ViInfographicVQA

Group 7 – Text Mining

May 18, 2025

INTRODUCTION

Vietnamese

Fully built in Vietnamese

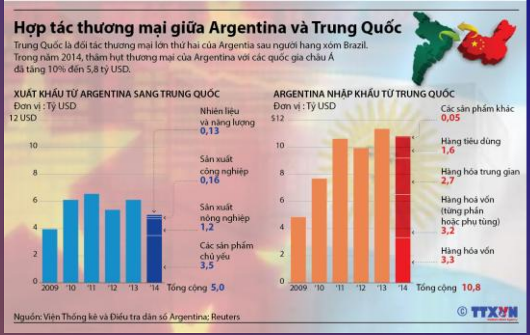
Infographic

Infographic images from different quality newspaper sources

VQA

Question-answering according to the content of the object

INTRODUCTION



Hợp tác thương mại giữa Argentina và Trung Quốc

Trung Quốc là đối tác thương mại lớn thứ hai của Argentia sau người hang xóm Brazil. Trong năm 2014, thâm hụt thương mại của Argentina với các quốc gia châu Á đã tăng 10% đến 5,8 tỷ USD.

Nguồn: Viện Thống kê và Điều tra dân số Argentina; Reuters

Question

Answer

DATASET STRUCTURE

Infographic Images

Consists of approximately 35,000 images crawled from various news sources.

QA pairs

Pairs of question-answers that cover the image’s content:

Text question
Non-Text question

TTXVN THANH NIÊN Nhân Dân

01 Data

PIPELINE

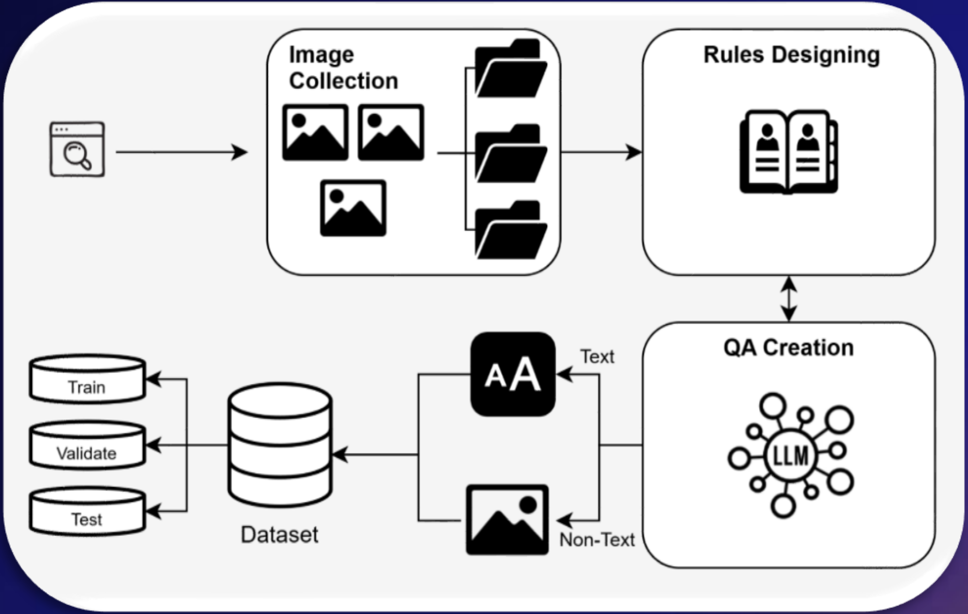
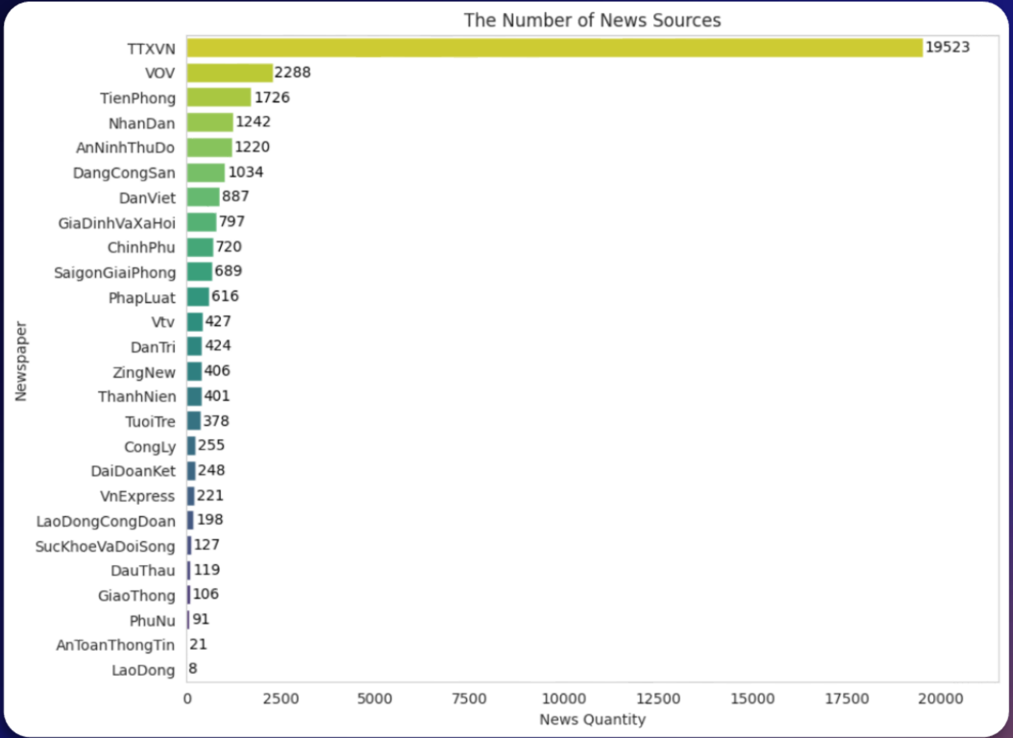


Image Collection
Rules Designing
QA Creation
LLM
Text
Non-Text
Dataset
Train
Validate
Test

INFOGRAPHIC COLLECTION



QA Generation

Rules and Constraints

- *Number of QAs:* About 5 pairs per image.
- *QA length:* Should not exceed 30 words.
- *Colors:* Restricted.
- *Question*
  - Avoid Yes/No and choice-based questions.
  - Ensure sufficient data.
  - No deep analysis or outside inference.
  - Include comparison for numerical questions.
  - Specify criteria for name-related questions.
- *Answer*
  - Should be a complete sentence.
  - Include a clear explanation.

QA Generation

Classification

Text QA

- ✓ Numerical data.
- ✓ Textual information.
- ✓ Any text present in the infographic.

Non-text QA

- ✓ Object.
- ✓ Colors.
- ✓ Chart shapes.
- ✓ Position on the map.

