Φval

CSC 245 - C418

Junjie Zhao, Zachary Tan

The Problem and Our Goal

Motivation:

There is a high barrier of entry to the realm of professional photograph. Getting feedback from professional humans is expensive, time consuming, and can be subject to bias, especially when the time of professionals is so valued.

Goal:

We aim to develop an AI tool that can generate high quality and precise evaluations for photographs through the lens of a photographer.

Idea:

We use the BLIP (Bootstrapping Language Image Pre-training**)** pre-trained to provide feedback about an image through the lense of a photographer.

Related Work

VLP - Vision-language Pre-training

- A pre-trained model on large-scale image-text pairs
- Perform very well on downstream vision and language tasks
- Reference: Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

KD - Knowledge distillation

- A technique that transform "knowledge" from a large size model to a small size model
- Reference: Hinton, G., Vinyals, O., and Dean, J. the knowledge in a neural network. arXiv:1503.02531, 2015.

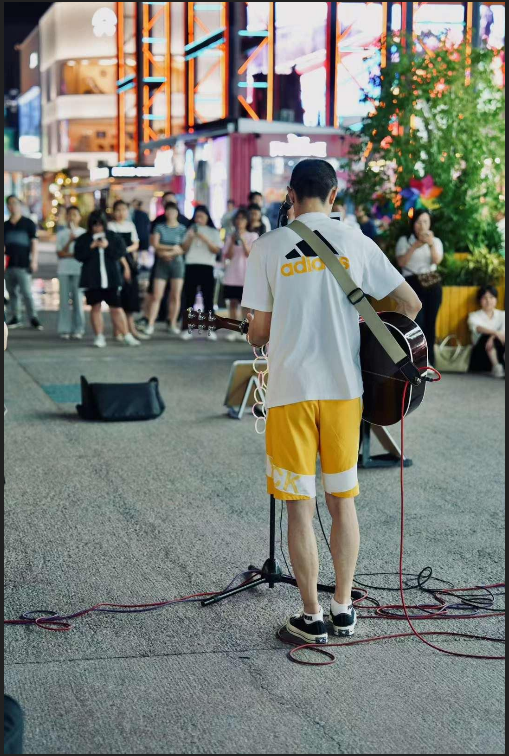
Application of BLIP

- Medical image analysis
- Visual question answering

Dataset

Train Dataset: 514 images (cleaned from 600 images)
Test Dataset: 40 images
Each image in our dataset has corresponding text in this format:

```
    {
        "image": "ID1.jpeg",
        "content": "The photograph captures an engaging street performance scene where
the musician is immersed in his     craft, with the audience in the background watching
attentively. This moment highlights the vibrancy and intimacy of urban life, creating a
genuine connection between the performer and the observers. The mood is both dynamic
and personal, reflecting the authenticity of public artistry.",
        "color": "The color palette in the image is vibrant and alive, with the bright
yellows of the musician's shorts standing out against the urban background filled with
glowing neon signs. The play of artificial and natural lighting creates a dynamic
harmony, making the scene feel warm and inviting while retaining the energy of the
street environment.",
        "composition": "The composition is thoughtfully structured, focusing on the
musician in the foreground while the slightly blurred audience creates a sense of depth
and context. The framing effectively draws the eye toward the central subject while
allowing the surrounding elements to add storytelling layers. The balance between the
performer and the audience creates a compelling narrative.",
        "quality": "The quality of the photograph is impressive, with sharp details on
the musician and well-managed exposure that captures the intricate light contrasts of
the night scene. There is minimal noise, and the image retains clarity, making it
visually pleasing. The technical precision complements the emotional and narrative
depth, resulting in a polished and professional image."
    },
```



Method

Model

- Default Model without Fine Tuned
- Default Model with Fine Tuned
- More Visual Information Model with Fine Tuned
- New Architecture Model with Fine Tuned

