Low-rank adaptation of large language models.** CoRR, abs/2106.09685, 2021
Prompt tuning	Brian Lester, Rami Al-Rfou, and Noah Constant. **The power of scale for parameter-efficient prompt tuning.** In EMNLP, 2021
ColBERT	Khattab, O., & Zaharia, M. (2020, July). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (pp. 39-48).
FLMR	Lin, W., Chen, J., Mei, J., Coca, A., & Byrne, B. (2024). Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, *36*.

Data Handling

- Image features are extracted from CLIP [ResNet 101 backbone]
- BART tokenizer are used
- Image size 224 x 224
- More about each dataset in the future slides

Dataset	Data source URL	Data statistics
VQA	https://visualqa.org/download.html	• 204,721 COCO images, • At least 3 questions (5.4 questions on average) per image, • 10 ground truth answers per question, • 3 plausible (but likely incorrect) answers per question
GQA	https://cs.stanford.edu/people/dorarad/gqa/download.html	• Contains 110K images 22M questions
NLVR	https://lil.nlp.cornell.edu/nlvr/	• 107,292 examples of human-written English sentences grounded in pairs of photographs
COCO Caption	https://cocodataset.org/#download	• Image/Captioning: 123.2k/566.8k

Data Handling : Dataset Split

Dataset	No of images / (No of Question answer or captions)		
	Train	Validation	Test
VQA	113.2k/605.1k	5.0k/26.7k	5.0k/26.3k
GQA	72.1k/943.0k	10.2k/132.1k	12.6k/156.4k
NLVR	103.2k/86.4k	8.1k/7.0k	8.1k/7.0k
COCO Caption	113.2k/566.8k	5.0k/5.0k	5.0k/5.0k

Data Handling: NLVR2 Dataset

NLVR Pair Example

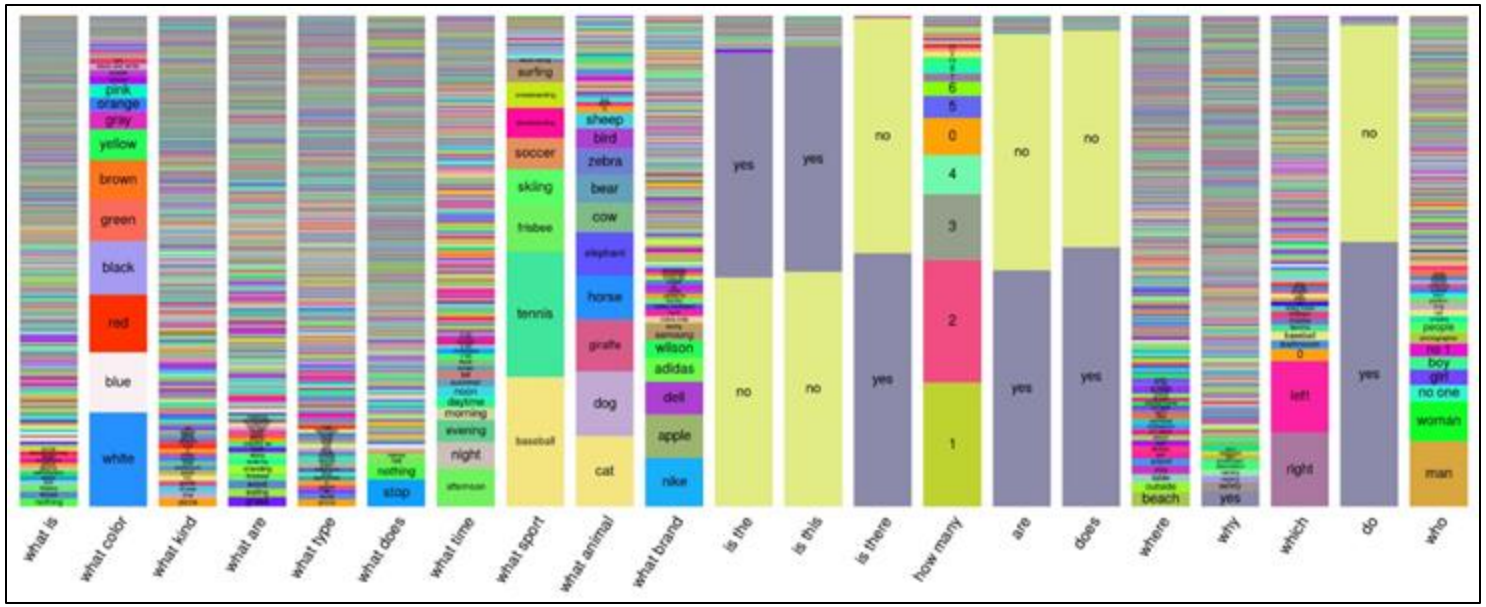
NLVR Pair and Related Question Example

Dataset Description

- NLVR2 contains 107,292 examples of human-written English sentences grounded in pairs of Photographs.
- The task is to determine whether a sentence is true about a visual input
- Solving the task requires reasoning about sets of objects, comparisons, and spatial relations, while including much more visually complex images.

Data Handling: VQA V2.0 Dataset

Distribution of the type of questions in VQA dataset



Example image and Related Question

Dataset Description

- Open-ended questions about images
- 204,721 COCO images
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question

Data Handling: GQA Dataset

Example image and Related Question

AA

AB

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**

Dataset Description

- Dataset contains Images and related compositional reasoning questions
- Questions are functional programs that represent scene semantics.
- Contains 110K images and 22M questions.
- Rectified flaw of previous dataset where, strong real-world priors displayed throughout the data

Data Handling: COCO Caption

Example image and Related Caption

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

Dataset Description

- **Number of Images: Over 120,000 images**
- **Number of Captions: Approximately 5 captions per image, totaling over 600,000 captions**
- **Content: Contains a wide variety of everyday scenes with diverse objects and activities**
- **Annotation: Each image is annotated with multiple captions describing different aspects of the scene**

Mathematical modelling of the problem

ADAPTOR

$$A^l(x) = U^l\left(GeLU\left(D^l(x)\right)\right) + x$$

$$\theta^D \in R^{\{d_i * d\}}$$

$$\theta^U \in R^{\{d * d_i\}}$$

$$Gaussian\ Error\ Linear\ Unit\ (GELU)$$

$$GELU(x) = x \cdot \Phi(x)$$

$$where\ \phi(x)\ is\ CDF\ of\ standard\ normal\ distribution$$

$$\text{GELU}(x) \approx 0.5x\left(1 + tanh\left[\frac{\sqrt{2}}{\pi}\left(x + 0.044715 \cdot x^3\right)\right]\right)$$

COMPACTER

$W_j = \sum_{i=1}^{n} A_i \otimes B_i^j = \sum_{i=1}^{n} A_i \otimes (S_i^j t_i^{j^T})$ Where n is hyperparameter

$A_i \in R^{n*n}$ and $B_i \in R^{\frac{k}{n}*\frac{d}{n}}$

B_i is low rank matrix $B_i = (S_i^j t_i^{j^T})$ Where $S_i \in R^{\frac{k}{n}*r}$ and $t_i \in R^{r*\frac{d}{n}}$ r is rank of matrix in general r=1

$A^l(x) = LPHM^{U^l}\left(GeLU\left(LPHM^{D^l}(x)\right)\right) + x$ LPHM = Low rank parameterized hypercomplex multiplication layers

HYPERFORMER

$$A_\tau^l(x) = LN_\tau^l\left(U_\tau^l\left(GeLU\left(D_\tau^l(x)\right)\right)\right) + x$$

Where LN_τ^l is conditional layer norm

$$LN_{\{\tau\}}^l\left(x_{\{\tau\}}^{\{i\}}\right) = \gamma_{\{\tau\}}^l \odot \frac{x_{\{i\}}^{\{\tau\}} - \mu_{\{\tau\}}}{\sigma_{\{\tau\}}} + \beta_{\{\tau\}}^l$$

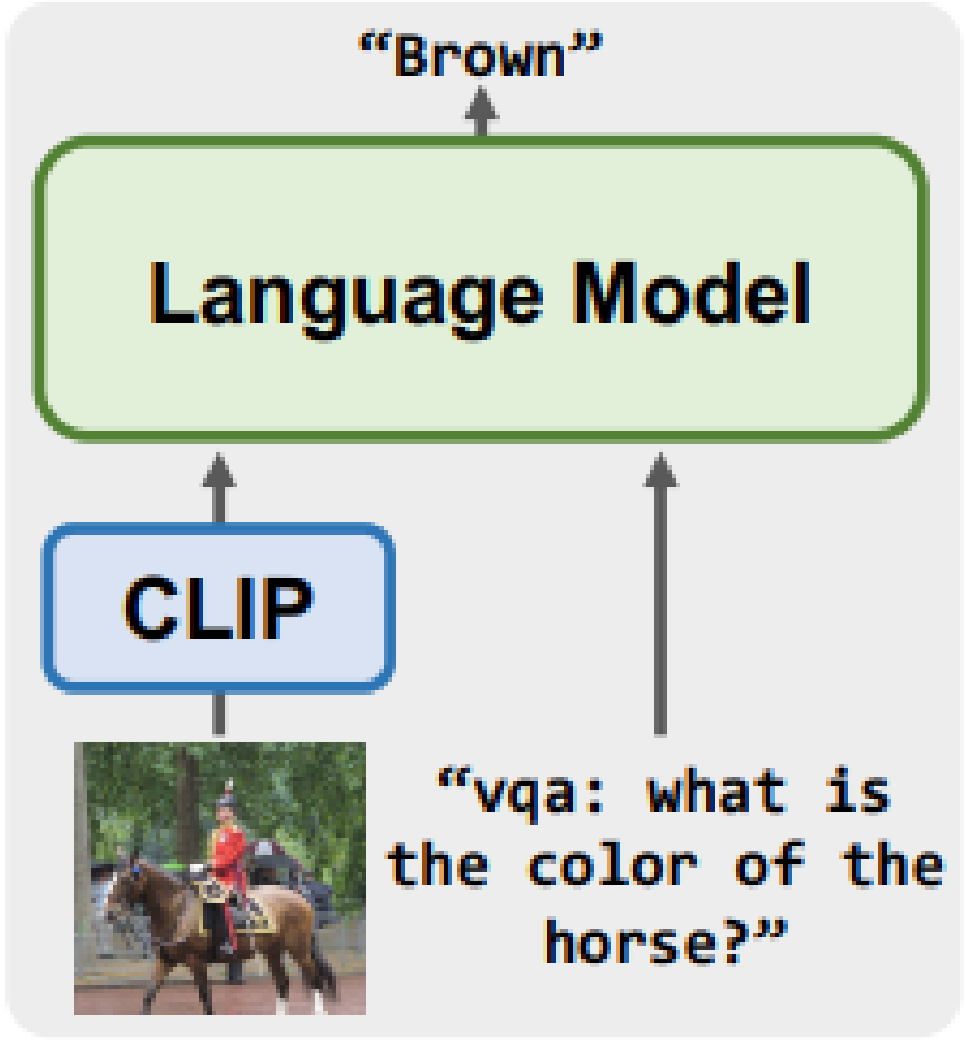
Where

$\gamma_{\{\tau\}}^l$ and $\beta_{\{\tau\}}^l$ are learnable parameter of same dimension $x_{\{\tau\}}^l$

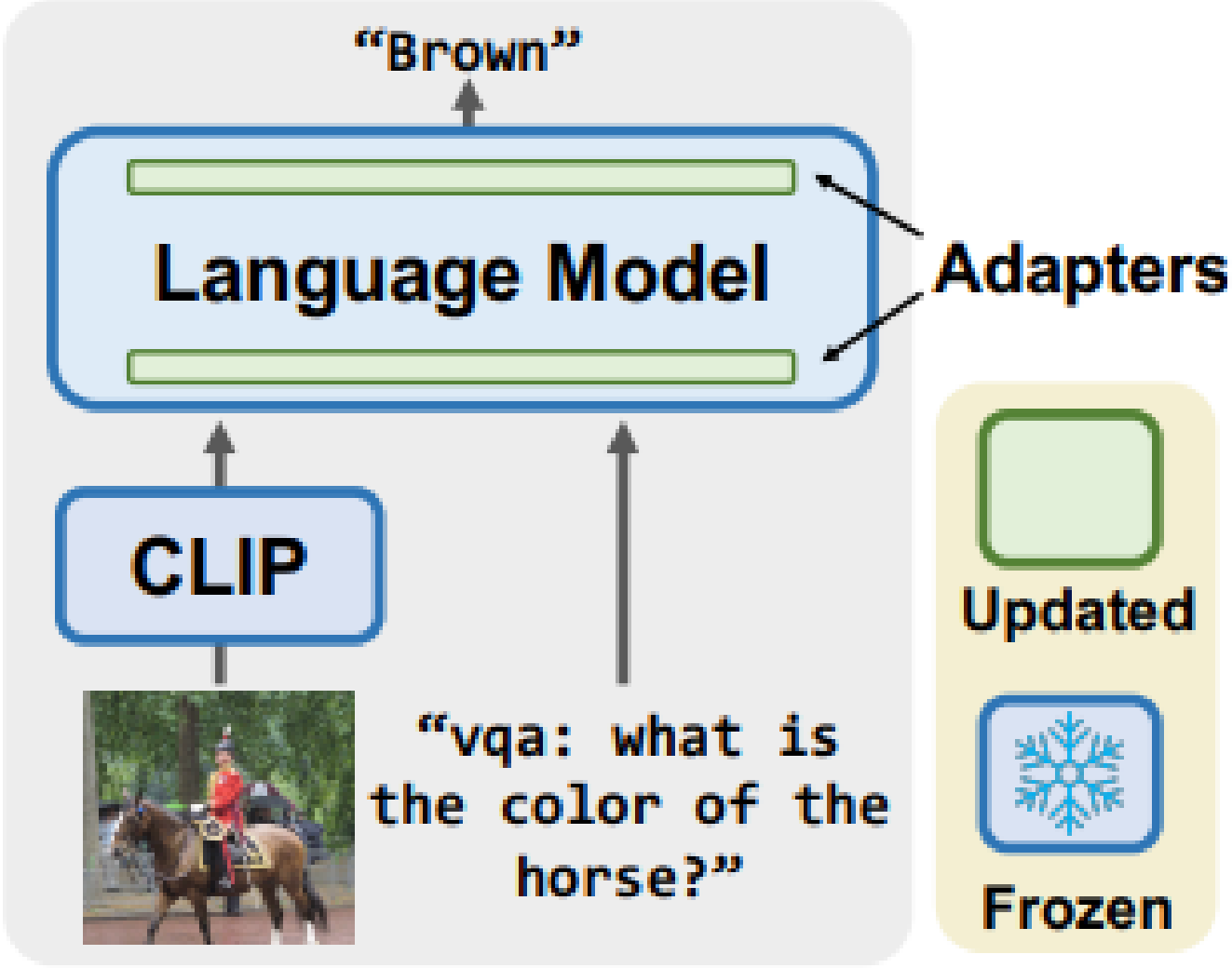
$\mu_{\{\tau\}}$ and $\sigma_{\{\tau\}}$ are mean and sd of $t^{\{th\}}$ task training data