QA Generation

Classification

Text QA

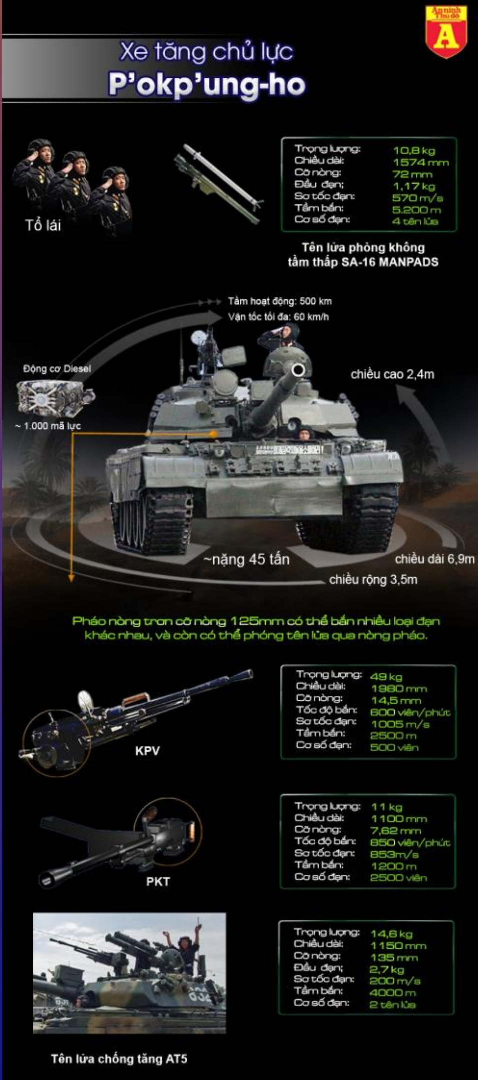
Q: Có bao nhiêu loại vũ khí được liệt kê có tầm bắn lớn hơn 2000 mét?

A: Có 3 loại vũ khí có tầm bắn lớn hơn 2000 mét: Tên lửa phòng không tầm thấp SA-16 MANPADS (5200m), KPV (2500m), và Tên lửa chống tăng AT5 (4000m).

Non-text QA

Q: Có bao nhiêu người đang đội mũ bảo hiểm trong hình minh họa 'Tổ lái' ở góc trên bên trái của infographic?

A: Có 3 người đang đội mũ bảo hiểm trong hình minh họa 'Tổ lái'.



Xe tăng chủ lực
P'okp'ung-ho

Tổ lái

Trọng lượng:	10,8 kg
Chiều dài:	1574 mm
Cỡ nòng:	72 mm
Đầu đạn;	1,17 kg
Sơ tốc đạn:	570 m/s
Tầm bắn:	5.200 m
Cơ số đạn:	4 tên lửa

Tên lửa phòng không tầm thấp SA-16 MANPADS

Tầm hoạt động: 500 km
Vận tốc tối đa: 60 km/h

Động cơ Diesel

~ 1.000 mã lực

chiều cao 2,4m

~nặng 45 tấn

chiều dài 6,9m

chiều rộng 3,5m

Pháo nòng trơn cỡ nòng 125mm có thể bắn nhiều loại đạn khác nhau, và còn có thể phóng tên lửa qua nòng pháo.

KPV

Trọng lượng:	49 kg
Chiều dài:	1980 mm
Cỡ nòng:	14,5 mm
Tốc độ bắn:	600 viên/phút
Sơ tốc đạn:	1005 m/s
Tầm bắn:	2500 m
Cơ số đạn:	500 viên

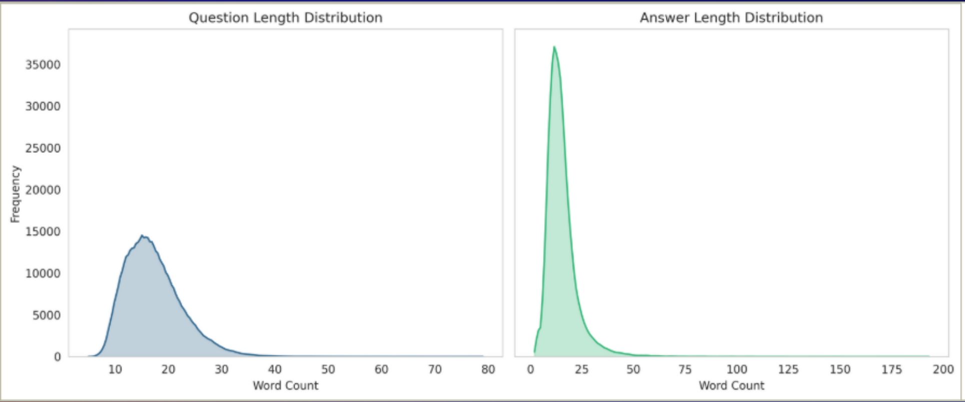
PKT

Trọng lượng:	11 kg
Chiều dài:	1100 mm
Cỡ nòng:	7,62 mm
Tốc độ bắn:	850 viên/phút
Sơ tốc đạn:	853m/s
Tầm bắn:	1200 m
Cơ số đạn:	2500 viên

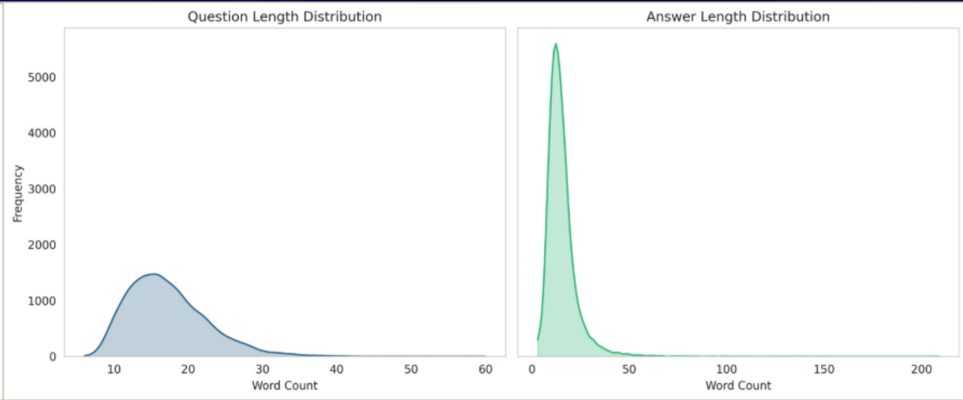
Trọng lượng:	14,6 kg
Chiều dài:	1150 mm
Cỡ nòng:	135 mm
Đầu đạn;	2,7 kg
Sơ tốc đạn:	200 m/s
Tầm bắn:	4000 m
Cơ số đạn:	2 tên lửa

