https://github.com/heungkyulee/SMILE_Flickr8k

Semipermeable
Max**I**mum
Likelihood **E**stimation

Enhancing Image Captioning through Richness Optimization

Group 7
Kim Dan / Lee Heungkyu / Jung Jiwon

https://github.com/heungkyulee/SMILE_Flickr8k

Overview

1. Paper Review

- Problems of Existing Image Captioning
- SMILE
- Paper Summary

2. Training Process

- Training Overview
- Dataset Preparation & Pre-Processing
- Model Architecture
- Hyperparameters
- Training
- Results

3. Demo

- Demo
- Results
- Challenges & Solutions
- Limitations

README

SMILE-based Image Captioning Project

This project is based on the BLIP and SMILE repositories. It fine-tunes an image captioning model using the Flickr8k dataset and provides a demo through Streamlit.

Table of Contents

1. Project Modifications
2. Flickr8k Dataset Download and Fine-tuning Process
3. Running the Demo with Streamlit

Paper Review

1. Project Modifications

The following changes were made from the original GitHub repositories:

- **Dataset Support**: Replaced the COCO dataset with the **Flickr8k dataset**.
- **Configuration File Updates**: Added the `configs/caption_flickr8k.yaml` file for configurations tailored to the Flickr8k dataset.
- **Custom Data Loader**: Added the `data/flickr8k_dataset.py` file to handle the Flickr8k dataset.
- **Evaluation Function Modifications**: Adjusted evaluation methods for the Flickr8k dataset.
- **Fine-tuning Script Modifications**: Updated `train_caption.py` to support fine-tuning with the Flickr8k dataset.
- **Streamlit Demo**: Added the `app.py` file to provide an image captioning demo using the trained model.

2. Flickr8k Dataset Download and Fine-tuning Process

Publish a Python Package to PyPI on release

https://github.com/heungkyulee/SMILE_Flickr8k

SLSA Generic generator — Configure

Generate SLSA3 provenance for your existing release workflows

Pylint — Configure

Lint a Python application with pylint.

More workflows Dismiss suggestions

https://github.com/heungkyulee/SMILE_Flickr8k

Problems of Existing Image Captioning

Existing MLE(Maximum Likelihood Estimation) **based models** produce **overly generic and simple captions.**

Human

A woman holding a birthday cake with lit candles

MLE-based models

A woman holding a cake with lit candles

https://github.com/heungkyulee/SMILE_Flickr8k

SMILE(Semipermeable Maximum Likelihood Estimation)

SMILE based models can produce **richer context** than existing image captioning models.

Optimization Conflicts

Richness Optimization

Conciseness Optimization

Human

A woman holding a birthday cake with lit candles

MLE-based models

A woman holding a cake with lit candles

SMILE

A pretty young lady that has some kind of white frosted birthday cake with lots of lit candles on top of it, surrounded by several other people looking onwardly at something in the distance.

https://github.com/heungkyulee/SMILE_Flickr8k

SMILE(Semipermeable Maximum Likelihood Estimation)

Key Ideas

- Preserve **Richness** Optimization
- **Block Conciseness** Optimization

How It Works

- During training, the model **predicts within a subset of the vocabulary**
- **Penalizes concise predictions** while allowing richer descriptions.

https://github.com/heungkyulee/SMILE_Flickr8k

Paper Summary

What they used

Base Model

- BLIP

Dataset

- MSCOCO
- Flickr30K

Results

Method	Semantic			Linguistic
	Descrip.	Acc.	F1	Flu.
CapEnrich	3.89	4.39	4.12	4.36
BLIP	3.41	4.57	3.91	**4.91**
BLIP-$\mathcal{L}_{\text{SMILE}}$	**4.67**	4.05	**4.34**	4.75
Human	3.53	**4.53**	3.97	**4.87**