Loss Function

- Cross-Entropy Loss

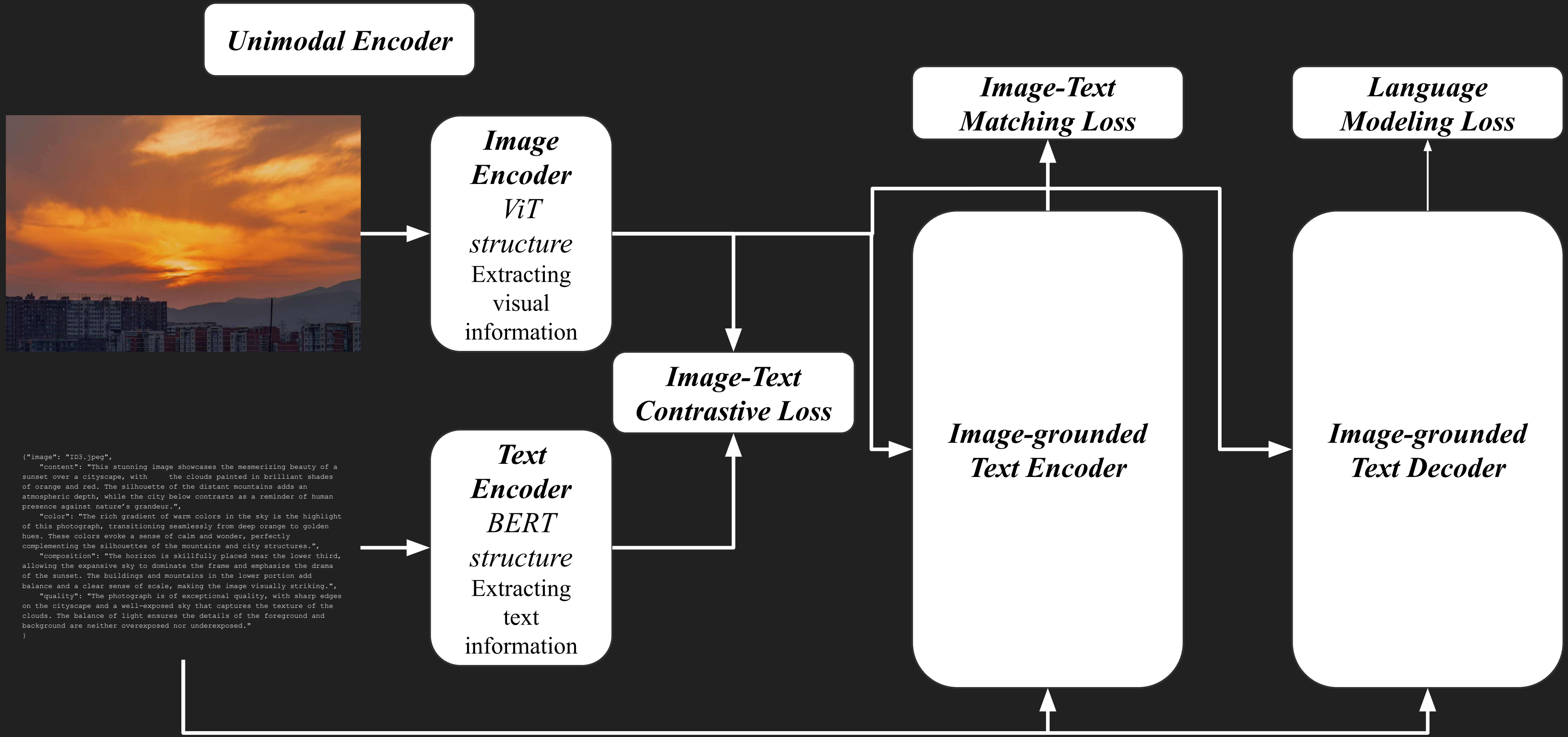
Optimizer

- AdamW
- We designed a dynamic learning rate. The learning rate started at 1e-5 and decay by 10% for every 10 epochs
- Epochs
    - After each epoch, the program will save and update the model named last.pth. If the epoch loss has reached a new minimum, the program will save and update the model named best.pth
    - We implemented early termination: If the epoch loss does not decrease by over 1e-3 for 10 epochs, the program will automatically terminate.