Methodology: VL-Adapter Architecture

ORIGINAL ARCHITECTURE



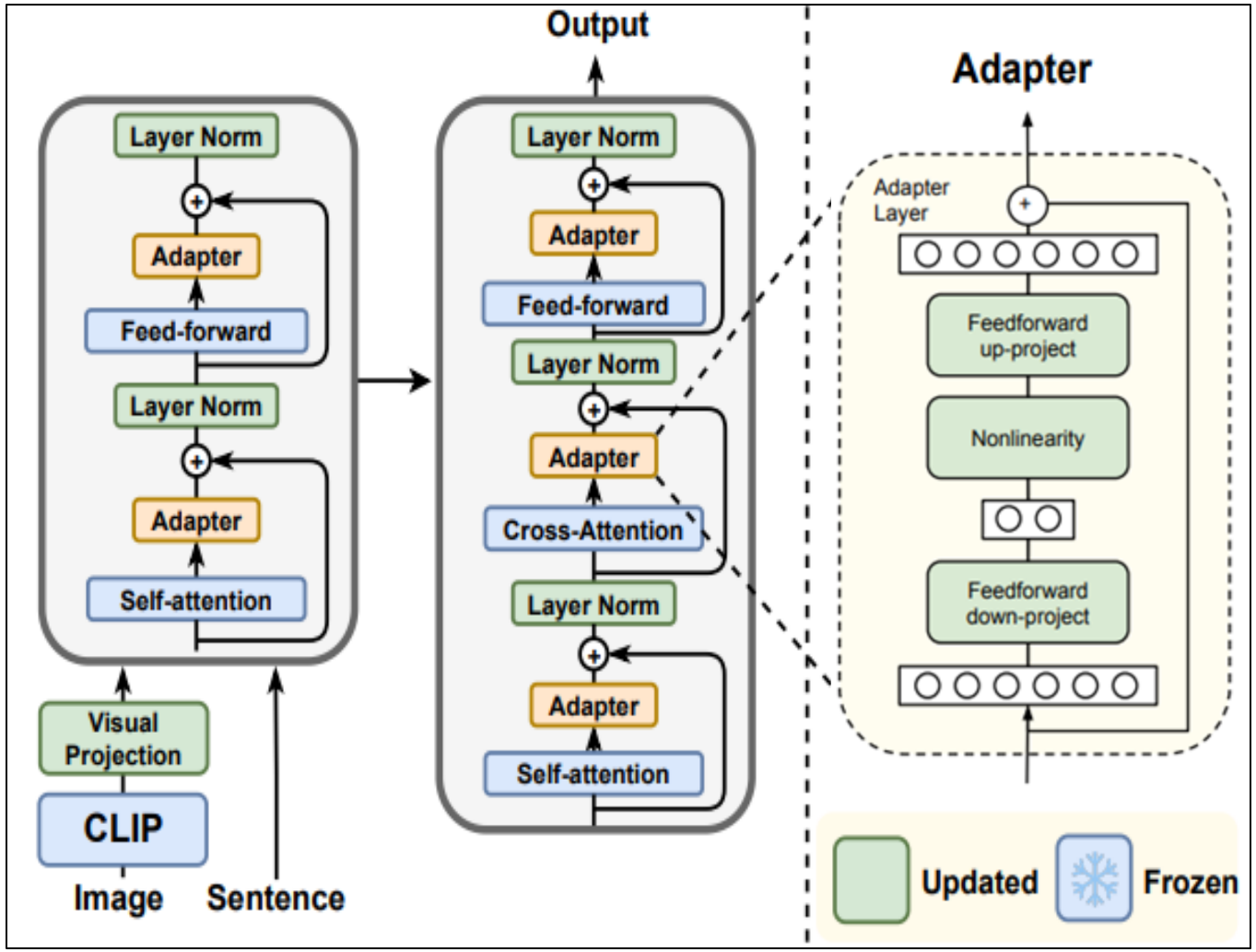
ADAPTER MODIFIED ARCHITECTURE



Source: Sung, Y. L., Cho, J., & Bansal, M. (2022). Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5227-5237)

Methodology: Adapter Module

ADAPTER MODULE ARCHITECTURE



ABOUT

1. Project the original d-dimensional features into a smaller dimension, m.
2. Apply a nonlinearity to the reduced features.
3. Project the features back to the original d dimensions.
4. Add skip connection to the output of step 3.

MATHEMATICAL FORMULATION

$$A^l(x) = U^l\left(GeLU\left(D^l(x)\right)\right) + x$$

$$\theta^D \in R^{\{d_i * d\}}$$

$$\theta^U \in R^{\{d * d_i\}}$$

$$Gaussian\ Error\ Linear\ Unit\ (GELU)$$

$$GELU(x) = x \cdot \Phi(x)$$

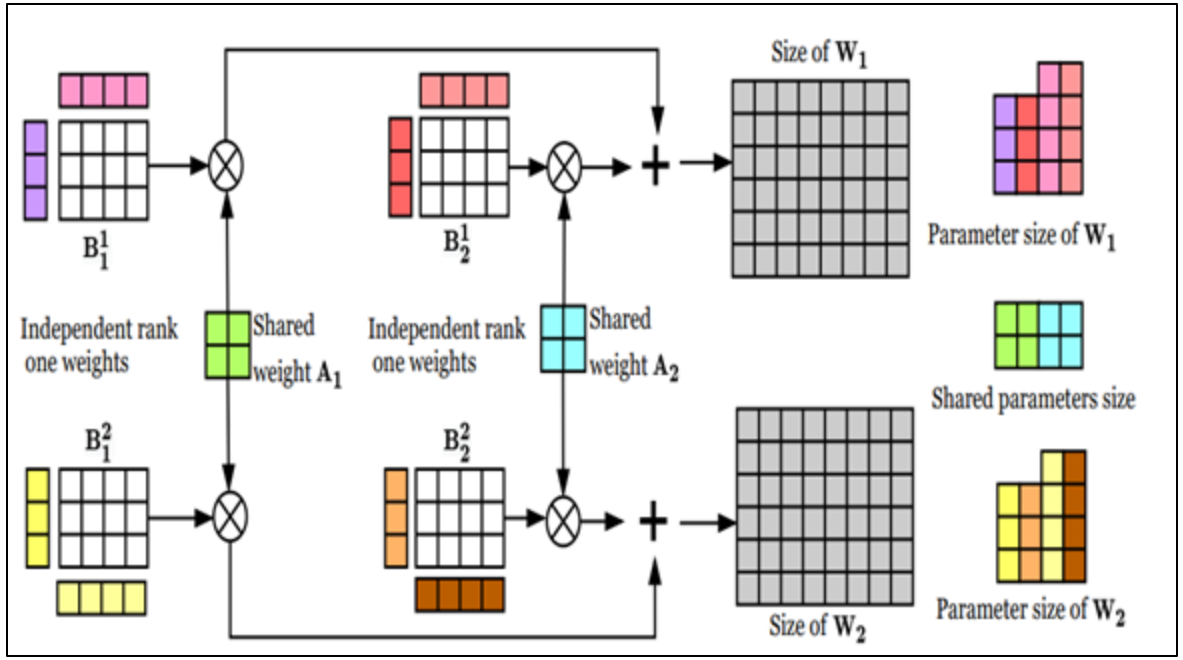
$$where\ \phi(x)\ is\ CDF\ of\ standard\ normal\ distribution$$

$$\text{GELU}(x) \approx 0.5x\left(1 + tanh\left[\frac{\sqrt{2}}{\pi}\left(x + 0.044715 \cdot x^3\right)\right]\right)$$

Image Source: Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning

Methodology: Compacter Module

COMPACTER ARCHITECTURE



MATHEMATICAL FORMULATION

$W_j = \sum_{i=1}^{n} A_i \otimes B_i^j = \sum_{i=1}^{n} A_i \otimes (S_i^j t_i^{jT})$ Where n is hyperparameter $A_i \in R^{n*n}$ and $B_i \in R^{\frac{k}{n}*\frac{d}{n}}$

B_i is low $rank$ $matrix$ $B_i = (S_i^j t_i^{jT})$ Where $S_i \in R^{\frac{k}{n}*r}$ and $t_i \in R^{r*\frac{d}{n}}$ r is rank of matrix in general r=1

$A^l(x) = LPHM^{U^l}\left(GeLU\left(LPHM^{D^l}(x)\right)\right) + x$ LPHM = Low rank parameterized hypercomplex multiplication layers

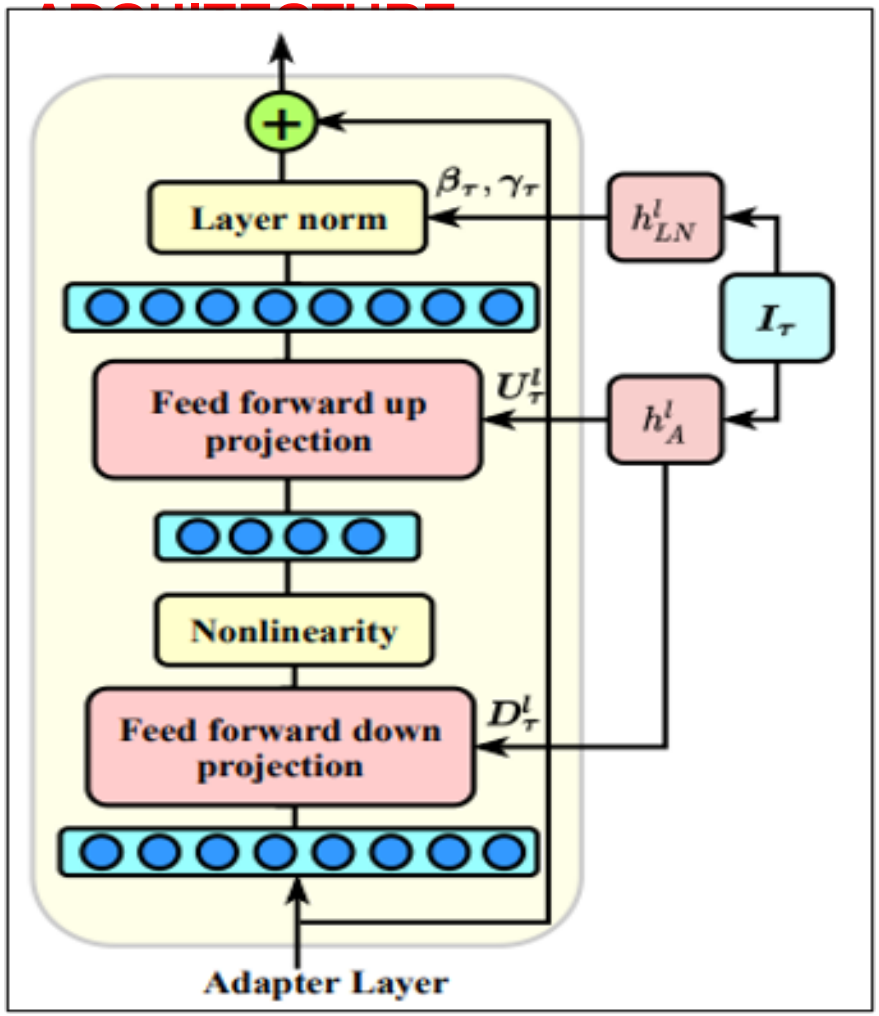
ABOUT

1. Each COMPACTER weight matrix is computed as the sum of Kronecker products between shared "slow" weights and "fast" rank-one matrices defined per COMPACTER layer
2. COMPACTER reduces parameter complexity to O(k+d) from the O(kd) complexity typically found in regular adapters

Image Source: Rabeeh Karimi mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In Thirty-Fifth Conference on Neural Information Processing Systems, 2021.