Row	Basic Model	Further Training	Caption Length	Lexical Diveristy	Self-Retrieval		CLIPScore	PPL
					R@1	R@5		
1	*F.S.*	-	9.6	0.6	1.9	6.3	73.9	56.2
2		$\mathcal{L}_{\text{SMILE}}$	15.4	0.7	0.7	2.5	63.7	86.7
3	*PT*	-	5.8	1.0	2.7	8.6	74.6	573.4
4		$\mathcal{L}_{\text{SMILE}}$	22.5	4.9	8.5	20.1	73.5	113.7
5	*PT+FT*	-	10.0	1.4	6.7	16.6	**77.2**	95.8
6		$\mathcal{L}_{\text{MLE}}$	10.0	1.4	6.5	16.6	**77.2**	67.6
7		$\mathcal{L}_{\text{SMILE}}$	**22.3**	**4.5**	**10.0**	**24.5**	75.0	95.6

https://github.com/heungkyulee/SMILE_Flickr8k

Paper Summary

Conclusions

- Addresses the shortcomings of MLE by guiding models to produce richer, more detailed captions.
- Simple and compatible with existing models

Limitations

- Limited applicability to large-scale text generation tasks.
- Relies on the quality and richness of the training dataset.

https://github.com/heungkyulee/SMILE_Flickr8k

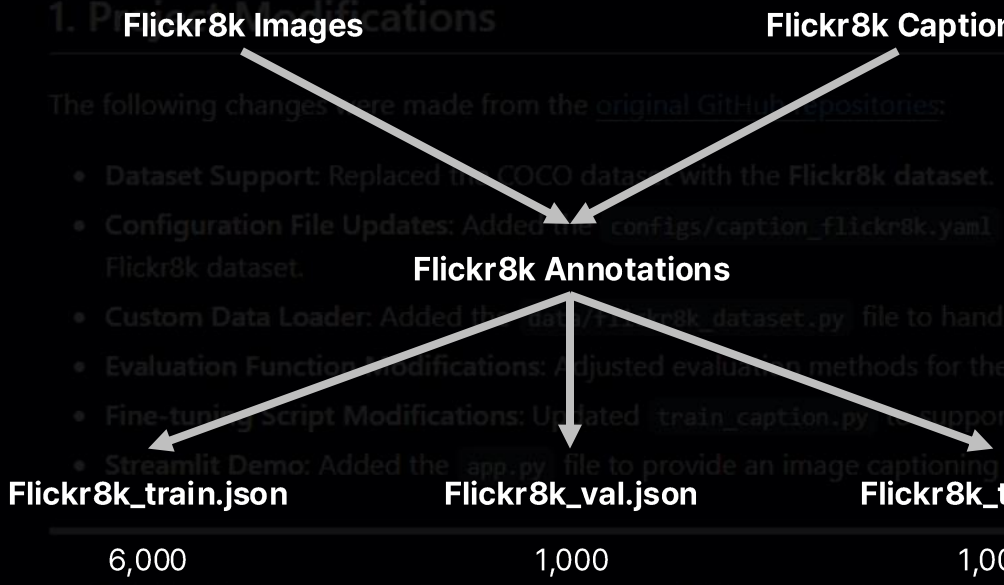
Training Process

https://github.com/heungkyulee/SMILE_Flickr8k

Training Overview

https://github.com/heungkyulee/SMILE_Flickr8k

Dataset Preparation & Pre-Processing



Tokenizing

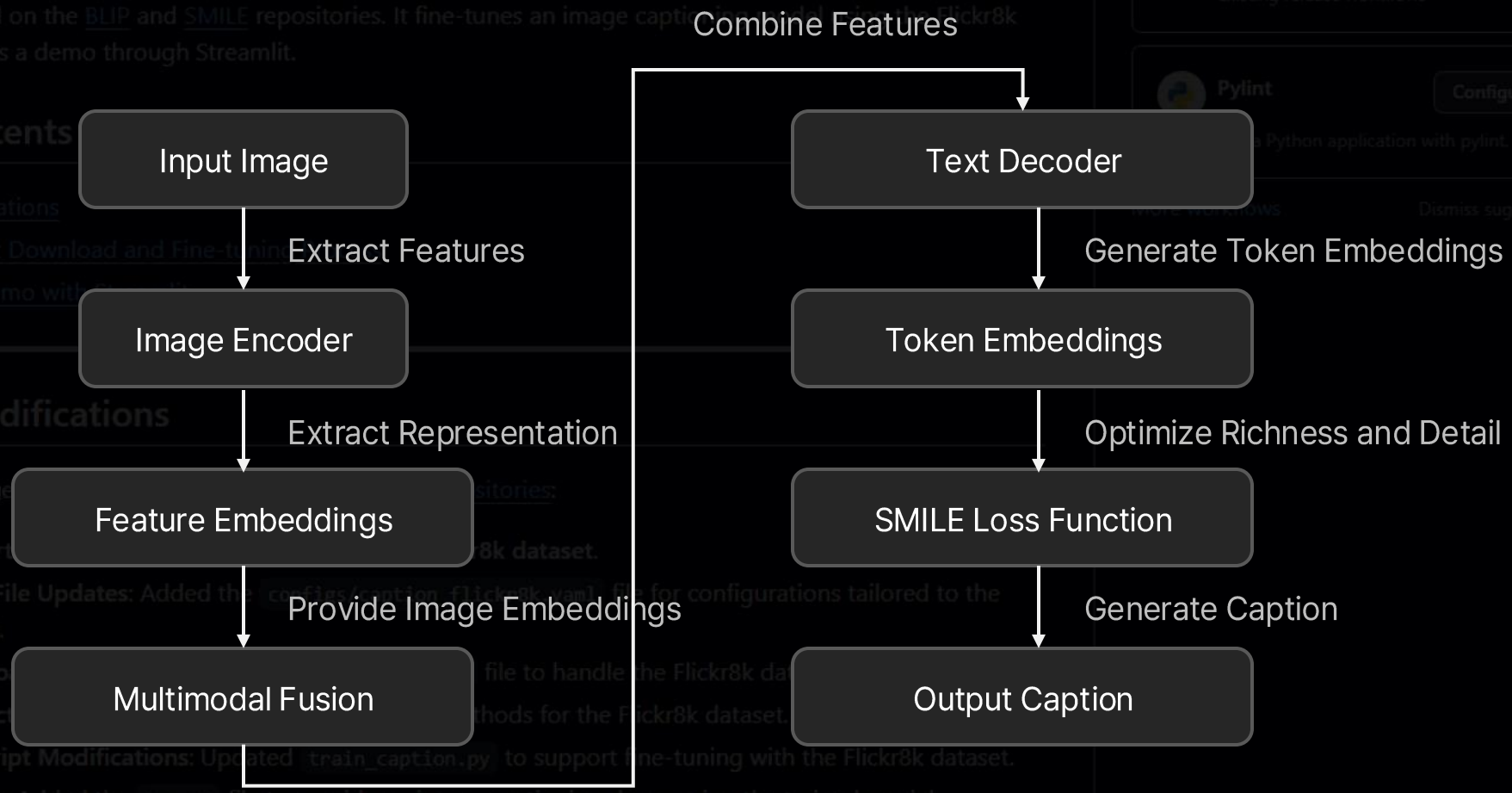
- BERT Tokenizer

Image Processing

- ViT(Vision Transformer)