Evaluation

- Human Evaluation
- BLEU and ROUGE

A Brief Outline of The Architecture of BLIP



Unimodal Encoder

Image Encoder
ViT structure
Extracting visual information

Text Encoder
BERT structure
Extracting text information

Image-Text Contrastive Loss

Image-Text Matching Loss

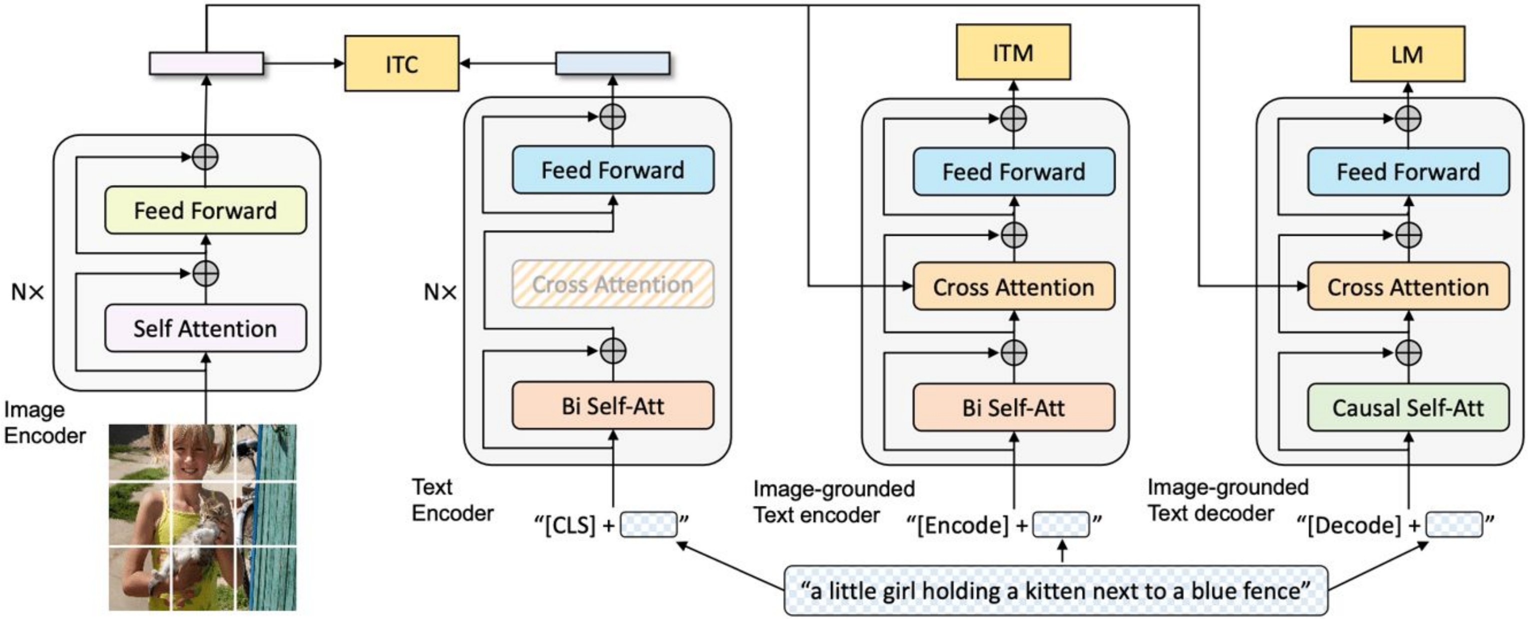
Language Modeling Loss

Image-grounded Text Encoder

Image-grounded Text Decoder

```
{"image": "ID3.jpeg",
    "content": "This stunning image showcases the mesmerizing beauty of a
sunset over a cityscape, with     the clouds painted in brilliant shades
of orange and red. The silhouette of the distant mountains adds an
atmospheric depth, while the city below contrasts as a reminder of human
presence against nature’s grandeur.",
    "color": "The rich gradient of warm colors in the sky is the highlight
of this photograph, transitioning seamlessly from deep orange to golden
hues. These colors evoke a sense of calm and wonder, perfectly
complementing the silhouettes of the mountains and city structures.",
    "composition": "The horizon is skillfully placed near the lower third,
allowing the expansive sky to dominate the frame and emphasize the drama
of the sunset. The buildings and mountains in the lower portion add
balance and a clear sense of scale, making the image visually striking.",
    "quality": "The photograph is of exceptional quality, with sharp edges
on the cityscape and a well-exposed sky that captures the texture of the
clouds. The balance of light ensures the details of the foreground and
background are neither overexposed nor underexposed."
}
```