Tên lửa chống tăng AT5

Dataset Preprocessing

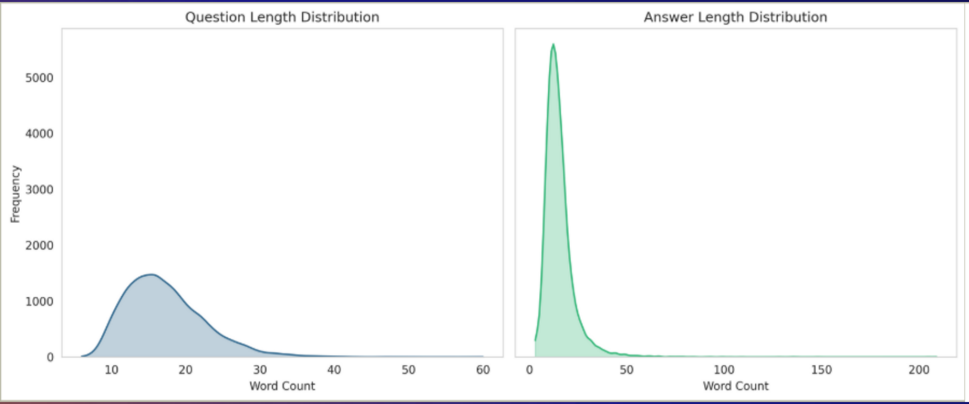


Training set

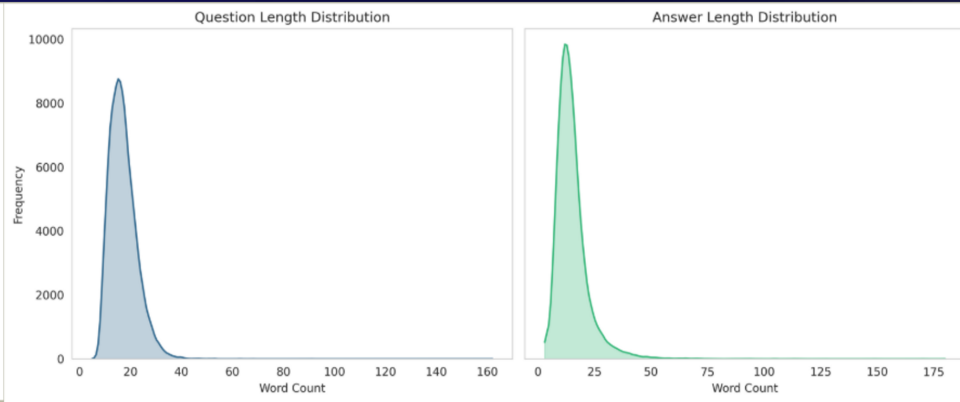


Validation set

Dataset Preprocessing



Validation set



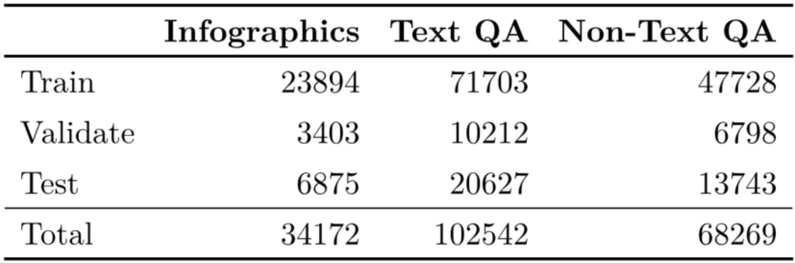
Test set

Dataset Analysis

Initial Statistics

Only take the length of:

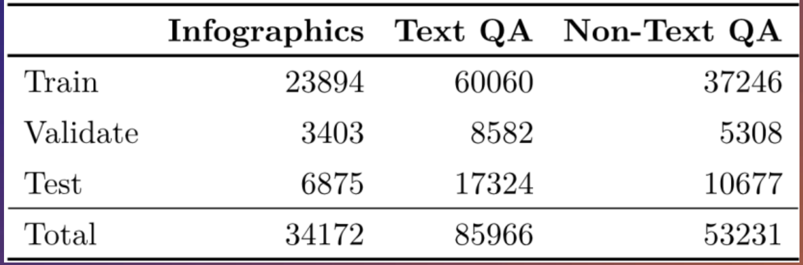
- Question: 10 – 35 words.
- Answer: 10 – 40 words.



	Infographics	Text QA	Non-Text QA
Train	23894	71703	47728
Validate	3403	10212	6798
Test	6875	20627	13743
Total	34172	102542	68269

Before

After



	Infographics	Text QA	Non-Text QA
Train	23894	60060	37246
Validate	3403	8582	5308
Test	6875	17324	10677
Total	34172	85966	53231

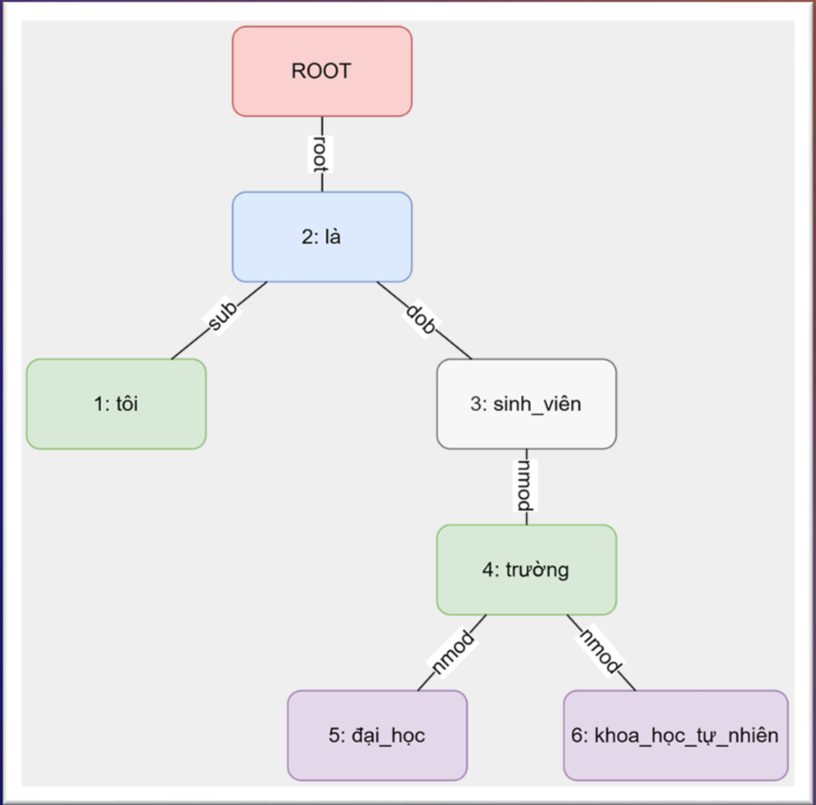
Dataset Analysis

QA Complexity

VnCoreNLP

A powerful and widely used NLP toolkit for Vietnamese text processing.

e.g.: tôi là sinh_viên trường đại_học khoa_học_tự_nhiên



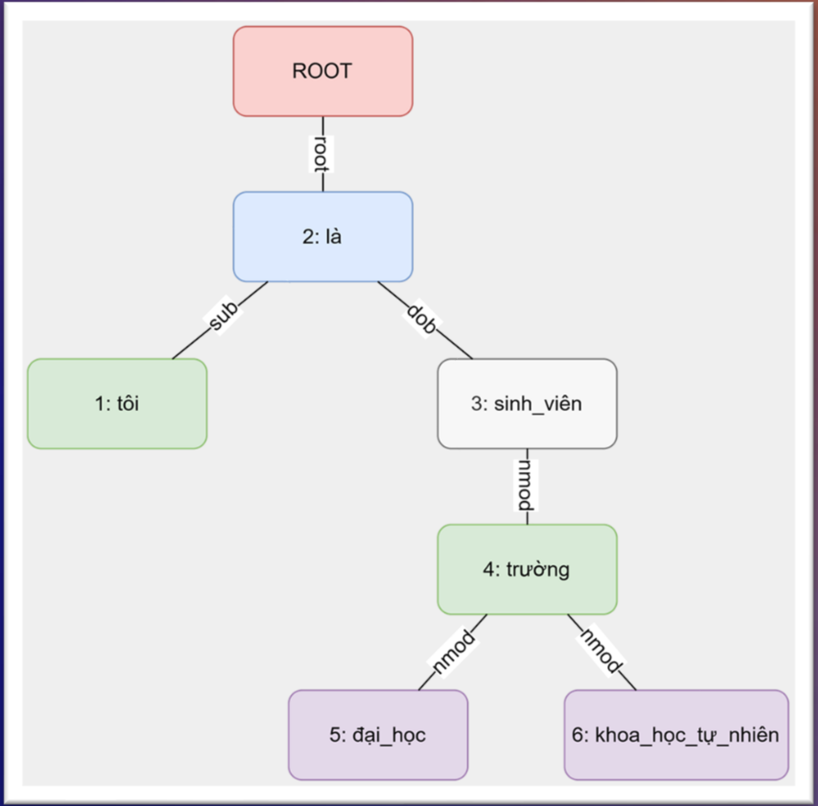
Source: https://github.com/vncorenlp/VnCoreNLP