Methodology: Hyperformer Module

HYPERFORMER



ABOUT

- The Adapter hypernetwork (h^l_A) generating task-specific adapter module weights (U^l_T and D^l_T)
- Conditional generation of adapter parameters based on input task embedding (I^T) Layer
- Normalization hypernetwork (h^l_{LN}) producing conditional layer normalization parameters ($\beta_T$and γ_T)

MATHEMATICAL FORMULATION

$$A^l_\tau(x) = LN^l_\tau\left(U^l_\tau\left(GeLU\left(D^l_\tau(x)\right)\right)\right) + x$$

Where LN^l_τ is conditional layer norm

$$LN^l_{\{\tau\}}\left(x^{\{i\}}_{\{\tau\}}\right) = \gamma^l_{\{\tau\}} \odot \frac{x^{\{\tau\}}_{\{i\}} - \mu_{\{\tau\}}}{\sigma_{\{\tau\}}} + \beta^l_{\{\tau\}}$$

Where

$\gamma^l_{\{\tau\}}$ *and* $\beta^l_{\{\tau\}}$ *are learnable parameter of same dimension* $x^l_{\{\tau\}}$

$\mu_{\{\tau\}}$ *and* $\sigma_{\{\tau\}}$ *are mean and sd of* $t^{\{th\}}$ *task training data*

Image Source: Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In Annual Meeting of the Association for Computational Linguistics, 2021

Experimental details

Model selection

VL-ADAPTER FULL FINE TUNING, ADAPTER MODULE, HYPERFORMER MODULE, COMPACTER MODULE, LOw Rank Adaptation of LLMs, Parameter-Efficient Prompt Tuning

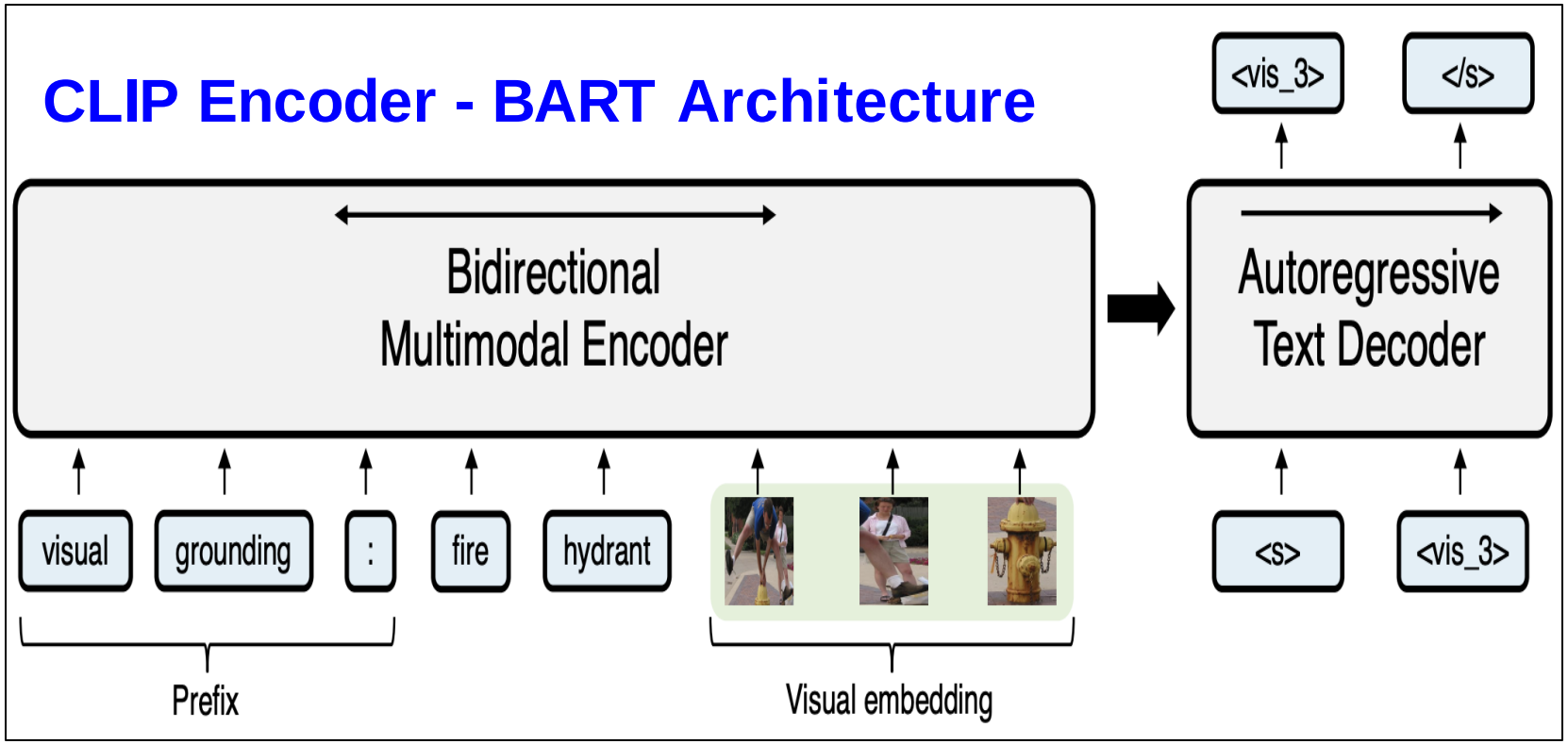
Hyperparameters

Parameter	Value	Parameter	Value	Parameter	Value
adam_beta1	0.9	adam_beta2	0.999	adam_eps	1e-06
batch_size	256	dropout	0.1	epochs	12
visual_feat_dim	2048	image_size	(448,448)	lr	0.0001
max_n_boxes	36	max_text_length	20	mid_dim	768
n_boxes	36	weight_decay	0.01	optimizer	adamw

Metrics

Bleu1, Bleu2, Bleu3, Bleu4, CIDEr, METEOR, SPICE

Baseline



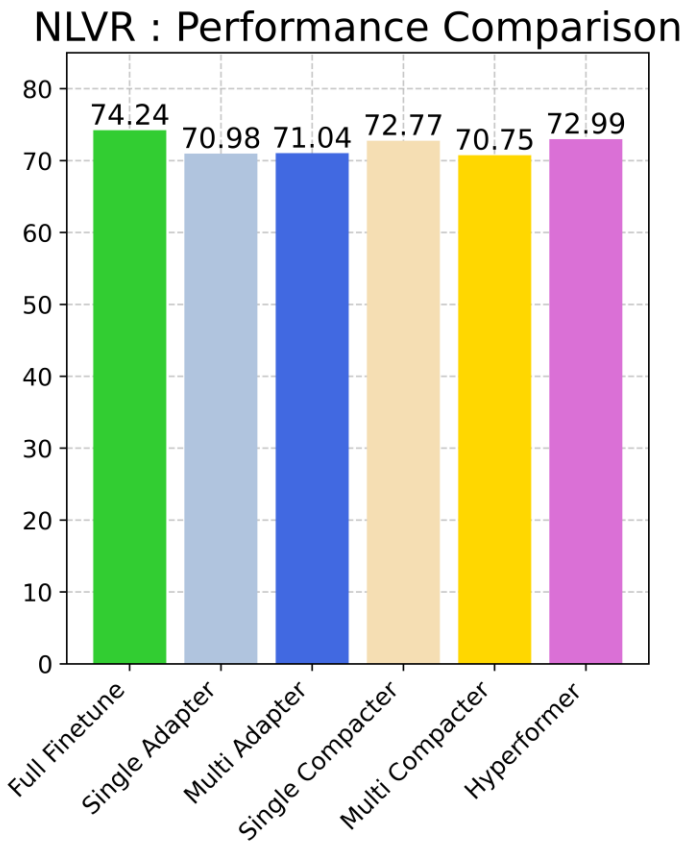
- Baseline: Full Finetuning
- Model: CLIP Encoder-BART
- For four Tasks: NLVR2, VQA, GQA, COCO Captioning
- Base Results are in Right side

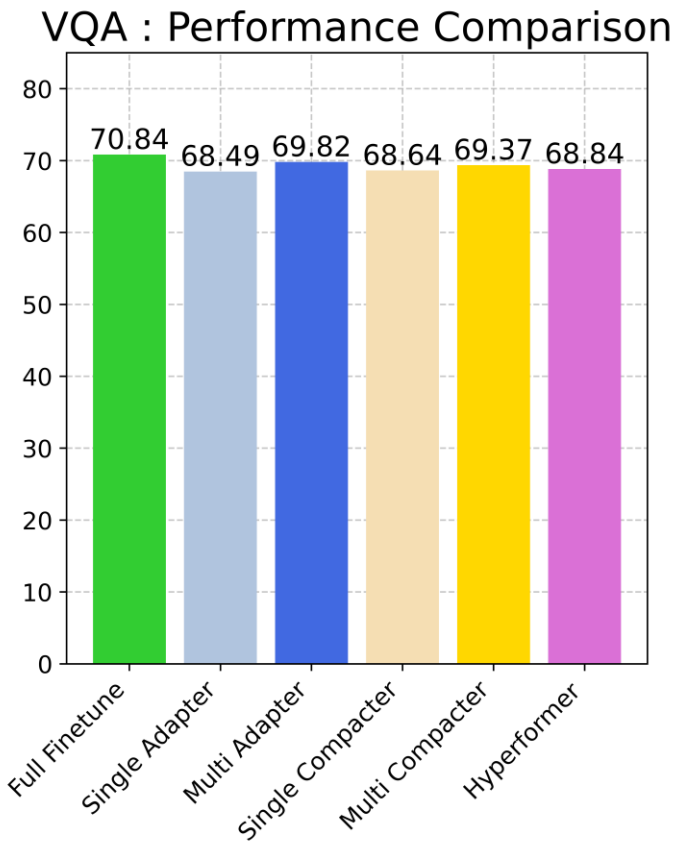
Task	Accuracy (CLIP Encoder - BART LM)
NLVR2	74.236
VQA	70.842
GQA	55.569
Bleu1	73.05
Bleu2	56.29
Bleu3	44.982
Bleu4	32.69
CIDEr	108.8
METEOR	28.17
SPICE	21.42

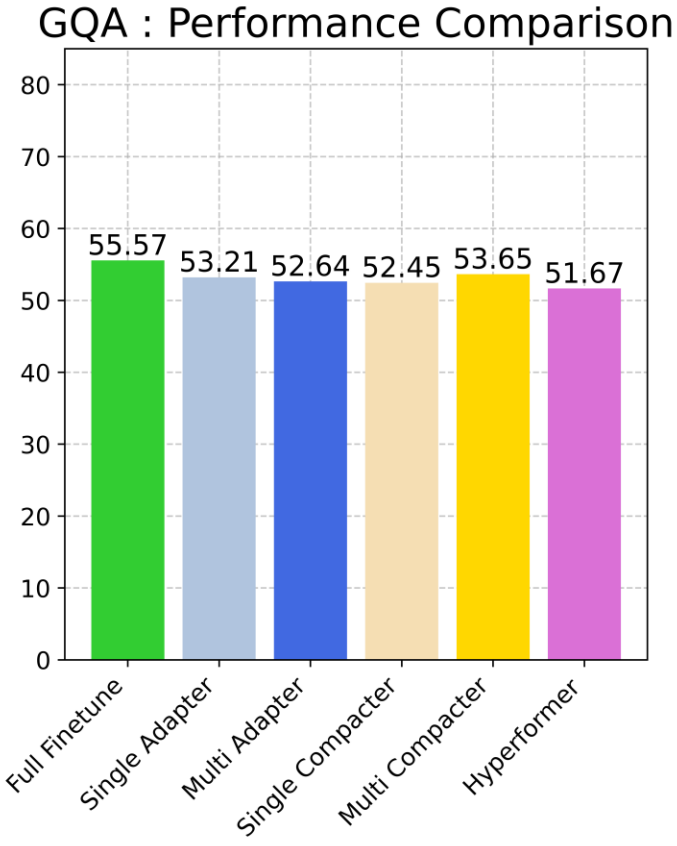
Source: Cho, Jaemin, Jie Lei, Hao Tan, and Mohit Bansal. "Unifying vision-and-language tasks via text generation." In *International Conference on Machine Learning*, pp. 1931-1942. PMLR, 2021.

Results and Analysis : Accuracy

Results of Experiments using three different Adapters







BASELINE : Full Finetuning

Results and Analysis : Parameter Reduction

Results of Experiments using three different Adapters

- Number of Parameters (Base model : full finetuning): 140M
- CLIP Parameter are not included as the parameters are frozen for all cases

Method	Updated Parameters (%)	VQA Accuracy (%)	GQA Accuracy (%)	NLVR2 Accuracy (%)
Full Fine-tuning	100.00	70.84	55.57	74.24
Single Adapter	4.18	68.49	53.21	70.98
Multiple Adapters	12.22	69.82	52.64	71.04
Single Compacter	2.70	68.64	52.45	72.77
Multiple Compacters	7.05	69.37	53.65	70.75
Hyperformer	5.79	68.84	51.67	72.99

Case study: NLVR

Example where model is **performing well**

write question related to image

in left image, people are doing meeting but in right image people are celebrating their birthday