https://github.com/heungkyulee/SMILE_Flickr8k

Model Architecture



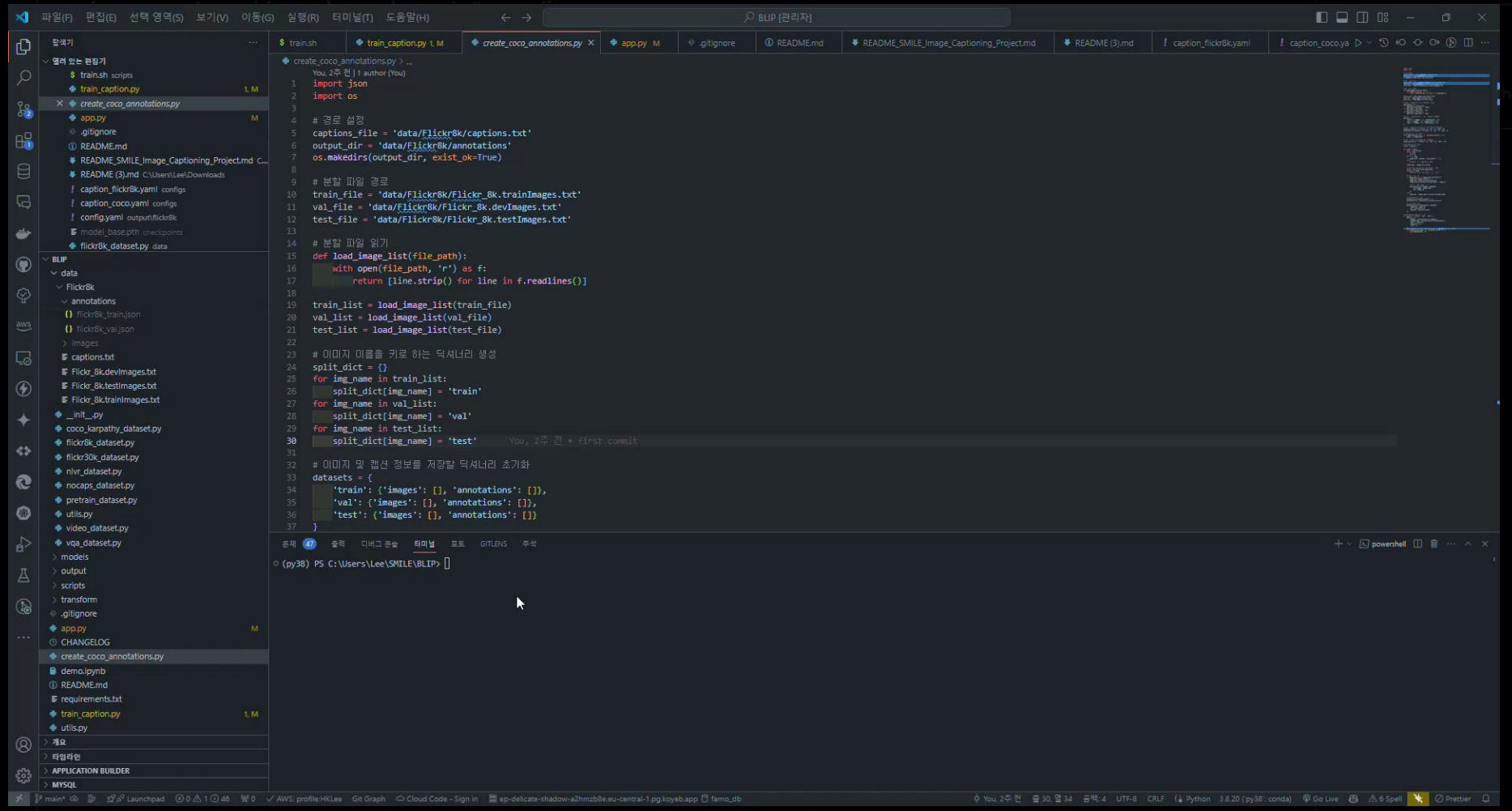
https://github.com/heungkyulee/SMILE_Flickr8k

Main Hyperparameters

Learning Rate	0.0001
Batch Size	16
Epochs	5
Max LR	1e-4
Min LR	1e-6

https://github.com/heungkyulee/SMILE_Flickr8k

Training (x60, 5 epochs, 12min/epoch)



https://github.com/heungkyulee/SMILE_Flickr8k

Fine-Tuning Result Analysis

Metric	Description	Base Model	
		blip_base	blip_smile_base
Caption Length (Cap.Len.)	The **average length** of **generated captions** in terms of word count	11.24	23.33
Lexical Diversity (Lex.Div.)	The **ratio of unique words** to total words in generated captions, reflecting **vocabulary richness**	0.06	4.33
Recall@1 (R@1)	The percentage of cases where the **correct caption** is **retrieved as the top result** in image-text matching.	10.18%	12.62%
Recall@5 (R@5)	The percentage of cases where the **correct caption** is among the **top 5 results** in image-text matching	28.75%	29.52%
ClipScore	The **semantic similarity score between images and captions** based on CLIP, expressed as a percent	30.27%	73.23%
Perplexity (PPL)	A measure of **language** model **fluency and predictaglity**; lower values indicate better performance	24792.13	110.34

README

SMILE-based Image Captioning Project

This project is based on the BLIP and SMILE repositories. It fine-tunes an image captioning model using the Flickr8k dataset and provides a demo through Streamlit.