Dataset Analysis

QA Complexity

PhoNLP

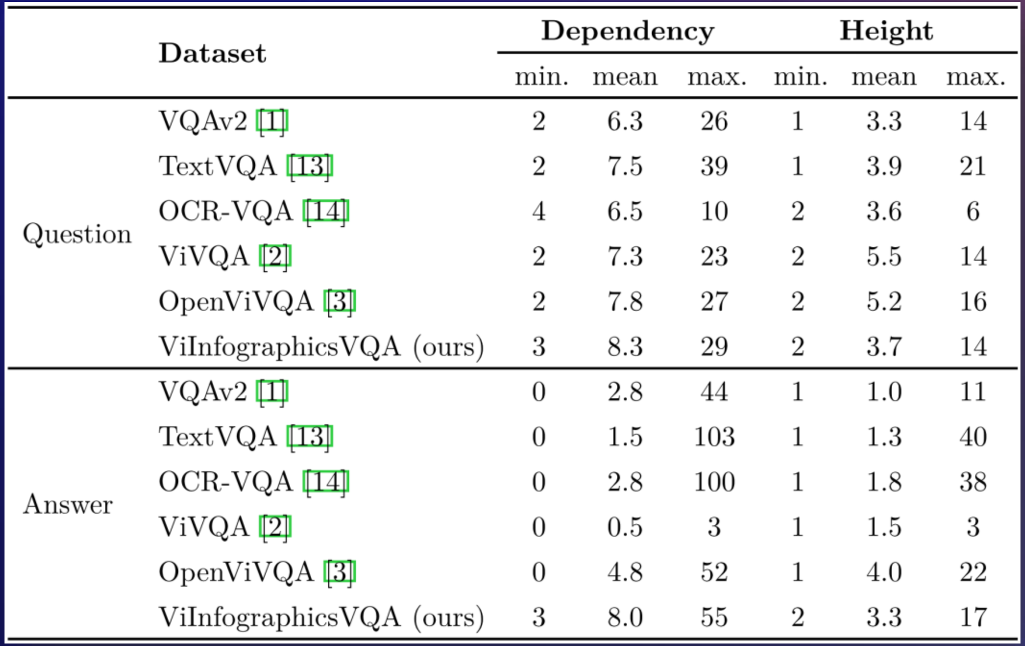
A BERT-based multi-task learning model developed by VinAI Research for joint part-of-speech tagging (POS), named entity recognition (NER), and dependency parsing in Vietnamese.



Source: *https://github.com/VinAIResearch/PhoNLP*

Dataset Analysis

QA Complexity



	Dataset	Dependency			Height		
		min.	mean	max.	min.	mean	max.
Question	VQAv2 [1]	2	6.3	26	1	3.3	14
	TextVQA [13]	2	7.5	39	1	3.9	21
	OCR-VQA [14]	4	6.5	10	2	3.6	6
	ViVQA [2]	2	7.3	23	2	5.5	14
	OpenViVQA [3]	2	7.8	27	2	5.2	16
	ViInfographicsVQA (ours)	3	8.3	29	2	3.7	14
Answer	VQAv2 [1]	0	2.8	44	1	1.0	11
	TextVQA [13]	0	1.5	103	1	1.3	40
	OCR-VQA [14]	0	2.8	100	1	1.8	38
	ViVQA [2]	0	0.5	3	1	1.5	3
	OpenViVQA [3]	0	4.8	52	1	4.0	22
	ViInfographicsVQA (ours)	3	8.0	55	2	3.3	17

Dataset Analysis

Text Normalization

Lowercase

01

VnCoreNLP

02

Excluding non-alphanumeric

03

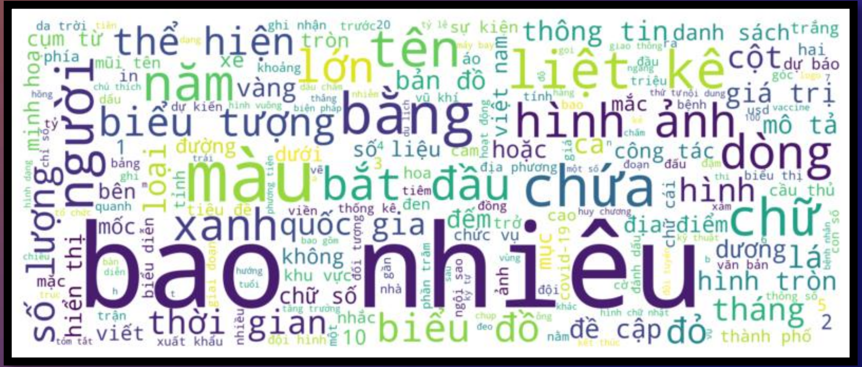
Exclude stopwords

04

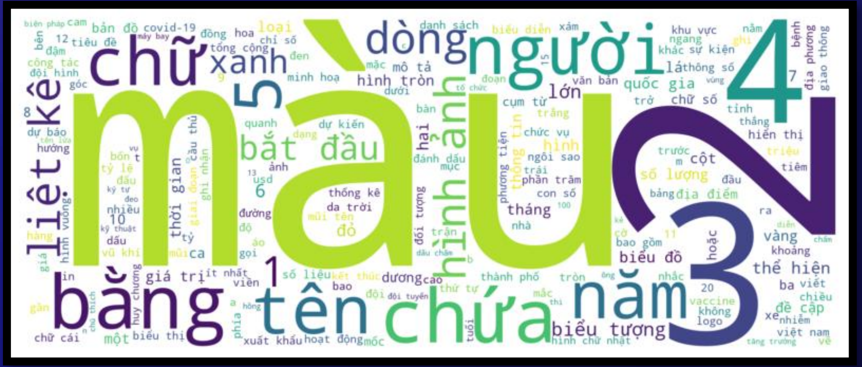
Source: *https://github.com/stopwords/vietnamese-stopwords*