Answer

True

Case study: NLVR

Example where model is **failing**

write question related to image

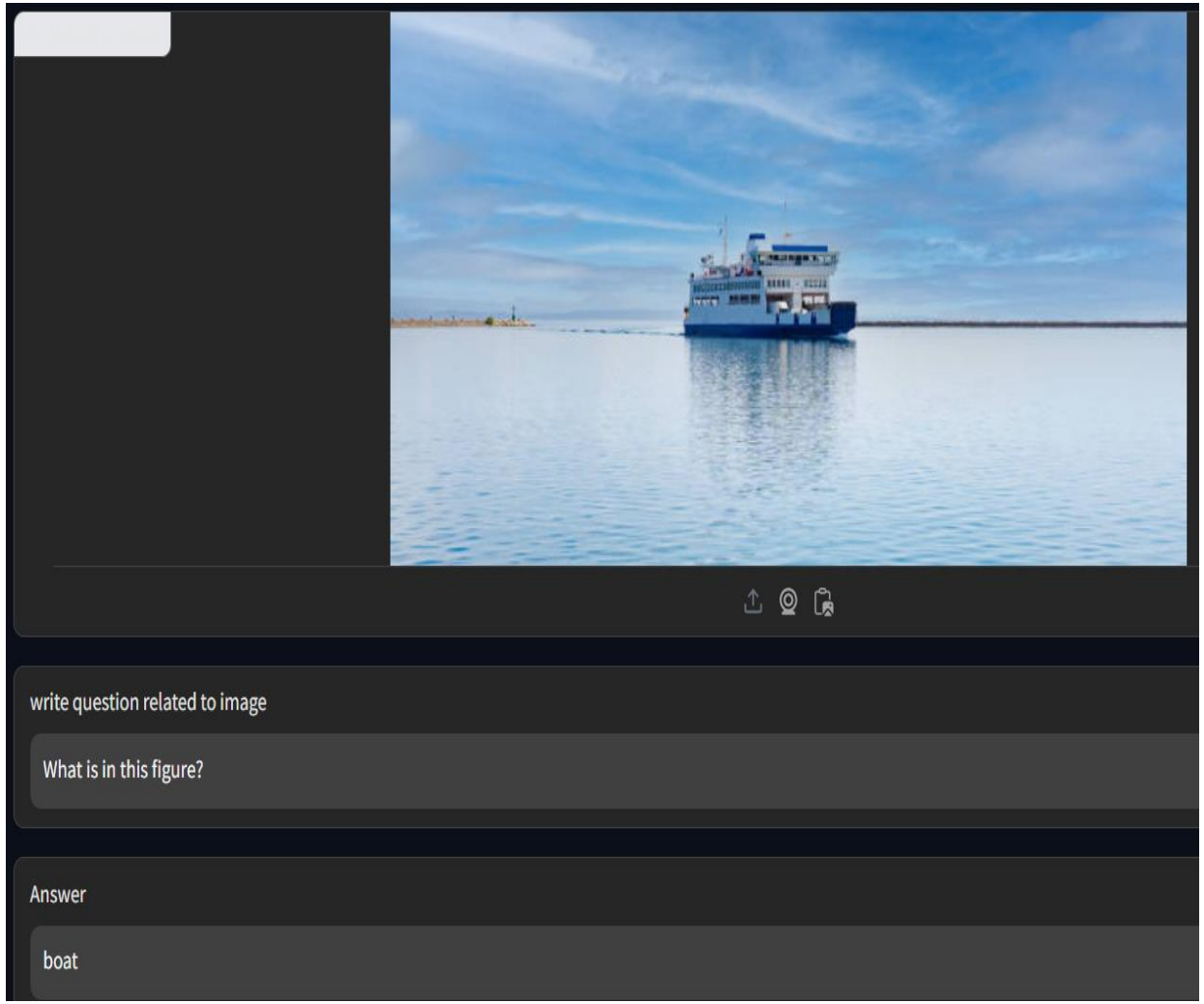
The number of people is greater in the right image compared to the left image.