A Brief Outline of The Architecture of BLIP



Different Model Iterations

Default Model without Fine Tuning

- Load default pre-trained model named “blip-image-captioning-base"
- Do generation on images from test dataset

New Architecture Model with Fine Tuned

- Load default pre-trained mode named “blip-image-captioning-base"
- Change the image size parameter from 224 x 224 to 384 x 384 in the image encoding part and give the model about 3 times more visual information
- Change the architecture of the model by adding four separate decoder layer for each of the evaluation criteria, aiming to improve them over all generation performance
- Do fined tuning using the image and text pair from train dataset
- Do generation on images from test dataset

Default Model with Fine Tuned

- Load default pre-trained mode named “blip-image-captioning-base"
- Do fined tuning using the image and text pair from train dataset
- The section heads “content”, “color”, “composition”, and “quality” are concatenated together and tokenized as an input
- Do generation on images from test dataset

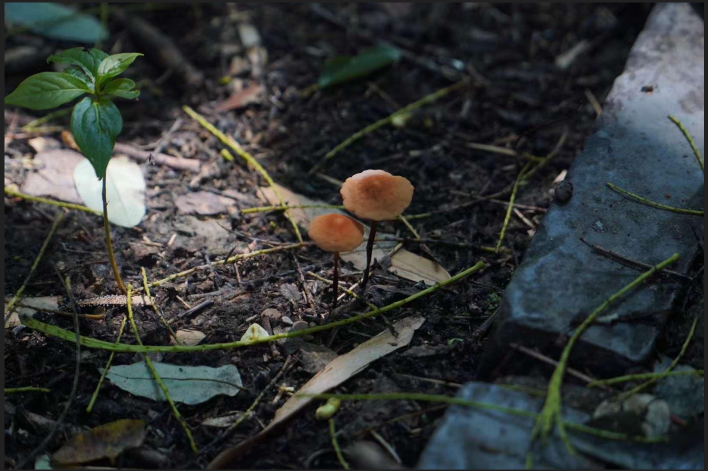
More Visual Information Model with Fine Tuned

- Load default model named “blip-image-captioning-base"
- Change the image size parameter form 224 x 224 to 384 x 384 in the image encoding part and give the model about 3 times more visual information
- Do fine tuning using the image and text pair from train dataset
- The section heads “content”, “color”, “composition”, and “quality” are concatenated together and tokenized as an input
- Do generation on images from test dataset

Results

Chat-GPT

Prompt: please evaluate this picture

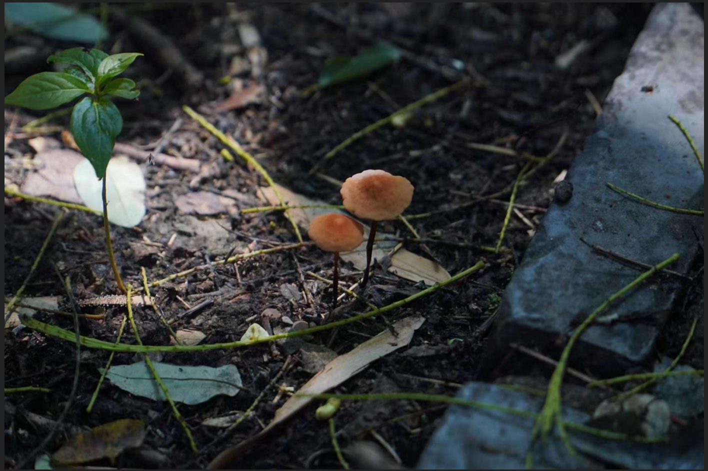


This image depicts a natural scene focused on small mushrooms growing in soil near a paved edge. Here’s an evaluation of the image:

1. **Composition**: The photo has a good balance between the natural elements (mushrooms, soil, and leaves) and the man-made edge (paved brick). The mushrooms serve as the focal point of the image, with light effectively drawing attention to them.
2. **Lighting**: The lighting is soft and natural, with a spotlight effect on the mushrooms. This highlights the main subject while leaving the surrounding environment slightly darker, adding depth and contrast.
3. **Focus and Clarity**: The mushrooms appear to be in focus, while the background has a subtle blur, which helps isolate the subject and create a sense of depth.
4. **Details**: The soil texture, surrounding leaves, and small branches provide context and detail to the environment. However, the image could benefit from slight sharpening to enhance the intricate details of the mushrooms.
5. **Framing**: The inclusion of the paved edge on the right adds structure to the composition. However, the frame might feel a bit tight; more negative space around the mushrooms could make the scene feel less constrained.