Dataset Analysis

Vocabulary



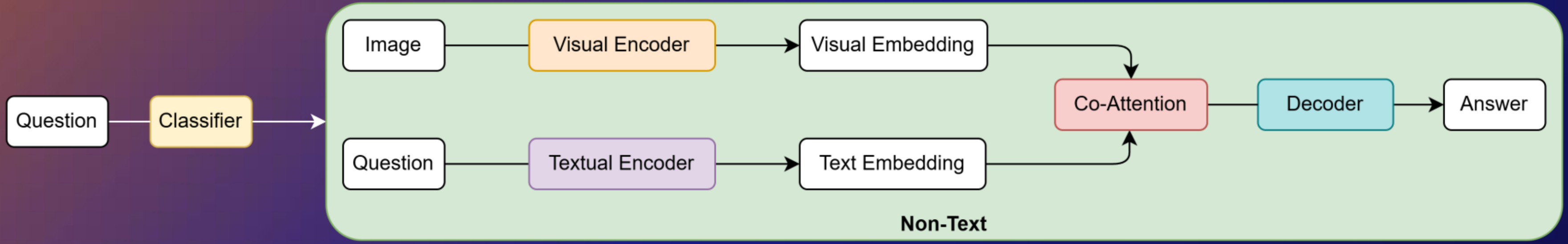
Question



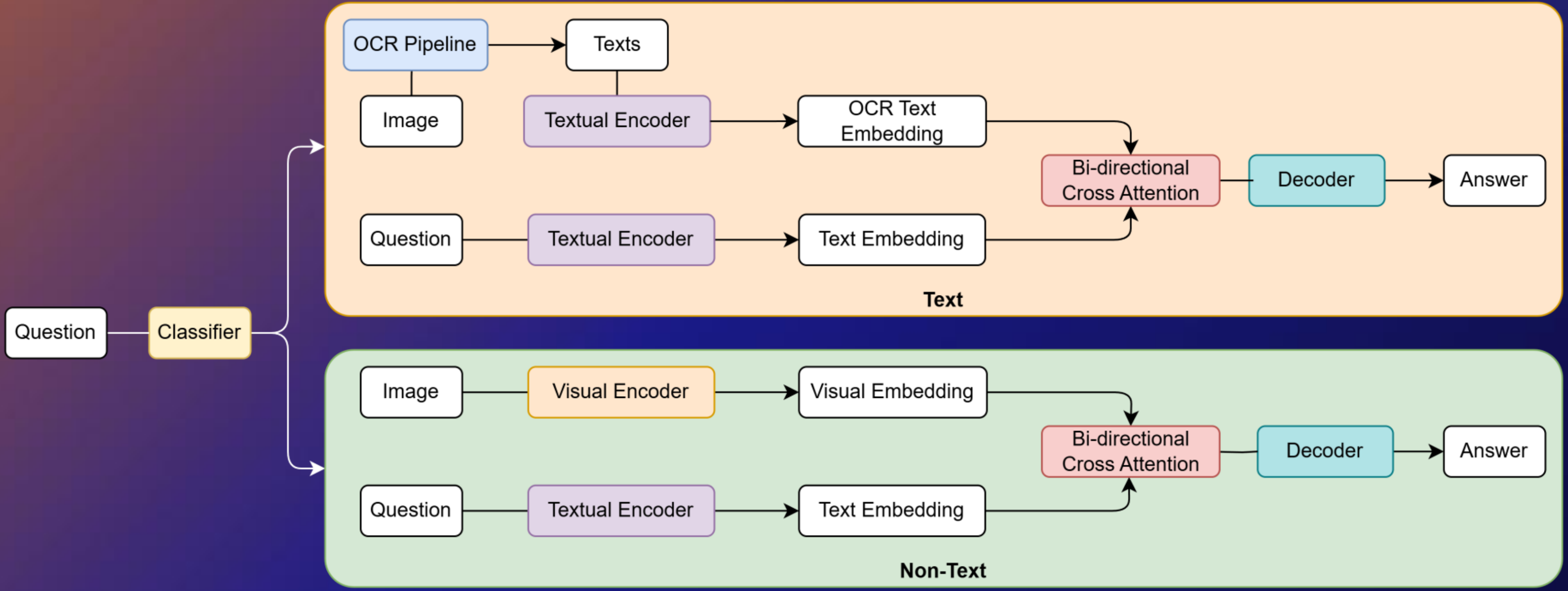
Answer

02 Architecture

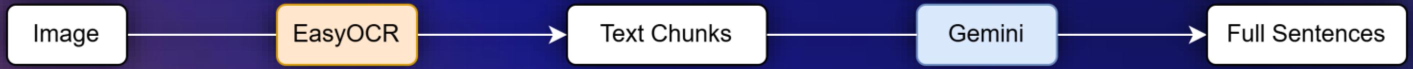
Approach 1

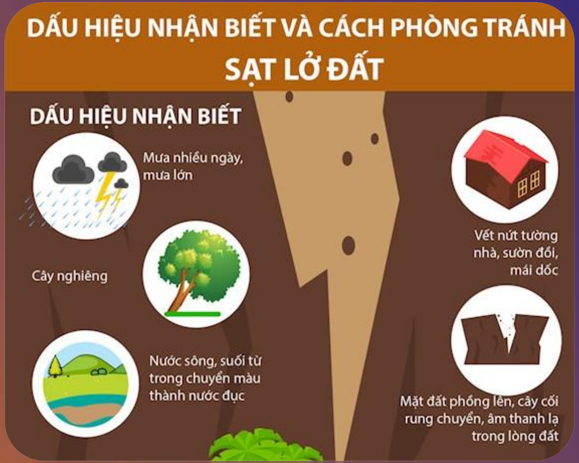

Question
Classifier
Image
Visual Encoder
Visual Embedding
Question
Textual Encoder
Text Embedding
Co-Attention
Decoder
Answer
Non-Text

Approach 2


OCR Pipeline
Texts
Image
Textual Encoder
OCR Text Embedding
Bi-directional Cross Attention
Decoder
Answer
Question
Textual Encoder
Text Embedding
Text
Question
Classifier
Image
Visual Encoder
Visual Embedding
Bi-directional Cross Attention
Decoder
Answer
Question
Textual Encoder
Text Embedding
Non-Text