Answer

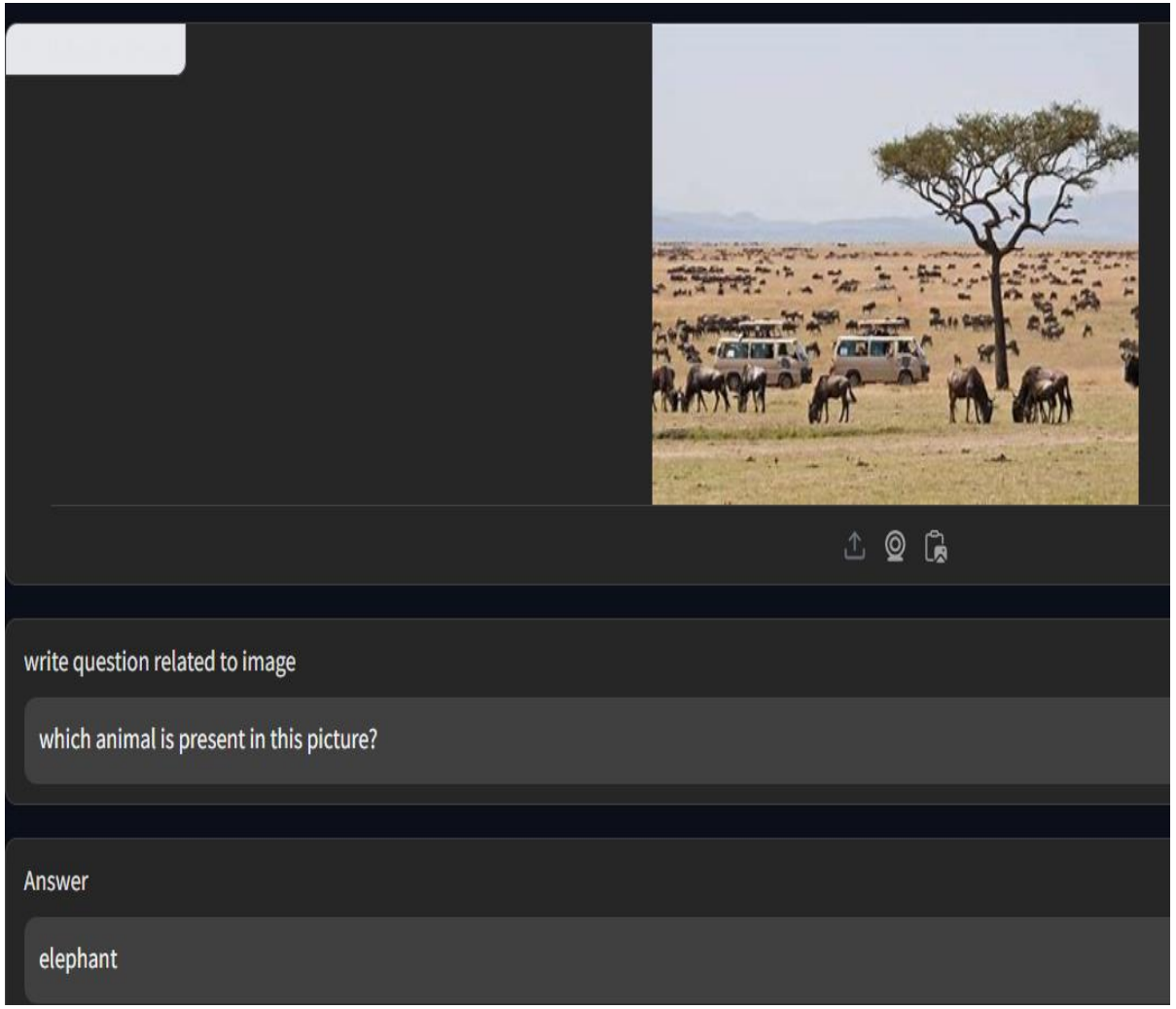
False

Case study: VQA

Example where model is performing well

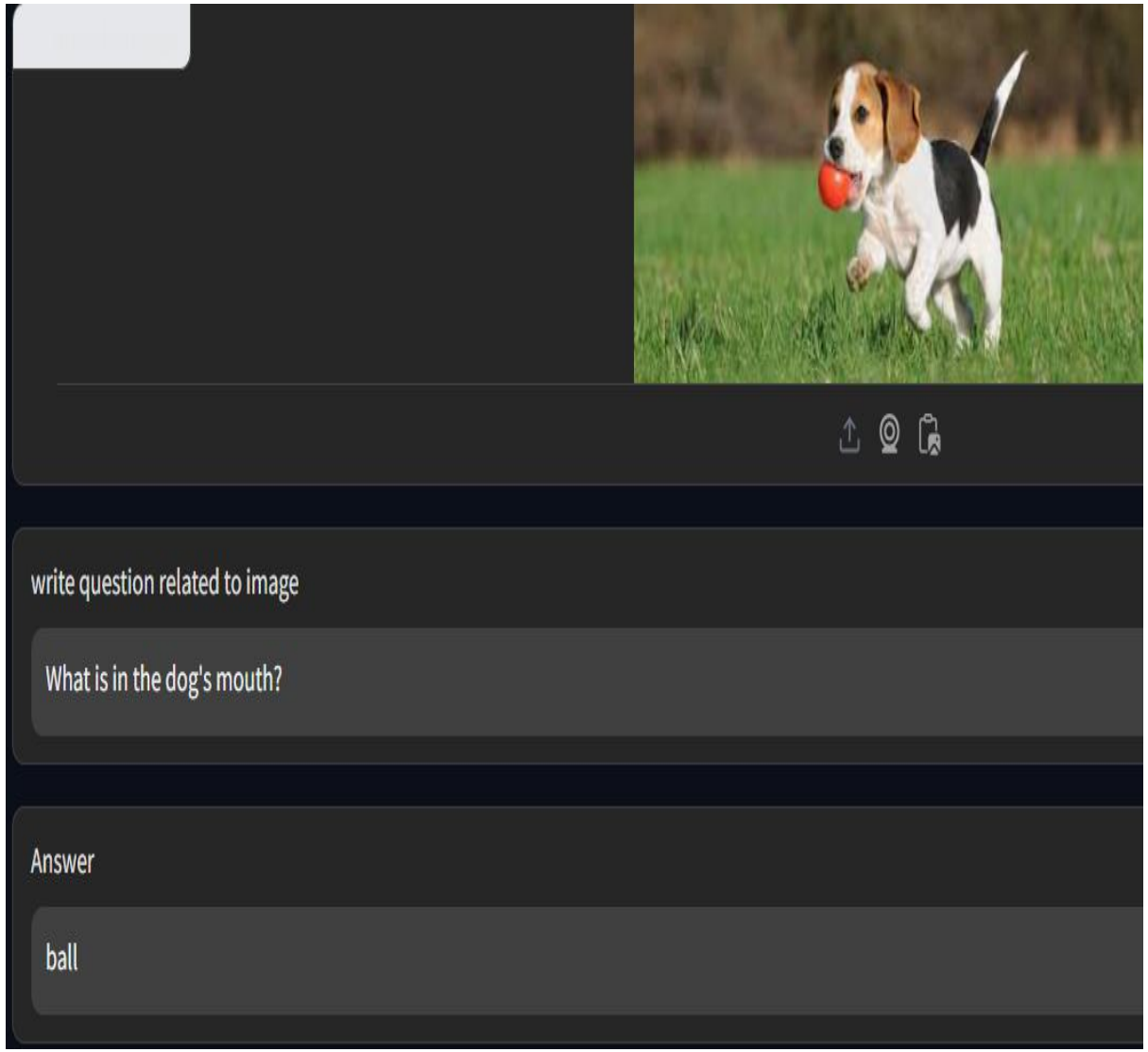


Example where model is failing

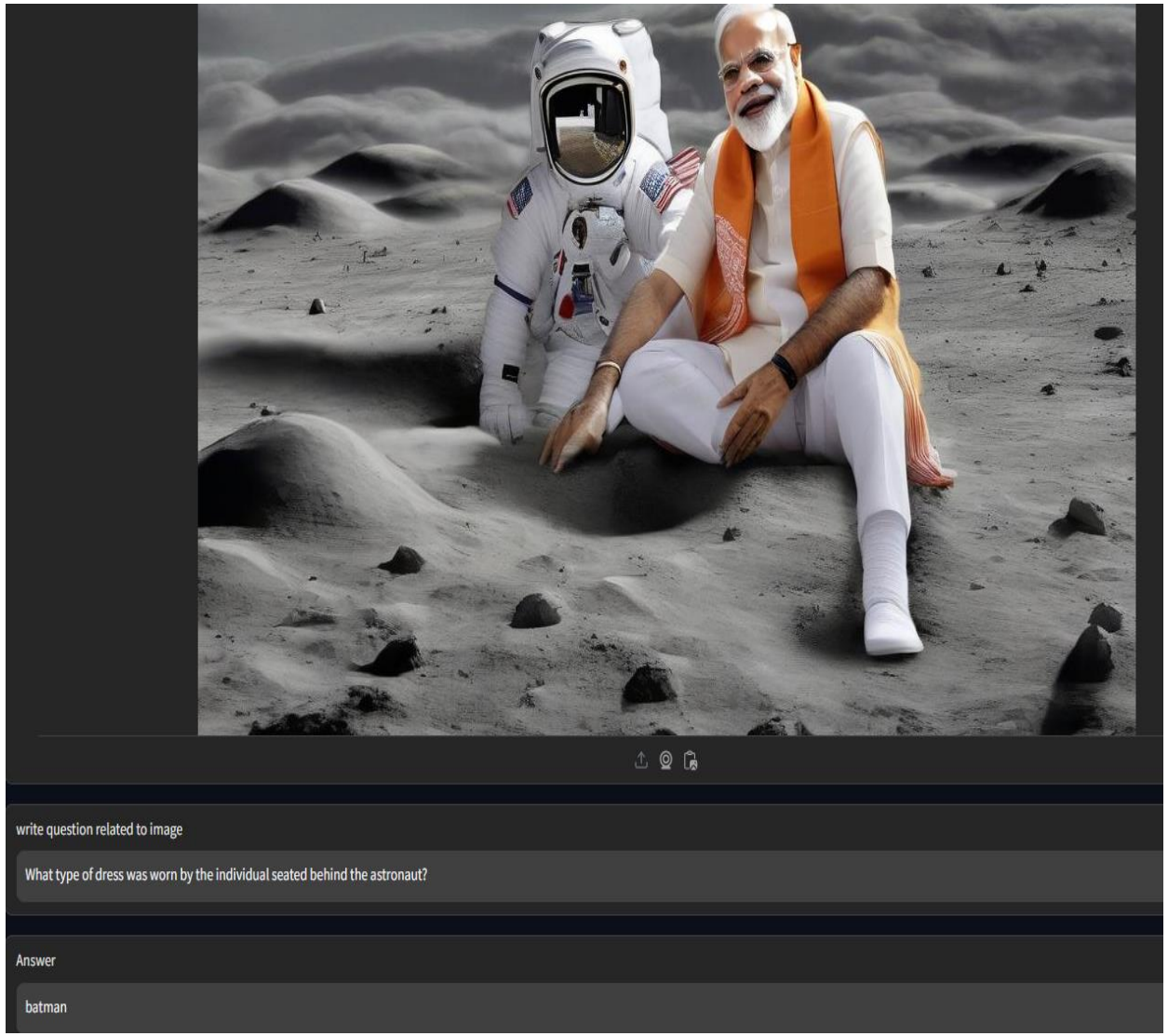


Case study: GQA

Example where model is performing well

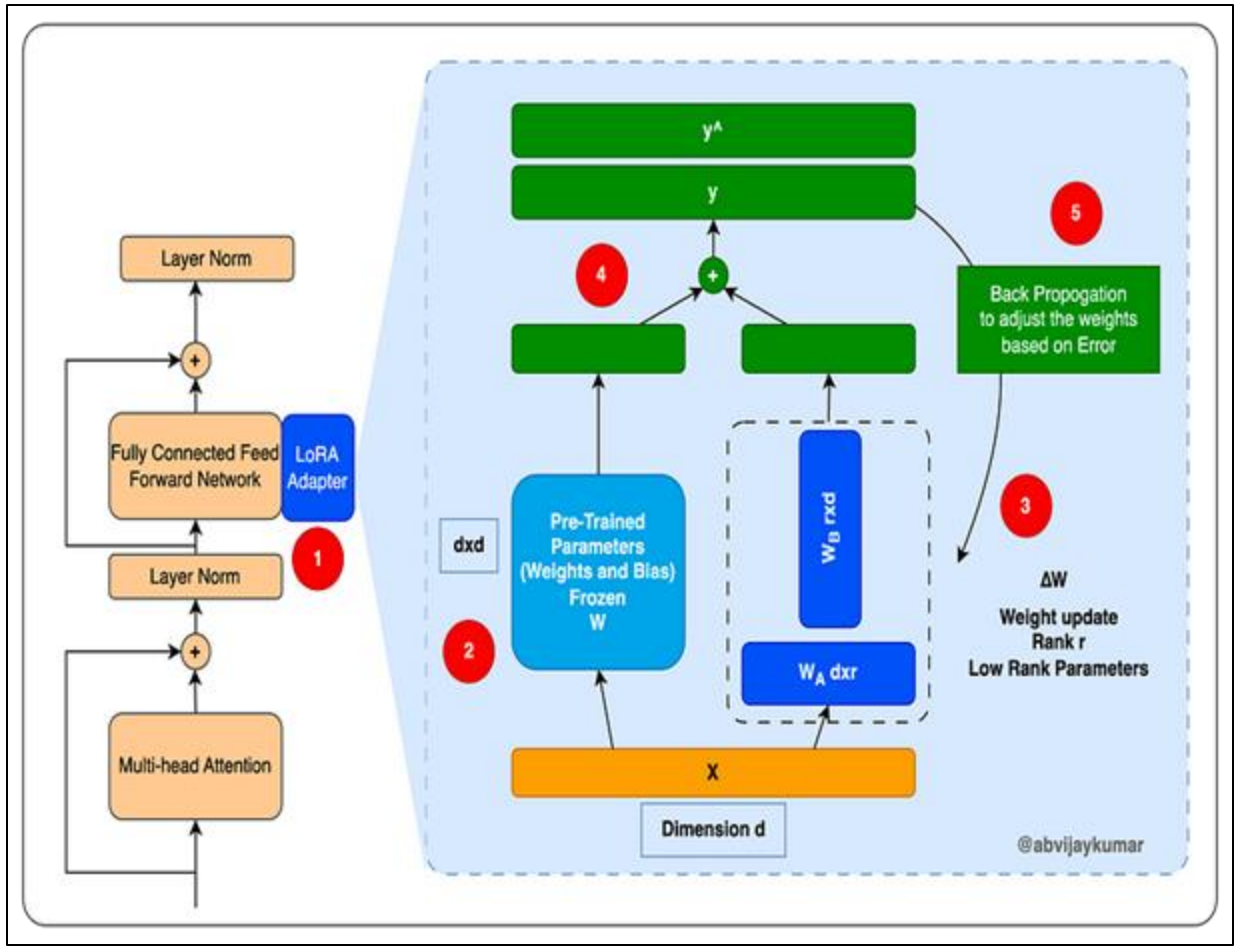


Example where model is failing



Methodology: LOw Rank Adaptation of LLMs

LORA ARCHITECTURE



ABOUT

1. For a pre-trained weight matrix $W_0 \in R^{d \times k}$, we constrain its update by representing the latter with a low-rank decomposition $W_0 + \Delta W = W_0 + BA$, where $B \in R^{d \times r}$, $A \in R^{r \times k}$, and the rank $r \leq min(d, k)$
2. During training, W_0 is frozen and does not receive gradient updates, while A and B contain trainable parameters.

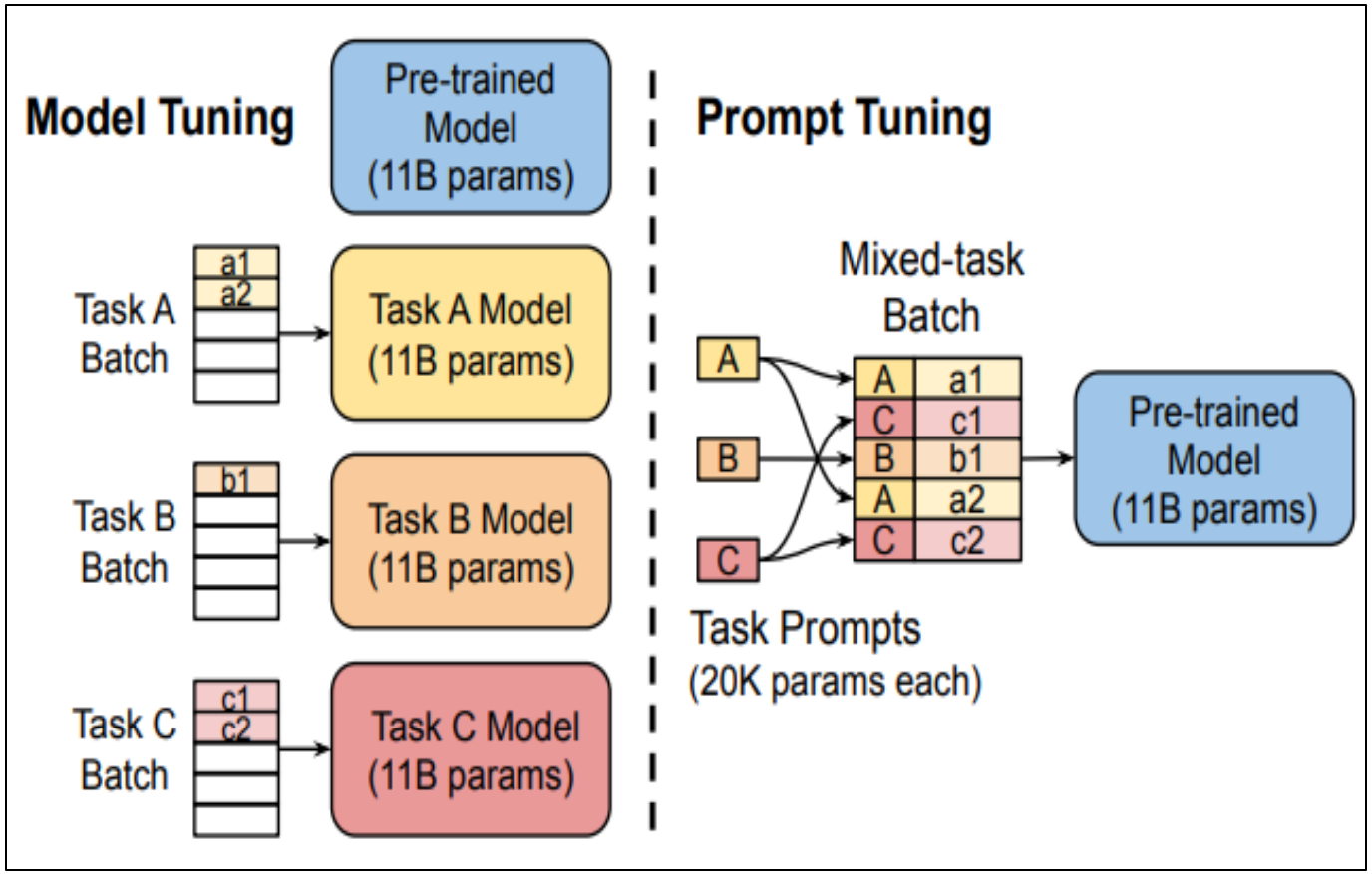
IMAGE SOURCE: https://abvijaykumar.medium.com/fine-tuning-llm-parameter-efficient-fine-tuning-peft-lora-qlora-part-2-d8e23877ac6f

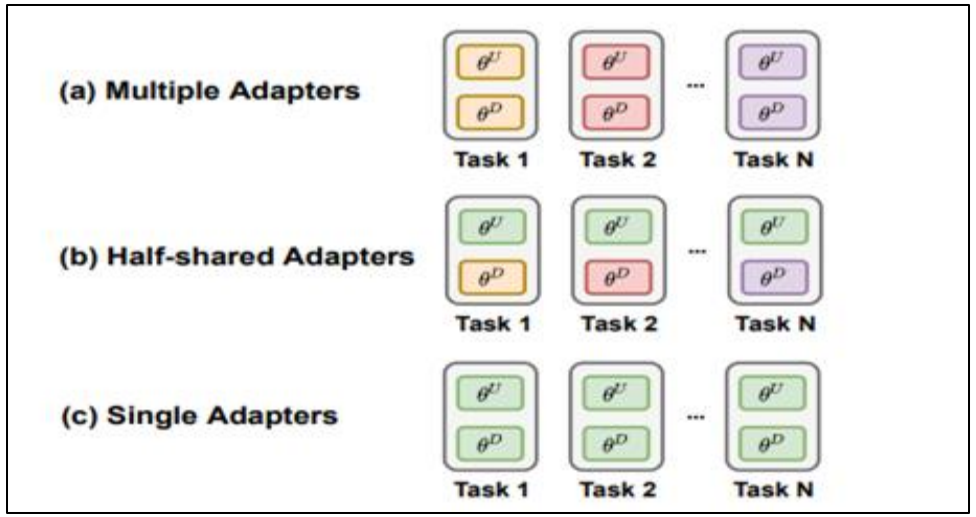
LORA PAPER : Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan AllenZhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models.

Methodology: Parameter-Efficient Prompt Tuning

ARCHITECTURE

20,480 parameters per task





ABOUT

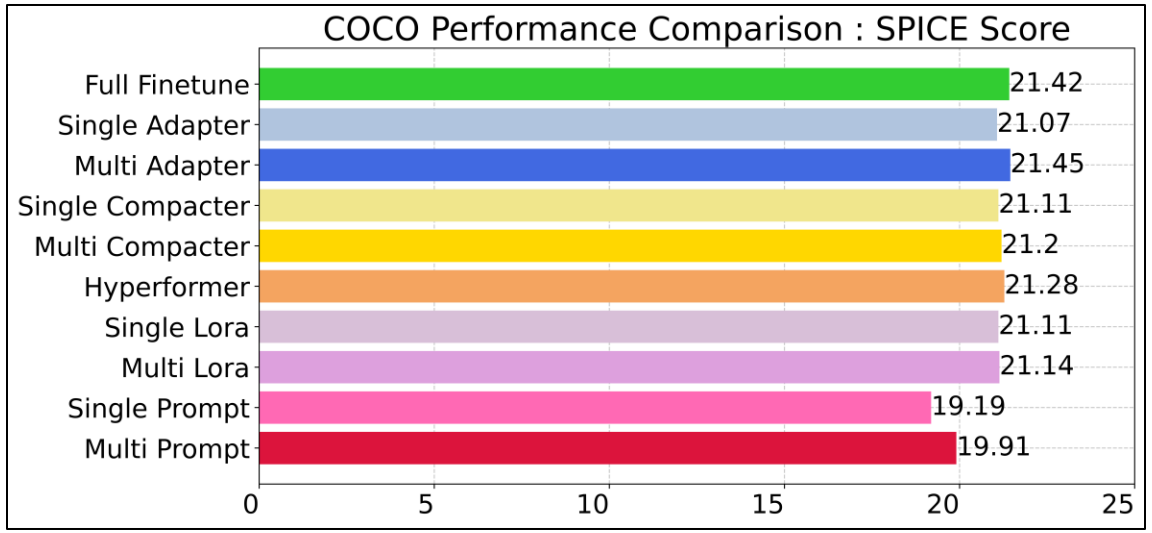
1. **Utilize prompt tuning as a simplification strategy for adapting language models by freezing the pre-trained model and introducing a "soft prompt" where only a specified number, k, of tunable tokens are added to the input text.**

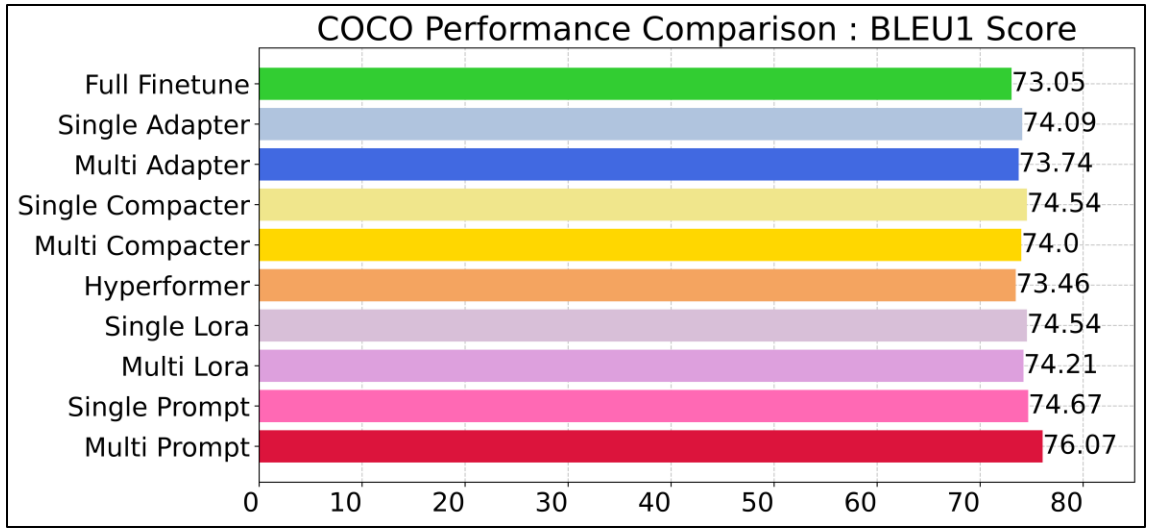
Source:

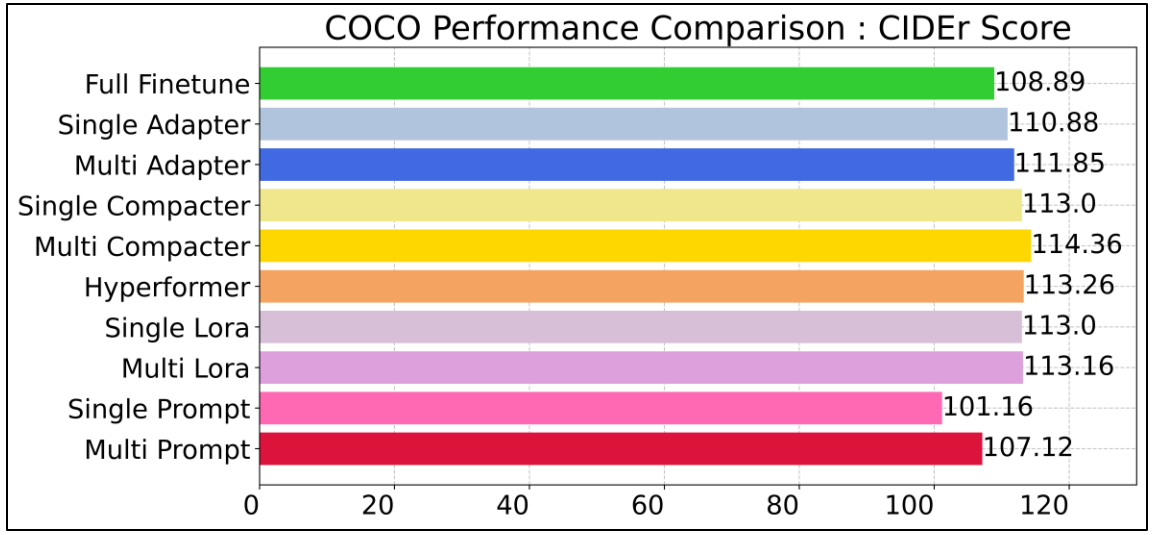
- Sung, Y. L., Cho, J., & Bansal, M. (2022). Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5227-5237)
- Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In EMNLP, 2021

Results and Analysis : Beating COCO

Comparision of Adapters to Other parameter reduction methods



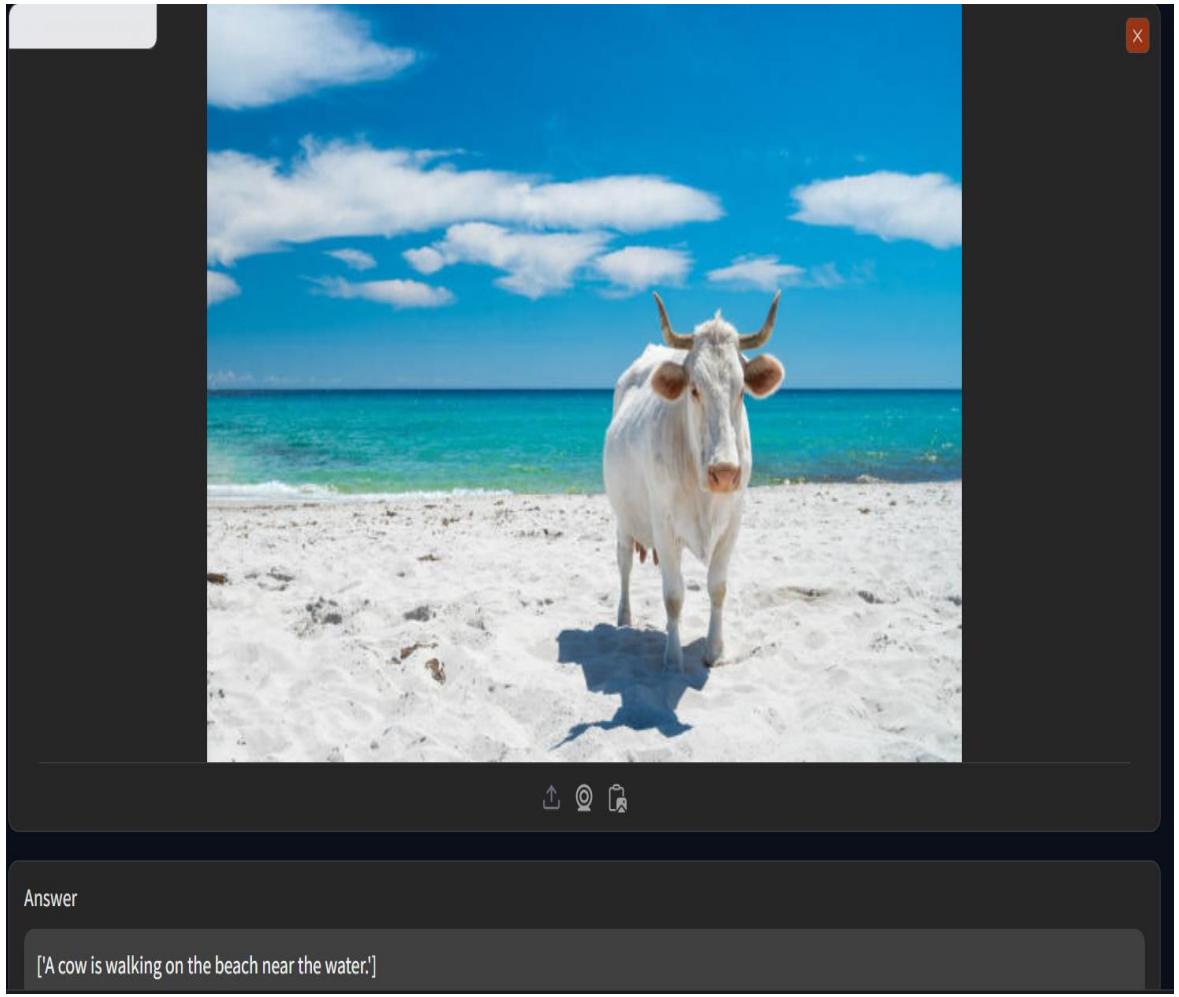




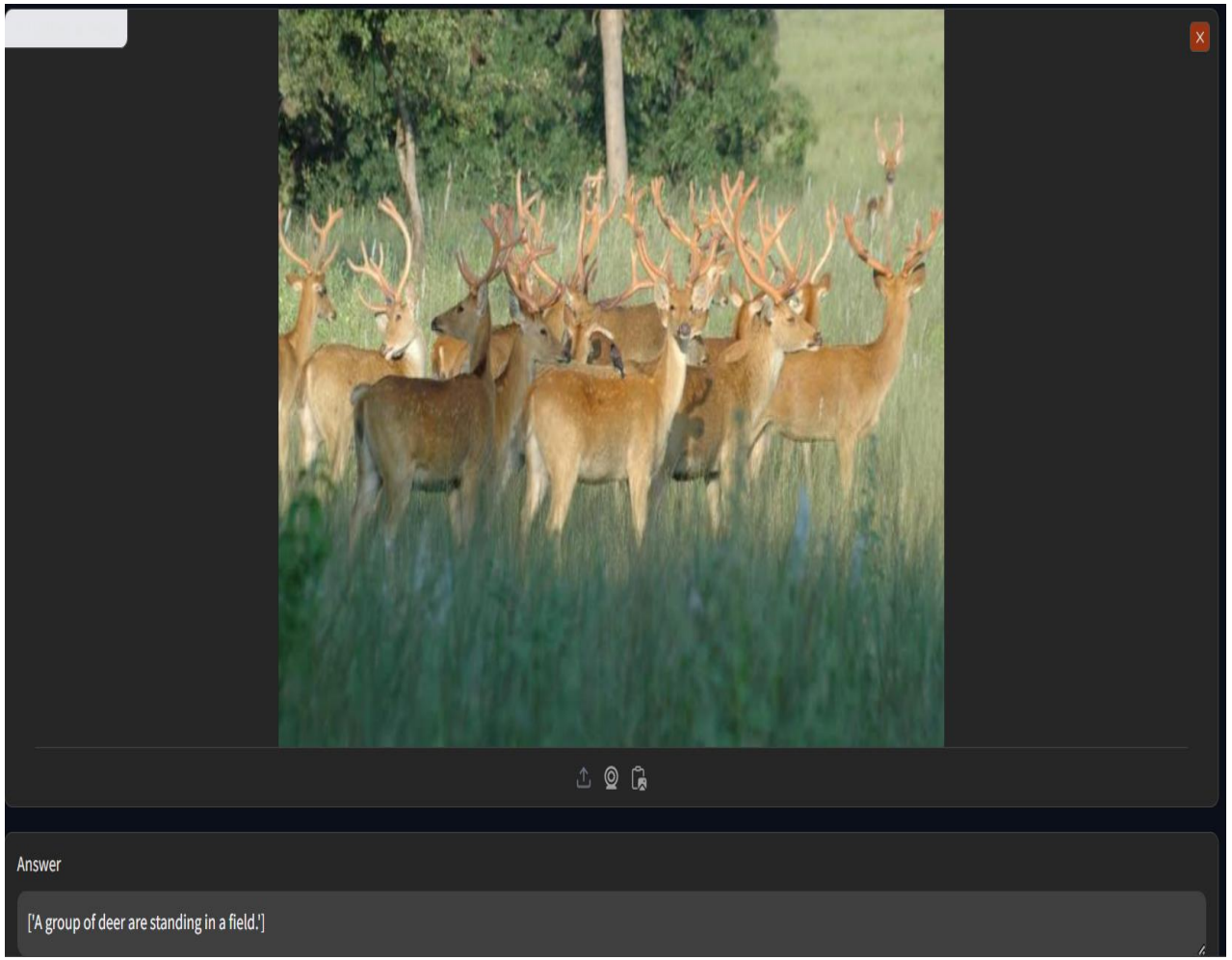
- Beat COCO captioning for full finetuning
- SPICE Score: Multi Adapter perform better than Full fine tuning, Best : Multi Adapter
- BLEU1 Score: All cases perform better than Full Finetuning, Best: Multi Prompt
- CIDEr Score: Except single and Multi Prompt, ALL beat full finetuning, Best: Multi Compacter

Case study: Captioning

Example where model is performing well



Example where model is failing



Results and Analysis : Comparative Study

Comparision of Adapters to Other parameter reduction methods

	NLVR	VQA	GQA	Bleu 1	Bleu 2	Bleu 3	Bleu 4	CIDEr	METEOR	SPICE
Full finetune	74.236	70.84	55.569	73.05	56.29	44.982	32.69	108.8	28.17	21.42
Single Adapter	70.977	68.49	53.212	74.09	57.67	44.852	33.91	110.8	28.18	21.07
Multi Adapter	71.035	69.82	52.64	73.74	57.37	43.973	33.71	111.8	28.57	21.45
Hyperformer	72.987	68.84	51.67	73.46	57.2	43.966	33.9	113.2	28.51	21.28
Single Compacter	72.772	68.64	52.4525	74.54	58.37	44.952	34.64	112.9	28.24	21.11
Multi Compacter	70.748	69.37	53.654	74	57.85	44.475	34.17	114.3	28.21	21.2
Single Lora	73.374	69.27	53.7285	74.5	58.24	44.854	34.61	113.5	28.4	21.25
Multi Lora	51.069	68.28	54.135	74.21	57.85	44.42	34.17	113.1	28.32	21.14
Single Prompt	51.801	46.89	37.1526	74.67	58.38	43.68	31.79	101.1	25.68	19.19
Multi Prompt	51.916	48.96	38.65	76.07	59.66	44.7709	32.8	107.1	26.21	19.91