Results

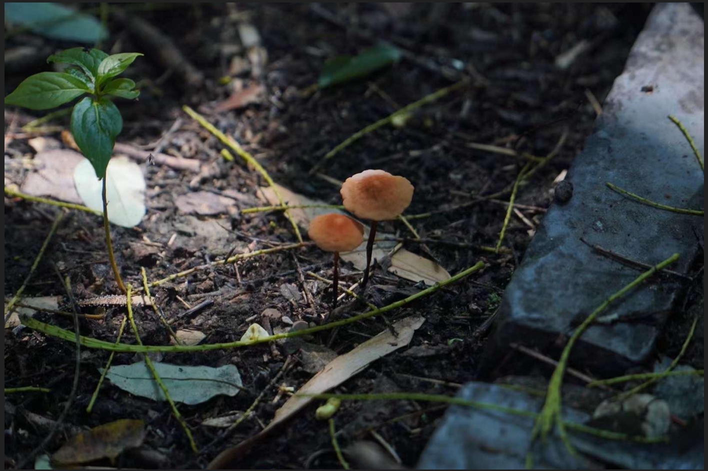
Default Model without Fine Tuned



a small mushroom sitting on the ground.

Results

Default Model with Fine Tuned



content : this delightful photograph features a small bird perched on a rock, its delicate posture juxtaposed against a lush green leaf. the intimate focus on the bird ’ s environment and its delicate gaze create a sense of connection with nature, making it both engaging and endearing.

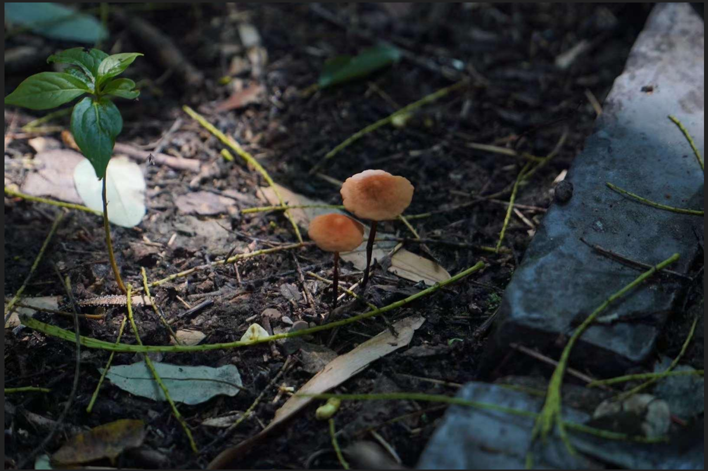
color : the image features a natural and muted hues of the bird ’ s orange against the earthy tones of the tree leaf. the soft glow of the light of the bird ’ s head adds a subtle yet eye, enhancing the overall appeal of the scene.

composition : the composition places the bird on a slightly off - center, creating a dynamic composition that draws the viewer ’ s attention to its environment. the surrounding elements create a natural and harmonious atmosphere, enhancing the sense of connection between the bird and its environment.

quality : the quality of the image is excellent, with clear details in the bird ’ s fur and the surrounding environment. the lighting is soft and even, highlighting the delicate glow of the bird ’ s feathers while the surrounding environment adds a touch of tranquility..

Results

Model with More Visual Information and Fine Tuning



content : this image captures a tranquil forest setting, focusing on small mushrooms amidst a forested environment. the juxtaposition of natural life and man - made structures evokes a sense of coexistence and human development, inviting the viewer to reflect a quiet coexistence of nature.

color : the warm orange of the mushrooms against the dark tones of the forest floor and the muted greens of the surrounding trees. the subtle glow of the mushrooms adds a fresh and natural color palette, enhancing the overall visual.

