SUPPLEMENTARY MATERIAL – PATHOLOGICAL VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

1 CODE

The code is availiable on Anonymous Github. https://anonymous.4open.science/r/f4d98701-c90c-42ab-a98c-80910a0cf05a/

The paths for different methods are listed in Table 1.

Table 1: Code for different methods

Name	Links
Method 1 without image	method1/src/tasks/pvqa.py[1]
Method 1	method1/src/tasks/pvqa.py[1]
Method 1 with ignoring	method1/src/tasks/pvqa_ignoring.py[1]
Method 1 with CMSSL-IQ	method1/src/pretrain/lxmert_pretrain.py[1]
Method 1 with CMSSL-IA	method1/src/pretrain/lxmert_pretrain.py[1]
Method 1 with SSL-QA	method1/src/pretrain/lxmert_pretrain.py[1]
Method 1 with joint pretraining	method1/src/pretrain/lxmert_pretrain.py[1]
Method 1 with joint pretraining+ignoring	method1/src/tasks/pvqa_ignoring.py[1]
Method 2 without image	method2/finetune_main.py[1]
Method 2	method2/finetune_main.py[1]
Method 2 with ignoring	method2/finetune_main_ignore.py[1]
Method 2 with with CMSSL-IQ	method2/pretrain_main.py[1]
Method 2 with CMSSL-IA	method2/pretrain_main.py[1]
Method 2 with SSL-QA	method2/pretrain_main.py[1]
Method 2 with joint pretraining	method2/pretrain_main.py[1]
Method 2 with joint pretraining+ignoring	method2/finetune_main_ignore.py[1]

[1] `https://anonymous.4open.science/r/f4d98701-c90c-42ab-a98c-80910a0cf05a/`

2 EXPERIMENTAL SETUP

The data split is performed to ensure the frequencies of question types in training, validation, and testing set to be consistent, and most words of the vocabulary exist in each set. The ratio of training set size, validation set size, and test set size is 3:1:1. We implement the methods using PyTorch and perform training on four GTX 1080Ti GPUs.

3 HYPERPARAMETER SETTINGS

The hyperparameter settings in each experiment are specified in Table 2-15.

Table 2: Hyperparameter settings for Method 1

Name	Value	Description
Optimizer	Adam	-
Learning rate	0.00005	Cosine scheduling
Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
Max epoch	200	-
Batch size	32	-

Table 3: Hyperparameter settings for Method 1 with ignoring

Name	Value	Description
Optimizer	Adam	-
Optimizer for ignoring variables	Adam	-
Learning rate	0.00005	Cosine scheduling
Learning rate for ignoring variables	0.1	Weight decay with a rate of 3e-4
Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
Beta for ignoring variables	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
Max epoch	120	-
Batch size	16	-

Table 4: Hyperparameter settings for Method 1 with CMSSL-IQ

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.0001	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	30	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Learning rate	0.00005	Cosine scheduling
	Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	32	-

Table 5: Hyperparameter settings for Method 1 with CMSSL-IA

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.0001	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	30	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Learning rate	0.00005	Cosine scheduling
	Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	32	-

Table 6: Hyperparameter settings for Method 1 with SSL-QA

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.0001	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	30	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Learning rate	0.00005	Cosine scheduling
	Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	32	-

Table 7: Hyperparameter settings for Method 1 with joint pretraining

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.0001	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	30	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Optimizer for ignoring variables	Adam	-
	Learning rate	0.00005	Cosine scheduling
	Learning rate for ignoring variables	0.1	Weight decay with a rate of 3e-4
	Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	32	-

Table 8: Hyperparameter settings for Method 1 with joint pretraining+ignoring

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.0001	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	30	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Optimizer for ignoring variables	Adam	-
	Learning rate	0.00005	Cosine scheduling
	Learning rate for ignoring variables	0.1	Weight decay with a rate of 3e-4
	Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
	Beta for ignoring variables	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	120	-
	Batch size	16	-

Table 9: Hyperparameter settings for Method 2

Name	Value	Description
Optimizer	Adam	-
Learning rate	0.01	Cosine scheduling
Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
Max epoch	200	-
Batch size	256	-

Table 10: Hyperparameter settings for Method 2 with ignoring

Name	Value	Description
Optimizer	Adam	-
Optimizer for ignoring variables	Adam	-
Learning rate	0.1	Cosine scheduling
Learning rate for ignoring variables	0.01	Weight decay with a rate of 3e-4
Beta	(0.9, 0.999)	Adam coefficients used for computing gradients and their squares
Beta for ignoring variables	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
Max epoch	180	-
Batch size	64	-

Table 11: Hyperparameter settings for Method 2 with CMSSL-IQ

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.1	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	120	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Learning rate	0.01	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	256	-

Table 12: Hyperparameter settings for Method 2 with CMSSL-IA

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.1	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	120	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Learning rate	0.01	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	256	-