OCR Pipeline

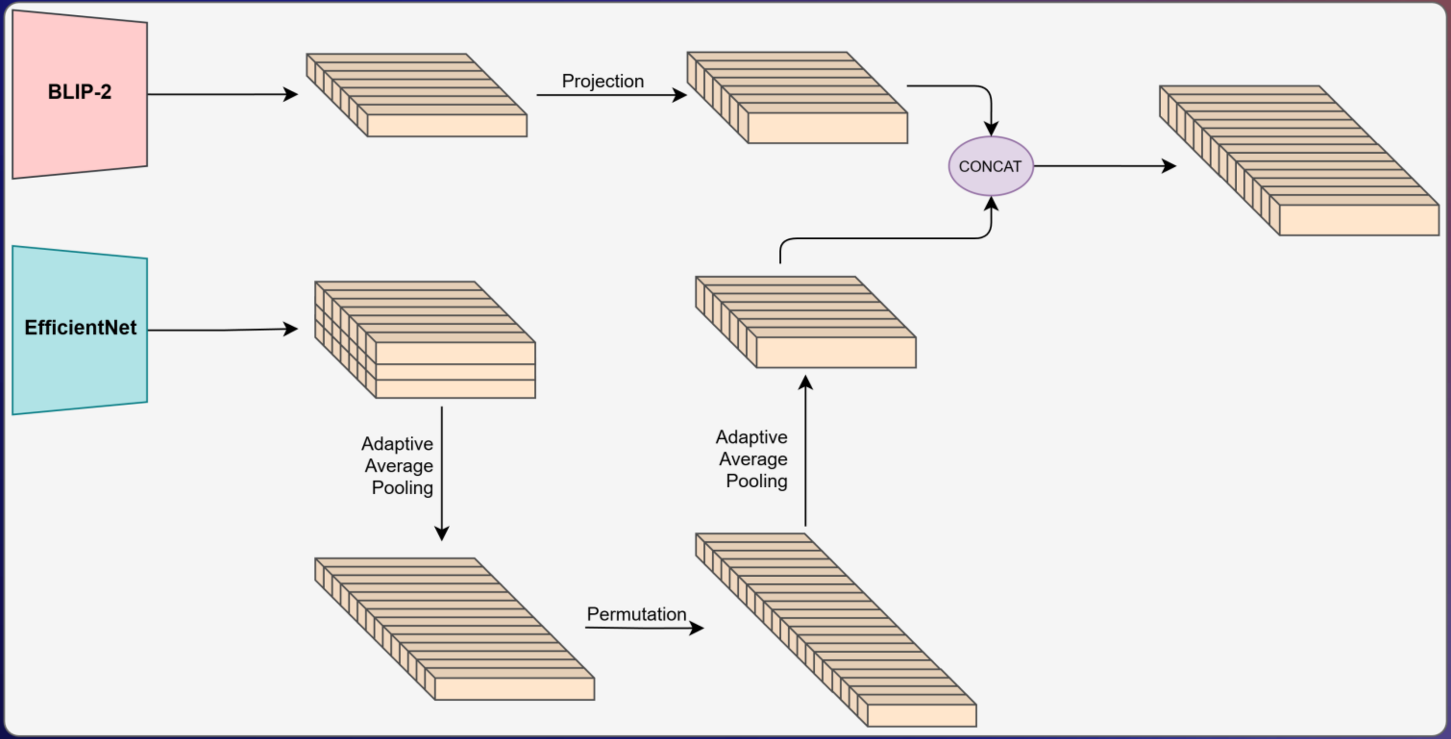



DẤU HIỆU NHẬN BIẾT VÀ CÁCH PHÒNG TRÁNH
SẠT LỞ ĐẤT
DẤU HIỆU NHẬN BIẾT
Mưa nhiều ngày, mưa lớn
Cây nghiêng
Nước sông, suối từ trong chuyển màu thành nước đục
Vết nứt tường nhà, sườn đồi, mái dốc
Mặt đất phồng lên, cây cối rung chuyển, âm thanh lạ trong lòng đất

- Dấu hiệu nhận biết và cách phòng tránh
- Sạt lở đất
- Dấu hiệu nhận biết
- Mưa nhiều ngày
- Mưa lớn
- ...

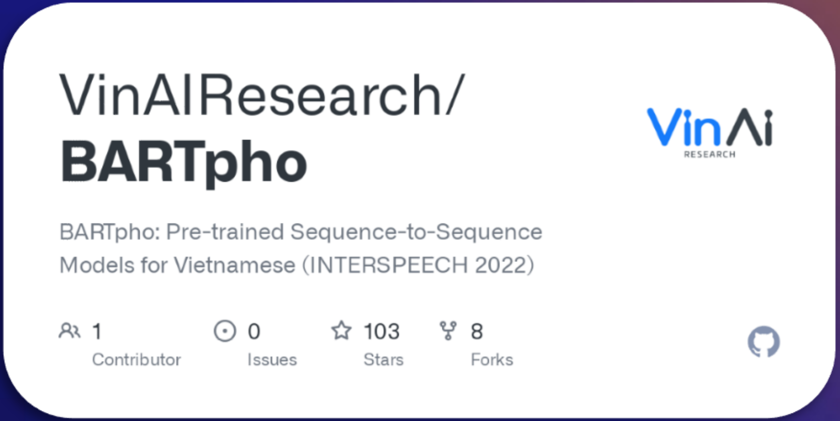
- Dấu hiệu nhận biết và cách phòng tránh sạt lở đất
- Dấu hiệu nhận biết
- Mưa nhiều ngày, mưa lớn
- ...

Visual Encoder



BARTpho

- BARTpho-syllable vs BARTpho-word.
- Seq2Seq model, support both encoder and decoder.
- PEFT *(LoRA).*
- *Encoder trick: concat smaller encoded semantic chunks.*

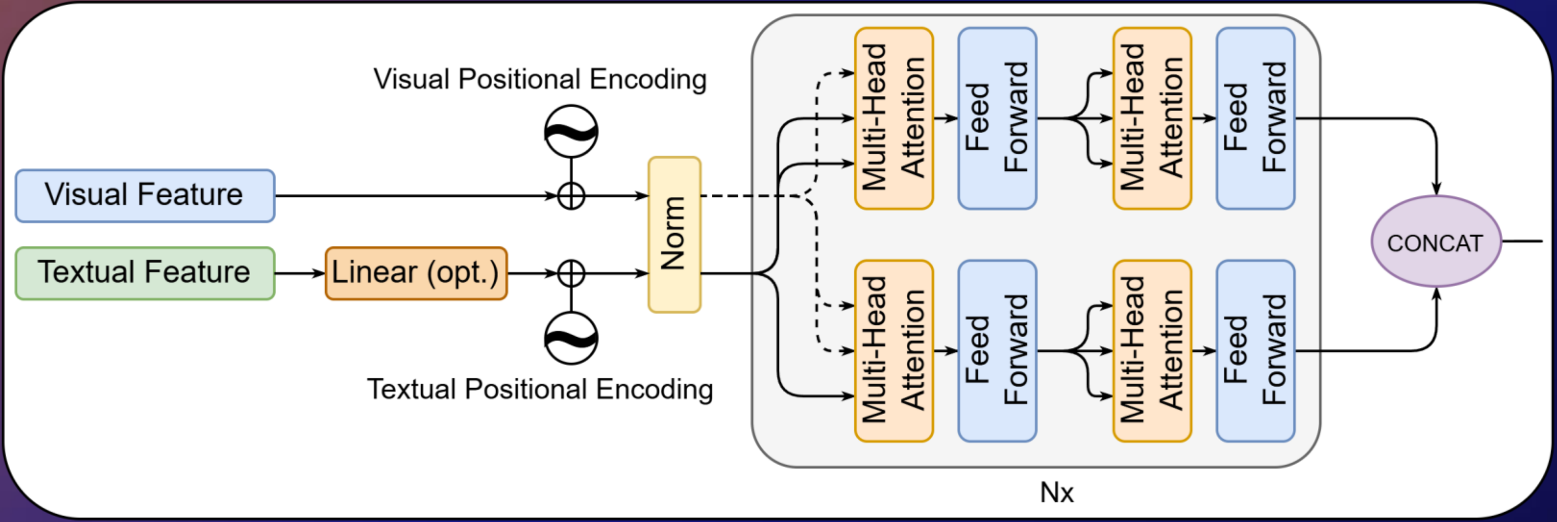


VinAIResearch/
BARTpho

BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese (INTERSPEECH 2022)

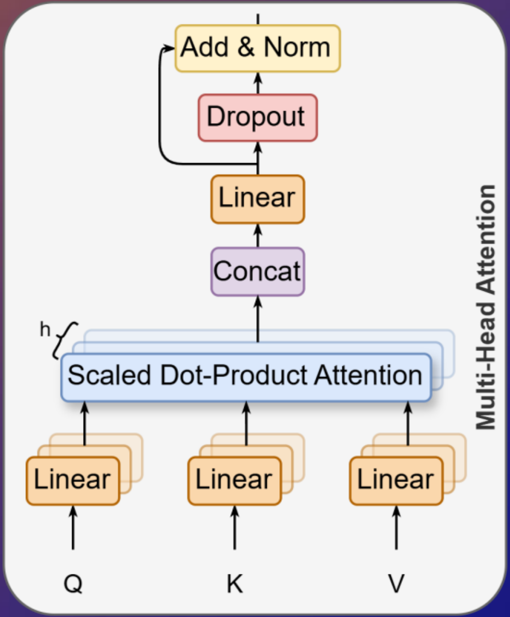
Source: *https://arxiv.org/abs/2109.09701*