Table of Contents

1. Project Modifications
2. Flickr8k Dataset Download and Fine-tuning Process
3. Running the Demo with Streamlit

Demo

1. Project Modifications

The following changes were made from the original GitHub repositories:

- **Dataset Support**: Replaced the COCO dataset with the **Flickr8k dataset**.
- **Configuration File Updates**: Added the `configs/caption_flickr8k.yaml` file for configurations tailored to the Flickr8k dataset.
- **Custom Data Loader**: Added the `data/flickr8k_dataset.py` file to handle the Flickr8k dataset.
- **Evaluation Function Modifications**: Adjusted evaluation methods for the Flickr8k dataset.
- **Fine-tuning Script Modifications**: Updated `train_caption.py` to support fine-tuning with the Flickr8k dataset.
- **Streamlit Demo**: Added the `app.py` file to provide an image captioning demo using the trained model.

2. Flickr8k Dataset Download and Fine-tuning Process

Publish a Python Package to PyPI on release

https://github.com/heungkyulee/SMILE_Flickr8k

SLSA Generic generator — Configure

Generate SLSA3 provenance for your existing release workflows

Pylint — Configure

Lint a Python application with pylint.

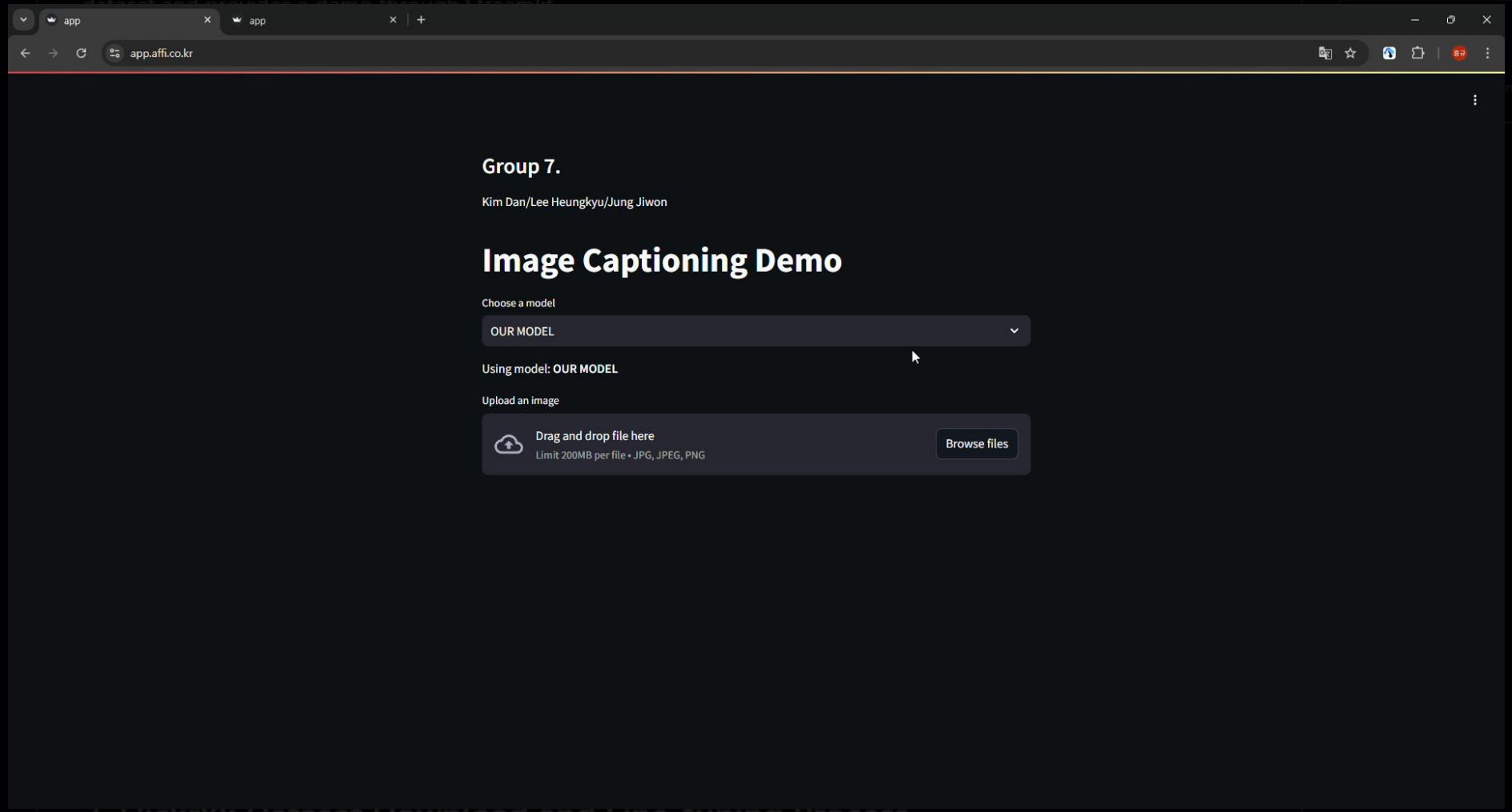
More workflows Dismiss suggestions

Demo QR Code

https://github.com/heungkyulee/SMILE_Flickr8k

Demo Video

https://github.com/heungkyulee/SMILE_Flickr8k



https://github.com/heungkyulee/SMILE_Flickr8k

Results

Familiar Images

SMILE(Ours)
an asian woman wearing a white tank top and holding up some fresh minty leaves to show off her natural beautyitangem

BLIP Base
an asian woman with blue eyes

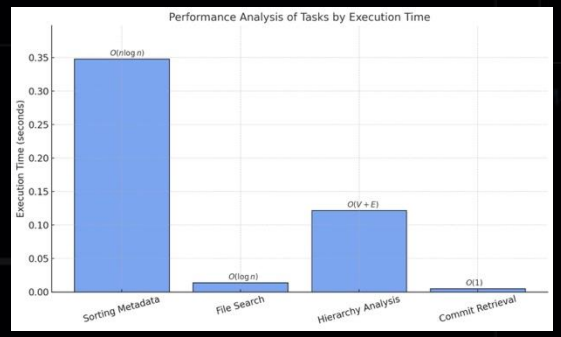
Capfilt Large
an asian woman with long brown hair and blue eyes

SMILE(Ours)
an orange, white and blue rocket surrounded by small stars that appear to be moving away from the vieweregale sourcerlingian

BLIP Base
a rocket ship in the desert

Capfilt Large
a rocket with planets in the background

Unfamiliar Images



SMILE(Ours)
a graph showing the percentage of performance in each race