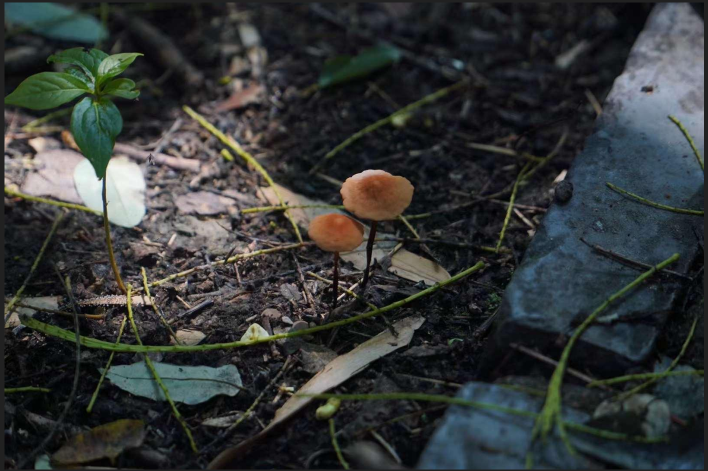
composition : the mushrooms are positioned slightly off - center, drawing the viewer ’ s eye into the scene while allowing the surrounding elements to frame the scene.

quality : the quality of the image is excellent, with sharp details that highlight the textures of the mushrooms and leaves.
the lighting is natural and even, ensuring the mushrooms are well - balanced and well - balanced, making the scene visually captivating..

Results

New Architecture with Fine Tuning

```
predicted_ids = top_k_sampling(logits, top_k=30, max_length=60)
```



content : this image photograph a a an ofling a a the the by perched natural of a lush sense. the, of the and backdrop of an traditional ju the textures into -. the,, moment of, and the its sense draws the water beauty.

color : the and features of tones of image dominate golden the the complement vibrant ’ and green of the and hue blurred, muted of complement the of. muted enhance. enhance the in of a adds that lighting the balance vi atmosphere overall depth. enhance the atmosphere

composition : the composition placement balancedlys cat,, drawing the a viewer, diagonal and the to framing subject adding. background its drawing elements to fore the creates reflection sense the to the sense from point thats engaging balance. and activity toward the both

quality : the quality of is sharp, high, impressive, intricate on cat textures s and sharply and in slightly the fur the s ensuring - on balanced well colors is ofs the even fields leaves subject the resulting professional the the a scene to to

Results

Default Model without Fine Tuned

BLEU: 0.000	r	p	f
rouge-1	0.543	0.998	0.699
rouge-2	0.110	0.880	0.195
rouge-1	0.531	0.978	0.684

Default Model with Fine Tuned

BLEU: 0.0184	r	p	f
rouge-1	0.900	0.920	0..909
rouge-2	0.692	0.757	0.719
rouge-1	0.856	0.876	0.865

BLEU Scores:
1-gram: 0.1172
2-gram: 0.0308
3-gram: 0.0099
4-gram: 0.0032

Model with More Visual Information and Fine Tuning

BLEU: 0.0157	r	p	f
rouge-1	0.931	0.844	0.883
rouge-2	0.789	0.476	0.584
rouge-1	0.891	0.806	0.844

BLEU Scores:
1-gram: 0.1171
2-gram: 0.0331
3-gram: 0.0087
4-gram: 0.0026

New Architecture with Fine Tuning

BLEU: 0.0029	r	p	f
rouge-1	0.875	0.897	0.885
rouge-2	0.555	0.737	0.632
rouge-1	0.823	0.844	0.832

BLEU Scores:
1-gram: 0.3558
2-gram: 0.0378
3-gram: 0.0009
4-gram: 0.0000

BLEU Score	Default Model with Fine-tuning	More Visual Information Model with Fine-tuning	New Architecture Model with Fine-tuning
1-gram	0.1172	0.1171	0.3558
2-gram	0.0308	0.0331	0.0378
3-gram	0.0099	0.0087	0.0009
4-gram	0.0032	0.0026	0.0000

Default Model with Fine-tuning

	r	p	f
rouge-1	0.900	0.920	0..909
rouge-2	0.692	0.757	0.719
rouge-1	0.856	0.876	0.865