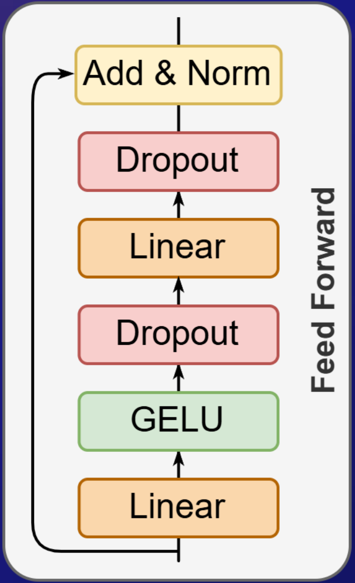
Bi-directional Cross Attention

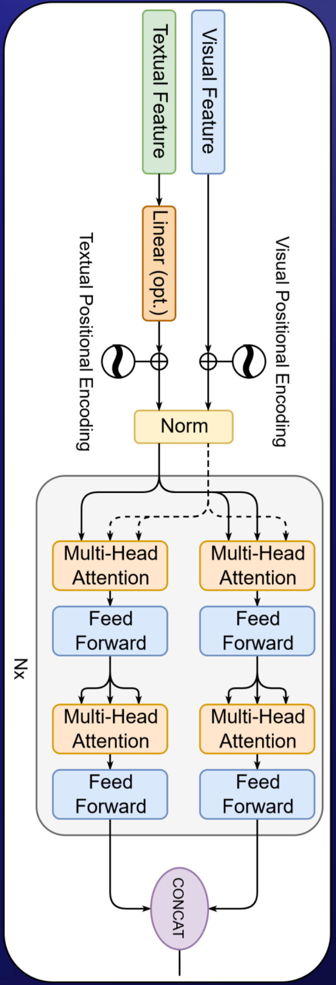


Source: *https://arxiv.org/abs/1606.00061*

Bi-directional Cross Attention







Source: https://arxiv.org/abs/1606.00061

03 Experiments

Metrics

ROUGE

- ROUGE-1
- ROUGE-2
- ROUGE-L

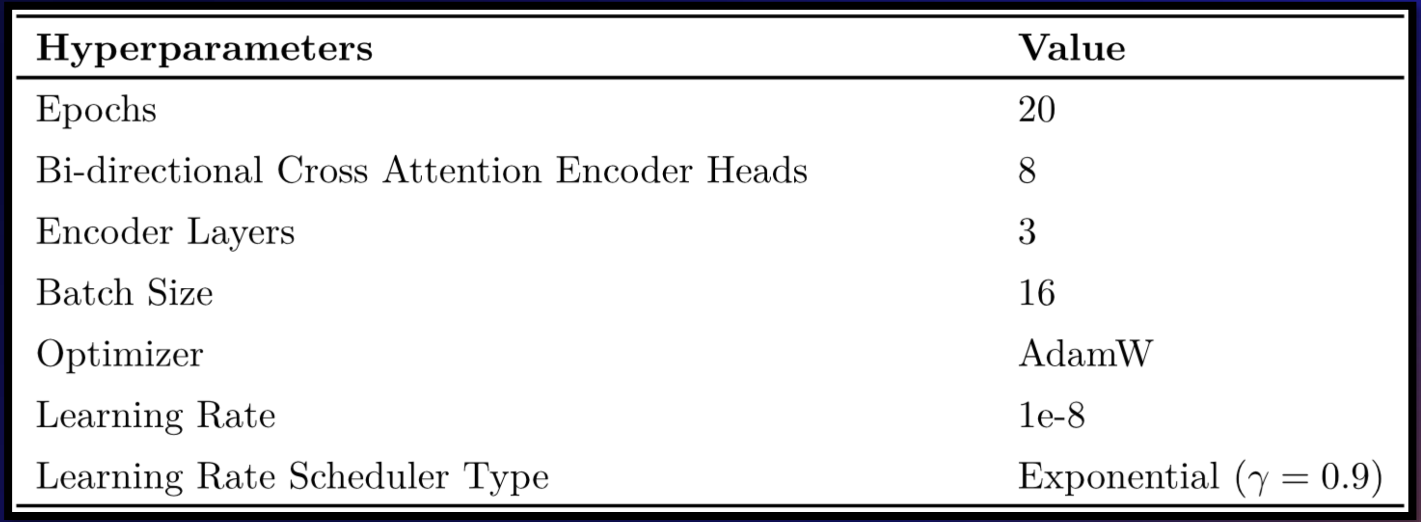
BLEU

- BLEU-4

BERTScore

- Precision
- Recall
- F1 Score

Settings

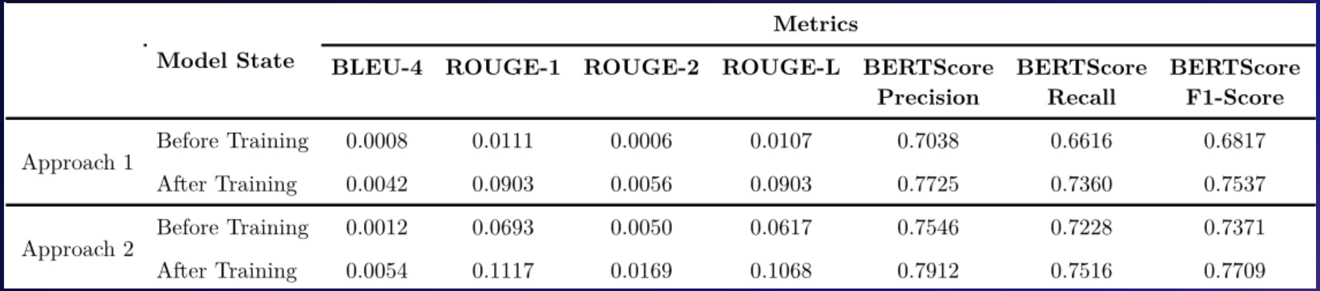


Hyperparameters	Value
Epochs	20
Bi-directional Cross Attention Encoder Heads	8
Encoder Layers	3
Batch Size	16
Optimizer	AdamW
Learning Rate	1e-8
Learning Rate Scheduler Type	Exponential ($\gamma = 0.9$)

Results

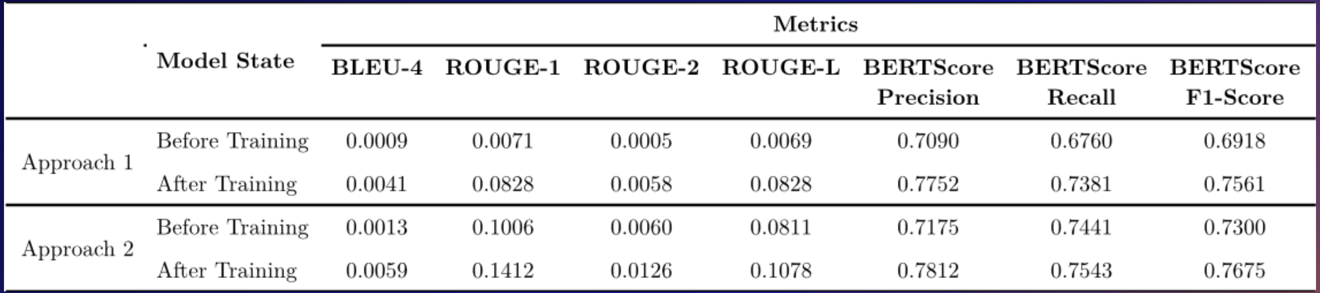
Validation Set

Text



	Model State	Metrics						
		BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore Precision	BERTScore Recall	BERTScore F1-Score
Approach 1	Before Training	0.0008	0.0111	0.0006	0.0107	0.7038	0.6616	0.6817
	After Training	0.0042	0.0903	0.0056	0.0903	0.7725	0.7360	0.7537
Approach 2	Before Training	0.0012	0.0693	0.0050	0.0617	0.7546	0.7228	0.7371
	After Training	0.0054	0.1117	0.0169	0.1068	0.7912	0.7516	0.7709