BLIP Base
a graph showing the performance of tasks by execution time

Capfilt Large
a bar chart with the percentage of performance

SMILE(Ours)
an empty bottle with the label removed from it

BLIP Base
a bottle of ala body mist

Capfilt Large
a bottle of d'alba body oil

https://github.com/heungkyulee/SMILE_Flickr8k

Challenges & Solutions

Challenges	Solutions	Now
Project Issues • Dependency conflicts • Dataset mismatches • Configuration, preprocessing, evaluation, and GPU environment inconsistencies	**Fix Project Issues** • Matched dependencies • Custom COCO-style annotations compatible with Flickr8K • Customized settings and training code for single GPU environments	**Successful Project Result** • Successfully trained on Flickr8K dataset in a local computing environment
Prolonged Training Time • Training took over 4 hours using the cloned project setup	**Optimize Training Time** • Reduced image size(384 -> 224) • Increased workers(8 -> 16) • Switched to smaller dataset(MSCOCO -> Flickr8K)	**95% Reduced Training Time** • Reduced training time to 12 minutes per epoch
Slow Model Loading for Demo • Initial model loading without a server exceeded 10 minutes per client.	**Optimize Model Loading Time** • Applied port forwarding and caching techniques	**99% Reduced Model Loading Time** • Reduced initial model load time from 10 minutes to just a few seconds

https://github.com/heungkyulee/SMILE_Flickr8k

Limitations

Text Quality

- Descriptions are richer compared to BLIP and Base Models, but long texts degrade in quality toward the end.

Object and Dataset Limitations

- Poor performance when handling two or more objects or non-Flickr8K images.
- Limited by Flickr8K dataset and insufficient annotations

Demo Setup Constraints

- Unable to upload the model via Github or GLF.
- Relied on inefficient local port forwarding for demonstration.
- Future goal: Explore AWS Bedrock or similar services for better scalability.

Q&A

Group 7

Kim Dan / Lee Heungkyu / Jung Jiwon

https://github.com/heungkyulee/SMILE_Flickr8k