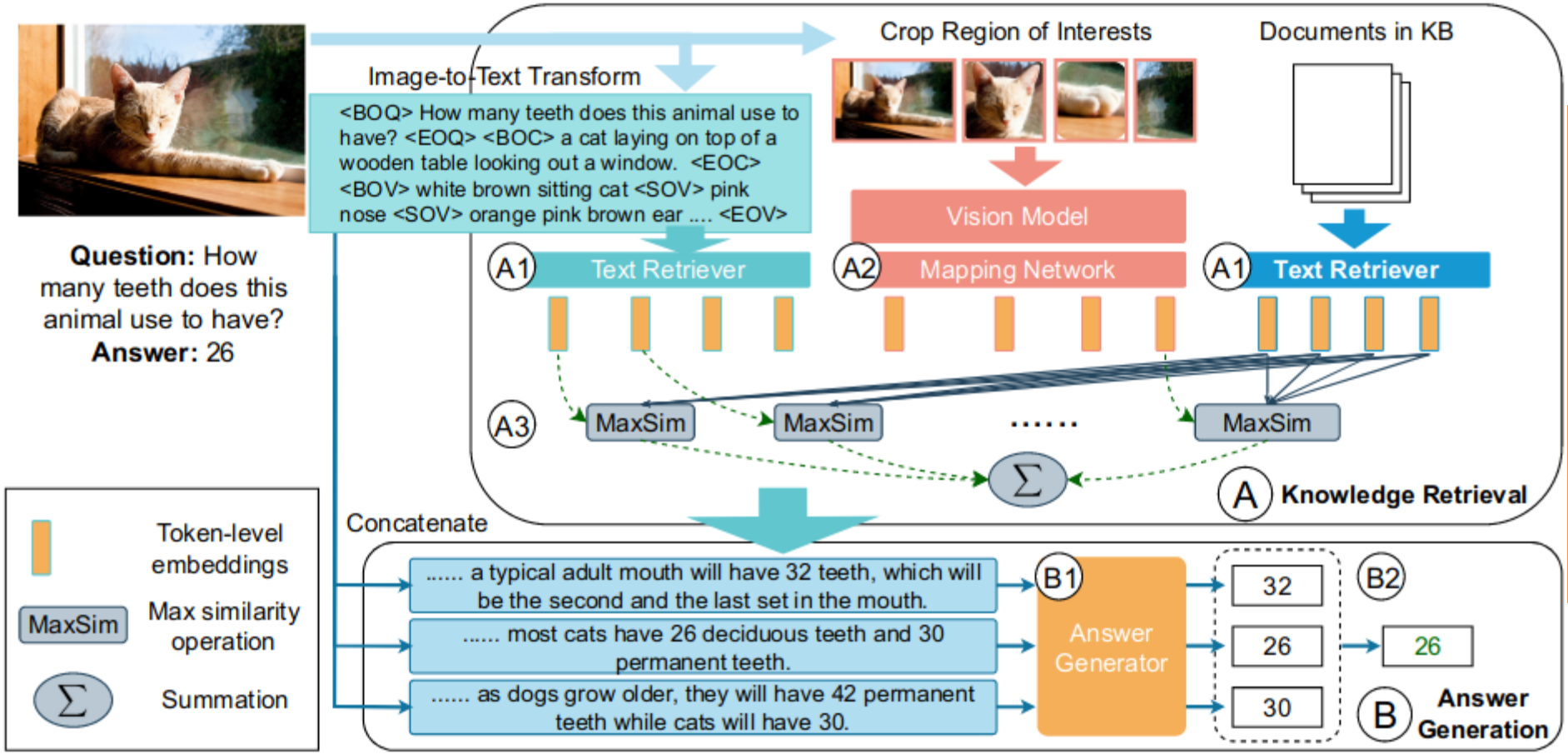
Adapters beat EVERYTHING even FULL FINETUNING

Adapters beat other parameter reduction methods [LORA & PROMPT TUNING]

LORA OR PROMPT TUNING beat Adapter

Methodology: Open Domain VQA

Retrieval based VQA



For each query token, the MAX operation selects the highest relevance score over all document tokens.

$$r((q, I), d) = \sum_{i=1}^{l_Q} \max_{j=1}^{l_D} Q_i D_j^T$$

A: Retrieval Module (A1: Text Models, A2: Vision Models, A3: Relevance score b/w query and document)

B: Answer Generation (language model or another smaller vision-text model or RAG generation)

Source: Lin, W., Chen, J., Mei, J., Coca, A., & Byrne, B. (2024). Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, *36*.

Results and Analysis: Retriever based VQA

- Retriever model was initialized with pretrained FLMR weights.
- A Vector database of over 168K Wikipedia passages was created using the document encoder or retriever.
- Moving information between GPU and CPU was bottleneck as vector database was in CPU.
- Retriever was evaluated on subset of data taken from test data of OK-VQA dataset.

Number of Samples	Recall @ 3
500	0.79025

Recall @ k

- It defines probability that a relevant document is retrieved by the query in first k documents.

$$\frac{\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}}{\{\text{relevant documents}\}}$$

Qualitative Results: Retriever based VQA

- it's an expensive sport, and whether you're in it for fun or business, you must spend abundantly. if racing is your hobby rather than a business, you can't claim a loss on form 1040 to offset other income, and your deductions cannot exceed your prize money; consistent financial losses are a clear indication that you lack a profit motive. if you use at least five vehicles, you can take only actual operating expenses.

- riders must use the same motorcycle throughout a race, with repairs made between heats if necessary. a cotton jersey, nylon pants padded at the knee and thigh, padded boots and gloves, a helmet and goggles, a plastic chest protector, and a kidney belt for support constitute the usual outfit of the motocross cyclist.motocross racingkinney jones britannica quiz sports fun facts quiz what sport's equipment was found in the tomb of an egyptian child buried about 3200 bce?

Retrieved Docs

What sport can you use this for?

Answers: "race", "race", "race", "race", "race", "race", "motocross", "motocross", "ride", "ride"

Qualitative Results: Retriever based VQA

Name the type of plant this is?

- if you have a houseplant, but you don't know the exact name, this guide is made for you!1. a succulent-cactus, a vine, a fern, or another type of herbaceous plant.3. your plant's name, you can download a mobile app like picture this, plantsnap or garden answers.take a picture of the plant, and these apps will tell you the name right away.
- preferred common name plantain taxonomic treedomain: eukaryota kingdom: plantae phylum: spermatophyta subphylum: angiospermae class: monocotyledonae there are no pictures available for this datasheet if you can supply pictures for this datasheet please contact: compendia cab international wallingford oxfordshire ox10 8de uk compend@cabi.org don't need the entire report? generate a print friendly version containing only the sections you need.

Retrieved Docs

Answers: "vine", "vine", "vine", "vine", "climb", "climb", "look like some kind of ivy", "look like some kind of ivy", "ficus", "ficus"

DEMO

DEMO LINK http://10.119.2.11:7072/

DEMO QR CODE

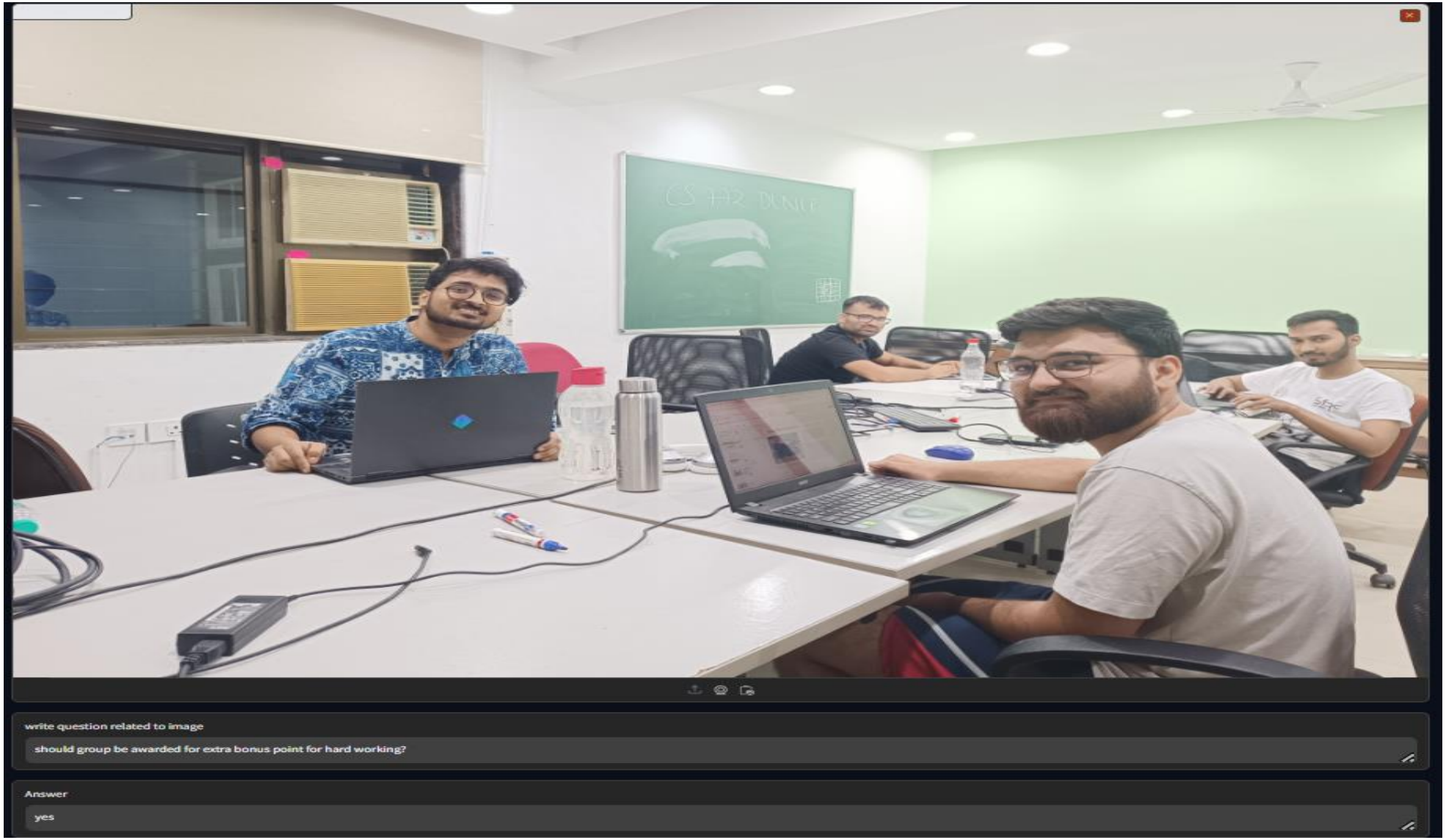
BONUS (Exceeds expectation)

- Performed **Exhaustive** analysis of all method
  Single, Multiple Adapters of different types
- Added COCO task later to perform **relative study**
  Beat COCO using Adapters
- Added
  Explored **SOTA [2023]** retrieval-based method
  Even **created a vector database** for the task
- Even did an ablation study
  Included in Appendix Slides

Bonus: Let Us see what the Model says

Question: Should group be awarded extra bonus for hard working?

Answer: Yes



Results and Analysis : Conclusion

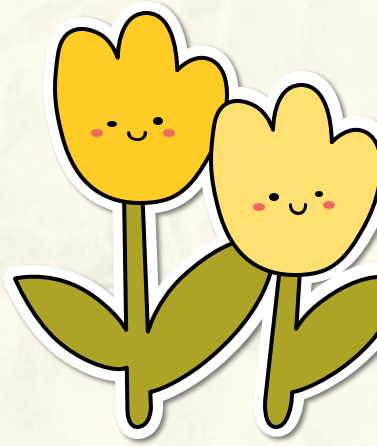
	NLVR	VQA	GQA	Bleu 1	Bleu 2	Bleu 3	Bleu 4	CIDEr	METEOR	SPICE
Full finetune	74.236	70.84	55.569	73.05	56.29	44.982	32.69	108.8	28.17	21.42
Single Adapter	70.977	68.49	53.212	74.09	57.67	44.852	33.91	110.8	28.18	21.07
Multi Adapter	71.035	69.82	52.64	73.74	57.37	43.973	33.71	111.8	28.57	21.45
Hyperformer	72.987	68.84	51.67	73.46	57.2	43.966	33.9	113.2	28.51	21.28
Single Compacter	72.772	68.64	52.4525	74.54	58.37	44.952	34.64	112.9	28.24	21.11
Multi Compacter	70.748	69.37	53.654	74	57.85	44.475	34.17	114.3	28.21	21.2
Single Lora	73.374	69.27	53.7285	74.5	58.24	44.854	34.61	113.5	28.4	21.25
Multi Lora	51.069	68.28	54.135	74.21	57.85	44.42	34.17	113.1	28.32	21.14
Single Prompt	51.801	46.89	37.1526	74.67	58.38	43.68	31.79	101.1	25.68	19.19
Multi Prompt	51.916	48.96	38.65	76.07	59.66	44.7709	32.8	107.1	26.21	19.91
T5 full fine tune	74.2409	67.35	57.3938	72.89257	56.86255	43.489	28.759	112.1811	25.643	21.516
T5 one Adapter	71.2345	67.325	57.199	72.82074	56.764	43.375	28.707	111.277	25.345	21.4971
T5 multi Adapter	70.2386	67.3	57.004	72.7489	56.66592	43.261	28.655	111.7291	25.125	21.47812
one less adapter	70.854	67.586	52.812	72.675	56.028	42.165	31.564	110.853	26.523	20.956
4 linear layer for adapter	70.689	69.564	53.194	73.879	56.321	45.653	32.89	110.654	27.586	21.027
Removed all adapter add one at end	69.9541	70.236	52.64	72.564	55.853	42.897	32.654	109.562	26.351	20.564