Non-text

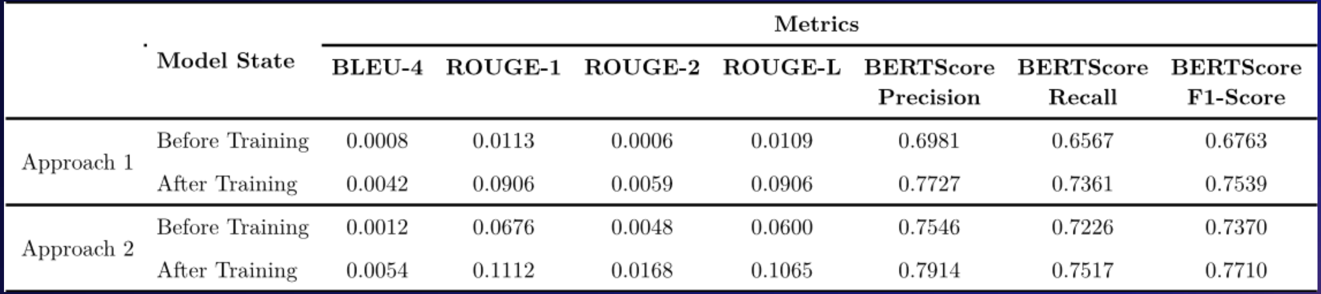


	Model State	Metrics						
		BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore Precision	BERTScore Recall	BERTScore F1-Score
Approach 1	Before Training	0.0009	0.0071	0.0005	0.0069	0.7090	0.6760	0.6918
	After Training	0.0041	0.0828	0.0058	0.0828	0.7752	0.7381	0.7561
Approach 2	Before Training	0.0013	0.1006	0.0060	0.0811	0.7175	0.7441	0.7300
	After Training	0.0059	0.1412	0.0126	0.1078	0.7812	0.7543	0.7675

Results

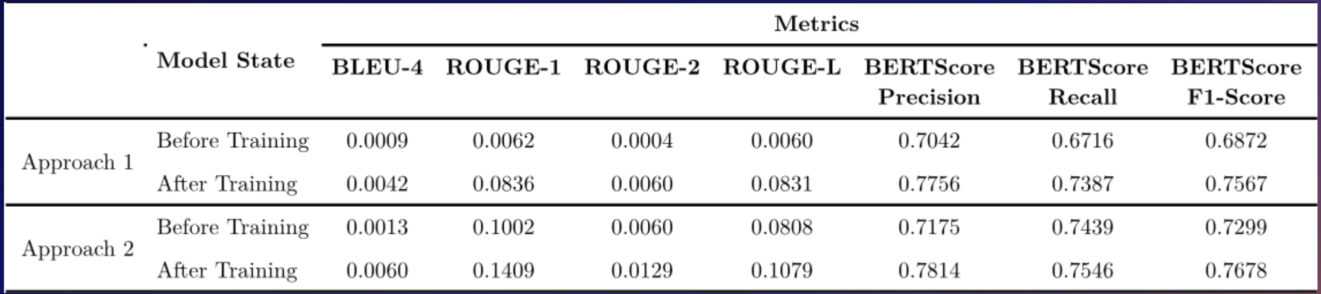
Test Set

Text



	Model State	Metrics						
		BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore Precision	BERTScore Recall	BERTScore F1-Score
Approach 1	Before Training	0.0008	0.0113	0.0006	0.0109	0.6981	0.6567	0.6763
	After Training	0.0042	0.0906	0.0059	0.0906	0.7727	0.7361	0.7539
Approach 2	Before Training	0.0012	0.0676	0.0048	0.0600	0.7546	0.7226	0.7370
	After Training	0.0054	0.1112	0.0168	0.1065	0.7914	0.7517	0.7710

Non-text

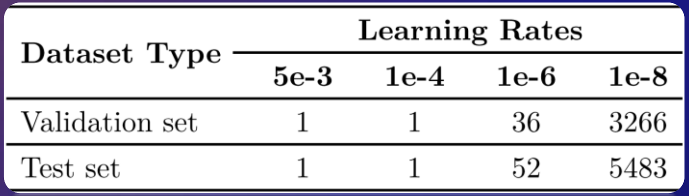


	Model State	Metrics						
		BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore Precision	BERTScore Recall	BERTScore F1-Score
Approach 1	Before Training	0.0009	0.0062	0.0004	0.0060	0.7042	0.6716	0.6872
	After Training	0.0042	0.0836	0.0060	0.0831	0.7756	0.7387	0.7567
Approach 2	Before Training	0.0013	0.1002	0.0060	0.0808	0.7175	0.7439	0.7299
	After Training	0.0060	0.1409	0.0129	0.1079	0.7814	0.7546	0.7678

04 Discussion

Discussion

Learning rate & Answer Variations



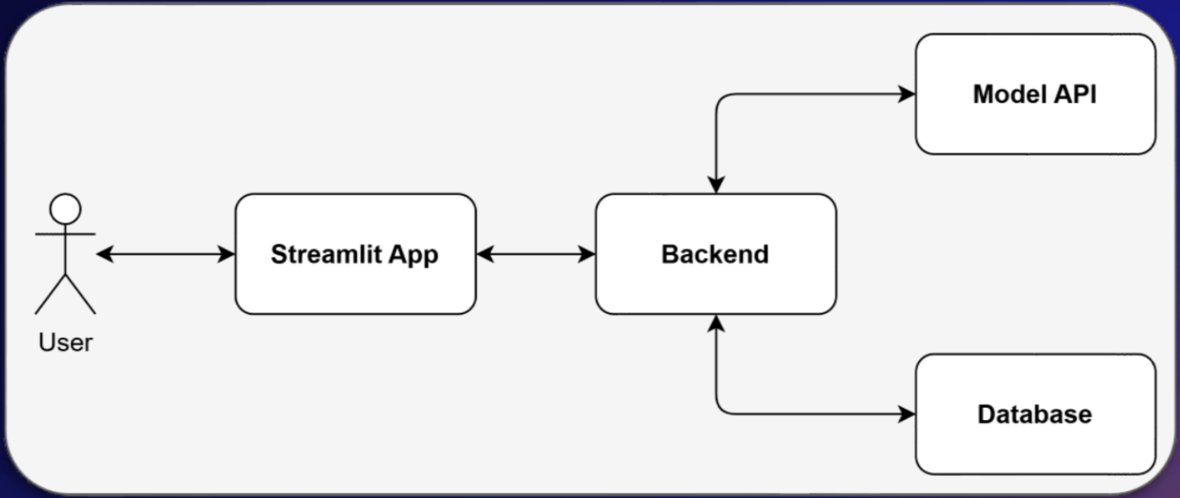
Dataset Type	Learning Rates			
	5e-3	1e-4	1e-6	1e-8
Validation set	1	1	36	3266
Test set	1	1	52	5483

Classifier’s Accuracy

93.21%

05 Application

Application



For more detail, please review our demo.

Thank you