More Visual Information Model with Fine-tuning

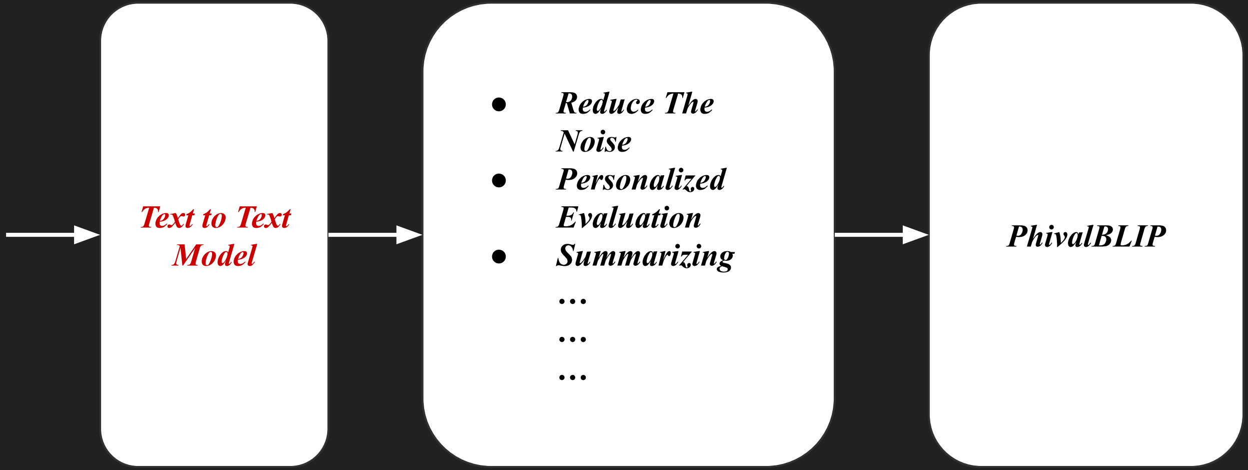
	r	p	f
rouge-1	0.931	0.844	0.883
rouge-2	0.789	0.476	0.584
rouge-1	0.891	0.806	0.844

New Architecture Model with Fine-tuning

	r	p	f
rouge-1	0.875	0.897	0.885
rouge-2	0.555	0.737	0.632
rouge-1	0.823	0.844	0.832

Ideas For Future Improvements

```
{"image": "ID3.jpeg",
    "content": "This stunning image showcases the
mesmerizing beauty of a sunset over a cityscape, with
the clouds painted in brilliant shades of orange and
red. The silhouette of the distant mountains adds an
atmospheric depth, while the city below contrasts as a
reminder of human presence against nature's grandeur.",
    "color": "The rich gradient of warm colors in the
sky is the highlight of this photograph, transitioning
seamlessly from deep orange to golden hues. These colors
evoke a sense of calm and wonder, perfectly
complementing the silhouettes of the mountains and city
structures.",
    "composition": "The horizon is skillfully placed
near the lower third, allowing the expansive sky to
dominate the frame and emphasize the drama of the
sunset. The buildings and mountains in the lower portion
add balance and a clear sense of scale, making the image
visually striking.",
    "quality": "The photograph is of exceptional
quality, with sharp edges on the cityscape and a
well-exposed sky that captures the texture of the
clouds. The balance of light ensures the details of the
foreground and background are neither overexposed nor
underexposed."
}
```



Text to Text Model

- ***Reduce The Noise***
- ***Personalized Evaluation***
- ***Summarizing***

…
…
…

PhivalBLIP

Conclusion

- Phival is a more specialized and smaller than Chat-GPT
- It demonstrates that image recognition models can be used for more specialized tasks
- A more general model (1 head) is sometimes better than dividing the model to deal with specialized tasks (4 heads)
- Model with more visual informance (greater image resize shape) will have better performance but consume much more computing resources
- It shows how BLIP can be used to generate target text output with a certain format
  - Can be upscaled, and increase accuracy with a better dataset

Potential issues

- The dataset is not very diverse and is skewed towards images of certain types such as environments, food, and animals
- It is very time consuming to create the data set