YZV 211E HW3 REPORT

Salih Demirözü 150230331

December 29, 2025

1 Introduction

This report presents a comparison of the BLIP model's results. The BLIP model is capable of producing captions for uploaded images with different parameters. The related parameters are temperature, max-length, min-length and number of variations.

2 Effect of Temperature

Figure 1: The image used

Temperature	Captions Generated
$T = 0.1$	a small cabin in the mountains
$T = 0.8$	a small cabin sits in the middle of a field and looks down
$T = 1.5$	an old hut sits alone amongst field of flowers near klapa in romania wall by assani

Figure 2: Outputs for different temperatures

As demonstrated in the table, increasing the temperature also increases the unpredictability or creativeness. Because temperature represents the degree of unpredictability of outcome. When T is equal to 1.5, the outcome is really detailed and more creative. However, for the case of being T is equal to 0.1, the caption seems like a simple daily expression. Higher temperatures are not always beneficial; as the chart illustrates, the results are absent and do not correspond well with the image.

3 Length Constraint Analysis

Figure 3: The image used

In this experiment, a simple image (Blue Sky) was used to observe the model's behavior when forced to generate a minimum length of 10 tokens. Three temperature (T) settings were tested to observe the variance in creativity and accuracy.

3.1 Experimental Results

Case 1: Low Temperature ($T = 0.1$)
Settings: 3 Captions, Min Length = 10

- a blue sky with white clouds and a few white clouds
- a clear blue sky with a few clouds in the distance
- a clear blue sky with a few clouds in it

Case 2: Medium Temperature ($T = 0.8$)
Settings: 4 Captions, Min Length = 10

- a blue sky with white clouds over hills in the background
- a clear blue sky and white clouds over a grassy hillside
- green pine trees in the background stand in the fore of the sky
- clouds in the sky above the ocean is clear blue

Case 3: High Temperature ($T = 1.5$)
Settings: 5 Captions, Min Length = 10

- very blue sky above beach ocean and lighthouse on beach
- thin clouds covered by a clear blue sky can be viewed through
- some white fluffy clouds in a blue sky,
- a lone white bench overlooking a city during the daytime
- a bright blue sky covered with some yellow patches of smothuring and clouds in an otherwise

3.2 Discussion of Outcomes

The results highlight a trade-off between redundancy and hallucination under length constraints. At **low temperature** ($T = 0.1$), the model satisfies the length requirement through repetition, lacking the creativity to generate new details.
As temperature increases to **medium** ($T = 0.8$), the model begins inferring contextually probable but absent features (e.g., hills, oceans) based on training priors.
Finally, at **high temperature** ($T = 1.5$), the model exhibits severe hallucinations and grammatical instability, fabricating entirely missing objects (e.g., "lighthouse," "city") to fulfill the token count.

4 Conditional Image Captioning Analysis

In this experiment, specific starting text was provided to guide the model's generation process.

- **Input:** "A dog holding"
  Output: A dog holding a large stick to his mouth.
- **Input:** "A golden"
  Output: A golden retrieve dog is holding a bamboo stick.
- **Input:** "A yellow"
  Output: A yellow puppy standing on the ground with a leaf in its mouth.

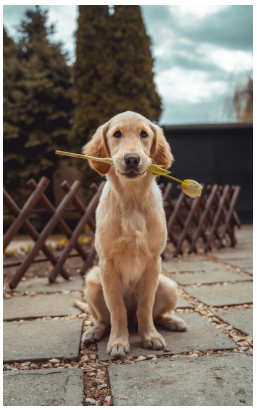

Figure 4: The image used is a puppy holding a flower

4.1 Observations on Conditional Generation

The results reveal a consistent failure mode known as *semantic hallucination*, where the model relies on statistical likelihood rather than visual evidence.

1. **Shape Bias:** The model focused on the green stem while ignoring the petals.
2. **Dataset Bias:** The model favored statistically probable associations (dogs holding sticks) over the actual visual evidence.

Additionally, splitting the token "Golden" caused a minor grammatical degradation ("retrieve dog"), showing that partial prompts can disrupt internal language representations.

Conclusion: Conditional prompting guided the sentence structure but failed to correct the underlying visual misinterpretation.