THANKING

Prof. Pushpak Bhattacharyya

SPECIAL THANKS TO

Tathagata Dey

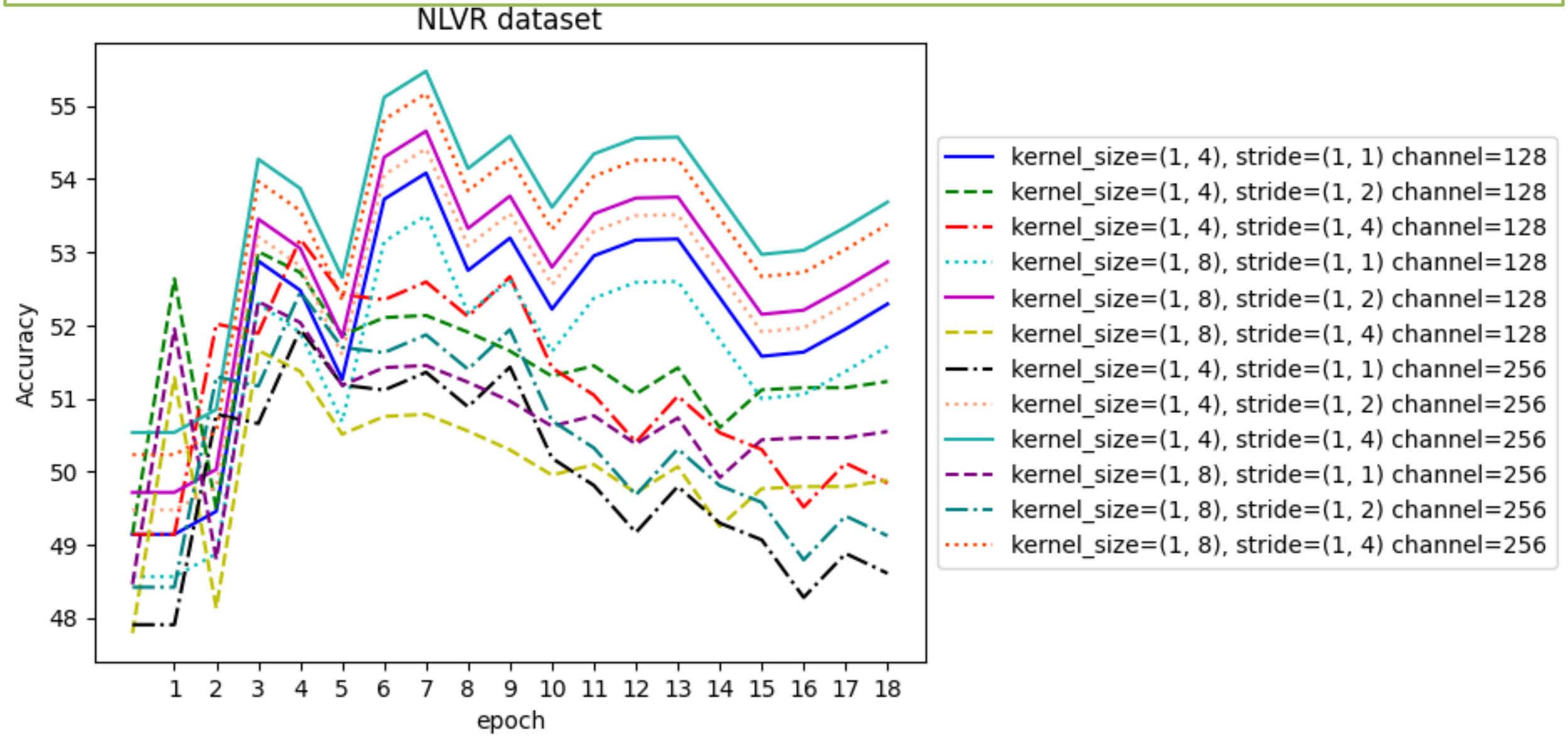


Evaluation Scheme

- Clarity of problem statement (5)
- Interestingness and complexity of the problem statement (5)
- Literature Survey (10)
- Data Handling (15)
- Problem Modelling (15)
- Results and Analysis (20)
  - Quantitative (10)
  - Qualitative (10)
- Demo (25)
- Bonus (5)

TOTAL - 100

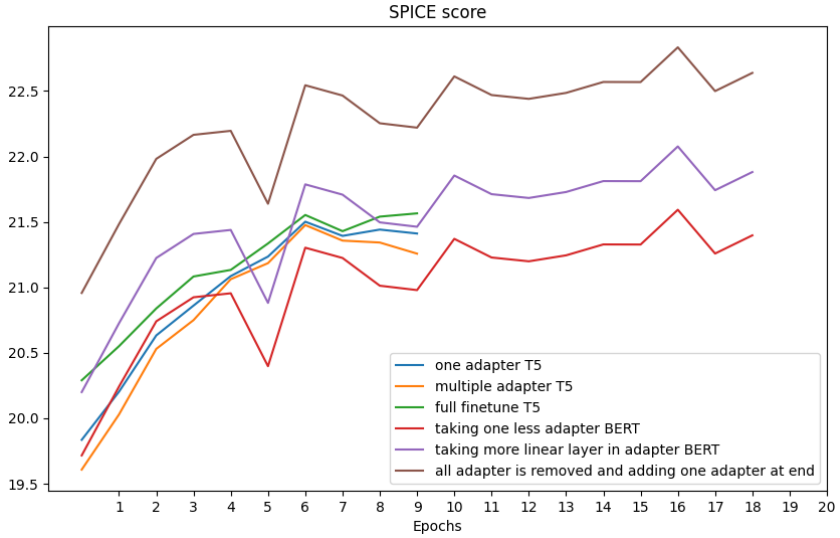
Results and Analysis : Ablation Study

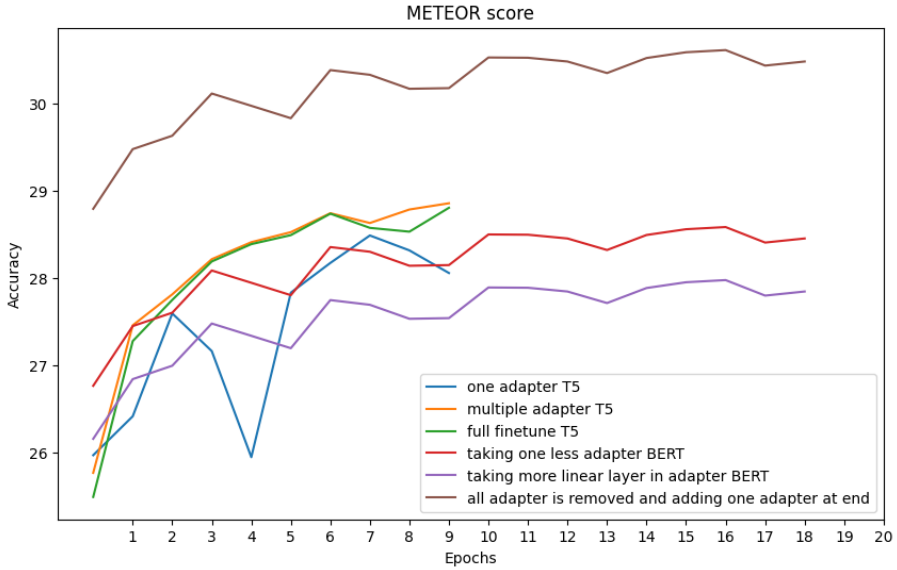


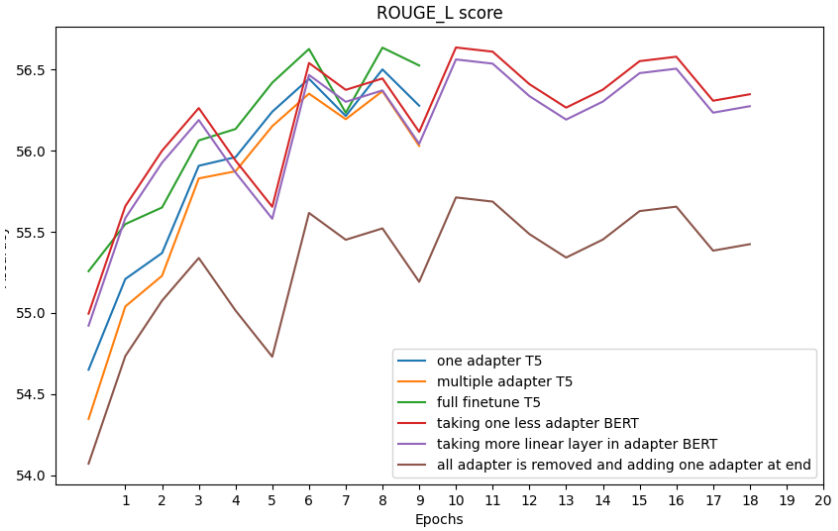
PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

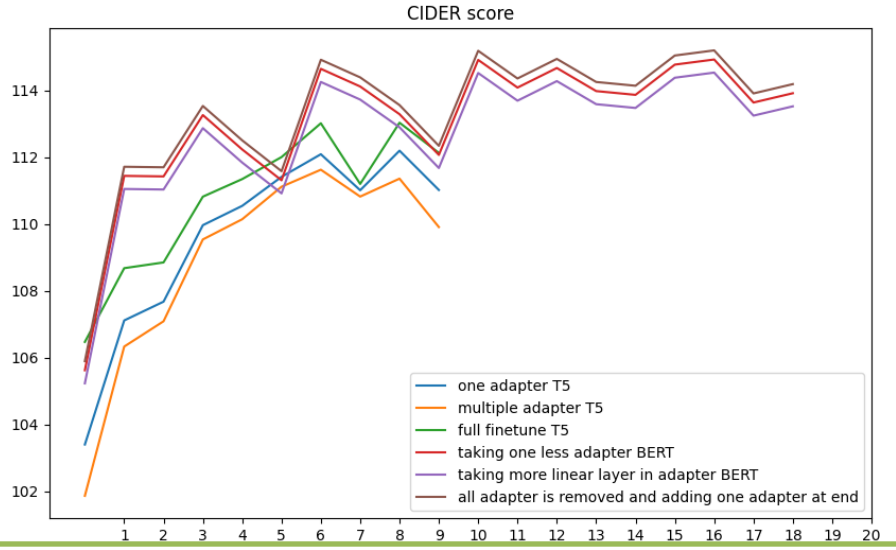
Results and Analysis : Ablation Study

RESULTS OF THE TRAINED MODULES ON THREE DATASETS









PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Hyper-network based adapter

Compute task embedding $I_\tau \in R^t$ for individual task using task projector network $h_I(.)$

$$I_\tau = h_I(Z_\tau)$$
$$Z_\tau \in R^{\{t'\}} \; is \; learnable \; parameter$$

$h_I(.)$ is MLP consisting two feed forward NN and ReLU activation

Consider Hypernetwork $h_A^l(.)$ for generating (U_τ^l, D_τ^l) :

$$\left(U_\tau^l, D_\tau^l\right) = \mathbf{h}_\mathbf{A}^\mathbf{l}(\mathbf{I}_\tau) = \left(\mathbf{W}^{\{\mathbf{U^l}\}}, \mathbf{W}^{\{\mathbf{D^l}\}}\right)\mathbf{I}_\tau$$

where $\mathbf{W}^{\{\mathbf{U^l}\}} \in R^{\{(d_i * d) * t\}}$, $W^{\{\mathbf{D}^l\}} \in R^{\{(d * d_i) * t\}}$

Another Hypernetwork $h_{\{LN\}}^l(.)$:

$$(\gamma_\tau^l, \beta_\tau^l) = h_{LN}^l(I_\tau) = \left(W^{\gamma^l}, W^{\beta^l}\right) I_\tau$$
$$where \;\; W^{\gamma^l}, W^{\beta^l} \in R^{h*t}$$

Source: Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In Annual Meeting of the Association for Computational Linguistics, 2021

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Evaluation Metric: BLEU

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram Precision $= 2/7$.[3]

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

$$w_n = 1/N$$

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Evaluation Metric: CIDEr

$$S_i = \{s_{i1}, \ldots, s_{im}\}$$

All word are mapped to their stem or root word in C_i and S_i
Performed TF-IDF weighting for each n-gram

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}\right)$$

$h_k(s_{ij})$ =number of times an ngram w_k occurs in reference sentence s_{ij}
Ω = vocabulary of all n-gram
I=set of all image in dataset

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Evaluation Metric: CIDEr

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g^n}(c_i) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}$$

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i, S_i)$$

$$w_n = 1/N$$

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Evaluation Metric: METEOR

$$P(Precision) = \frac{\#mapped\_unigrams}{\#unigrams\_in\_candidate}, R(Recall) = \frac{\#mapped\_unigrams}{\#unigrams\_in\_reference}$$

$$Fscore = \frac{10PR}{R + 9P}$$

$$\text{Penalty} = 0.5 * \left[\frac{\#chunks}{\#unigrams\_matched} \right]^3$$

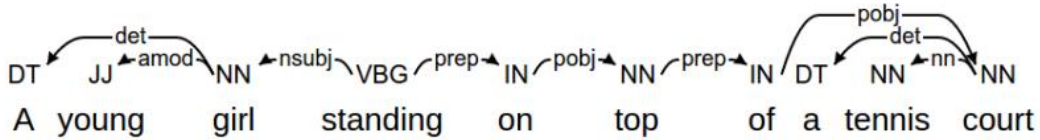
METEOR Score = Fscore*(1-Penalty)

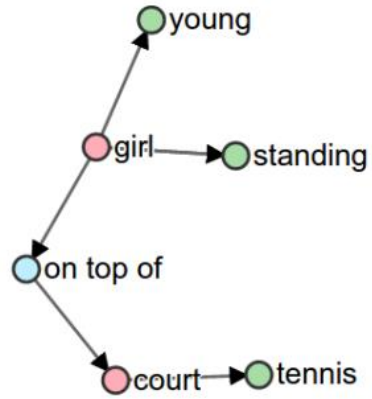
PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Evaluation Metric: SPICE

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

$$T(G(c)) \triangleq O(c) \cup E(c) \cup K(c)$$





{ (girl), (court), (girl, young), (girl, standing)
(court, tennis), (girl, on-top-of, court) }

Fig. 1. Illustrates our method's main principle which uses semantic propositional content to assess the quality of image captions. Reference and candidate captions are mapped through dependency parse trees (top) to semantic *scene graphs* (right)—encoding the objects (red), attributes (green), and relations (blue) present. Caption quality is determined using an F-score calculated over tuples in the candidate and reference scene graphs

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Evaluation Metric: SPICE

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)}$$

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

Evaluation Metric: SPICE

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)}$$

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide

SOTA

Task	Accuracy (CLIP Encoder - BART LM)	Accuracy (Fast RCNN Encoder - BART LM)
NLVR2	74.236	72.3
VQA	70.842	67.8
GQA	55.569	57.3

Method	Updated Params (%)	VQA Karpathy test Acc. (%)	GQA test-dev Acc. (%)	NLVR^2 test-P Acc. (%)	COCO Cap. Karpathy test CIDEr	Avg.
VL-BART [7]	100.00	**67.8**	**57.3**	72.3	109.4	76.7
CLIP-BART + Full fine-tuning	100.00	67.6	56.7	**73.0**	**112.9**	**77.6**

Task	Model	Accuracy
NLVR2	BEiT-3	92.58
VQA	PaLI	84.3
GQA	CFR	72.1

PLEASE NOTE: This slide is not part of the presentation. This is an APPENDIX slide