Table 13: Hyperparameter settings for Method 2 with SSL-QA

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.1	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	120	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Learning rate	0.01	Cosine scheduling
	Beta	(0.5, 0.999)	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	256	-

Table 14: Hyperparameter settings for Method 2 with joint pretraining

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.0001	Cosine scheduling
	Beta	**(0.5, 0.999)**	Adam coefficients used for computing gradients and their squares
	Max epoch	30	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Learning rate	0.00005	Cosine scheduling
	Beta	**(0.9, 0.999)**	Adam coefficients used for computing gradients and their squares
	Max epoch	200	-
	Batch size	256	-

Table 15: Hyperparameter settings for Method 2 with joint pretraining+ignoring

	Name	Value	Description
Pretraining	Optimizer	Adam	-
	Learning rate	0.0001	Cosine scheduling
	Beta	**(0.5, 0.999)**	Adam coefficients used for computing gradients and their squares
	Max epoch	120	-
	Batch size	256	-
Finetuning	Optimizer	Adam	-
	Optimizer for ignoring variables	Adam	-
	Learning rate	0.00005	Cosine scheduling
	Learning rate for ignoring variables	0.1	Weight decay with a rate of 3e-4
	Beta	**(0.9, 0.999)**	Adam coefficients used for computing gradients and their squares
	Beta for ignoring variables	**(0.5, 0.999)**	Adam coefficients used for computing gradients and their squares
	Max epoch	180	-
	Batch size	64	-

4 FILES CONTAINING TRAINED MODELS

Files containing trained models are listed in Table 16.

Table 16: Trained models for different methods

Name	Links
Faster RCNN pretrained on BCCD	faster_rcnn_res101_bccd.pth[2]
Method 1 without image	Method 1/m1_no_image.pth[2]
Method 1	Method 1/m1.pth[2]
Method 1 with ignoring	Method 1/m1_ignoring.pth[2]
Method 1 with CMSSL-IQ	Method 1/m1_IQ.pth[2]
Method 1 with CMSSL-IA	Method 1/m1_IA.pth[2]
Method 1 with SSL-QA	Method 1/m1_QA.pth[2]
Method 1 with joint pretraining	Method 1/m1_joint.pth[2]
Method 1 with joint pretraining+ignoring	Method 1/m1_joint_ignoring.pth[2]
Method 2 without image	Method 2/m2_no_image.pth[2]
Method 2	Method 2/m2.pth[2]
Method 2 with ignoring	Method 2/m2_ignoring.pth[2]
Method 2 with CMSSL-IQ	Method 2/m2_IQ.pth[2]
Method 2 with CMSSL-IA	Method 2/m2_IA.pth[2]
Method 2 with SSL-QA	Method 2/m2_QA.pth[2]
Method 2 with joint pretraining	Method 2/m2_joint.pth[2]
Method 2 with joint pretraining+ignoring	Method 2/m2_joint_ignoring.pth[2]

[2]https://drive.google.com/drive/folders/1tTibOSHNcmMc9rXD7xIwWV4F41ee3A0w?usp=sharing

5 TRAINING TIME

Table 17 shows the training time for each experiment using GTX 1080ti GPUs with the hyper-parameter settings described above.

Table 17: Training time for different experiments

Name	Training time
Method 1 without image	5 hour
Method 1	13 hour
Method 1 with ignoring	34 hour
Method 1 with CMSSL-IQ	21 hour
Method 1 with CMSSL-IA	24 hour
Method 1 with SSL-QA	23 hour
Method 1 with joint pretraining	36 hour
Method 1 with joint pretraining+ignoring	57 hour
Method 2 without image	1.5 hour
Method 2	3 hour
Method 2 with ignoring	17 hour
Method 2 with CMSSL-IQ	4 hour
Method 2 with CMSSL-IA	4 hour
Method 2 with SSL-QA	3 hour
Method 2 with joint pretraining	11 hour
Method 2 with joint pretraining+ignoring	25 hour

6 PARAMETER NUMBERS

Table 18 shows the parameter numbers for different models.

Table 18: Parameter number of different model

Model	Parameter numbers
Method 1 without image	120 million
Method 1	238 million
Method 1 with ignoring	238 million
Method 1 with CMSSL-IQ	257 million
Method 1 with CMSSL-IA	257 million
Method 1 with SSL-QA	257 million
Method 1 with joint pretraining	257 million
Method 1 with joint pretraining+ignoring	258 million
Method 2 without image	137 million
Method 2	84 million
Method 2 with ignoring	84 million
Method 2 with CMSSL-IQ	93 million
Method 2 with CMSSL-IA	93 million
Method 2 with SSL-QA	137 million
Method 2 with joint pretraining	138 million
Method 2 with joint pretraining+ignoring	138 million