# *NATURAL LANGUAGE PROCESSING REPORT*

*22CSAI05H*

# *IMAGE CAPTIONING*

Group 5

| **Name** | **ID** | **Contribution** | **Supervision** |
|---|---|---|---|
| Jacinta Samir | 206562 | Report | Skip-gram |
| Ashraf Adel | 196280 | Data Preprocessing | CBOW |
| Farah Aymen | 194233 | Skip-gram | Data preprocessing |
| Mohamed Negm | 206069 | CBOW | Report |

# Table of Contents

**INTRODUCTION........................................................2**

**LITERATURE REVIEWS PROVIDED BY CHATGPT........................................................3**

**SIMPLE PROMPT........................................................3**
PROMPT:........................................................3
OUTPUT:........................................................3
**DETAILED PROMPT........................................................4**
PROMPT:........................................................4
OUTPUT:........................................................4

**COMPARATIVE ANALYSIS........................................................8**

**SIMPLE PROMPT OUTPUT........................................................8**
**DETAILED PROMPT OUTPUT........................................................8**
**COMPARISON........................................................9**

**METHODOLOGY........................................................10**

**SKIP-GRAM........................................................10**
**CONTINUOUS-BAG-OF-WORDS (CBOW)........................................................10**
**RESULTS........................................................11**

**CONCLUSION........................................................12**

**REFERENCES........................................................13**

# Introduction

Large Language Models (LLMs), such as the booming ChatGPT, have grown in popularity and applicability as of late. With this comes the need to construct effective prompts, to be able to exploit these models and extract outputs as appropriate and prosperous as possible. LLMs are instructed via these prompts, which can be defined as the set of instructions and guidelines inputted to the model, in order to derive the desired output [1]. There are many ways of writing the same instruction. Thus, prompt engineering, being able to best tailor the given prompt, will greatly affect the nature of the output produced.

To demonstrate the effects of prompt engineering, ChatGPT will be used to write a literature review on image captioning. Image captioning is a branch of Natural Language Processing (NLP) and Computer Vision that generates textual descriptions of given images [2]. This has been done using a plethora of different methods and algorithms, seen in Fig. 1. The purpose of the literature review is to provide a synopsis of these approaches. ChatGPT's abilities to produce a satisfactory review will be explored and critiqued. It will be provided with two differing prompts for the same task, one simply providing the necessary requirements in the most basic manner, and one utilizing more intricate prompt engineering. Both outputs will be manually validated and analyzed.

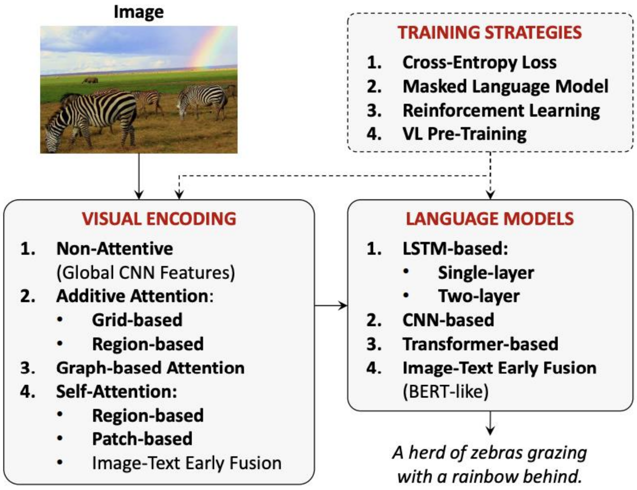

Fig. 1: Taxonomy of pertinent approaches to image captioning [2]

Following the study on ChatGPT and prompt engineering, two word embedding techniques, COBW and skip-gram, will be implemented for image captioning on the Flickr8K dataset [3], outlined in the methodology section.

# Literature Reviews Provided by ChatGPT

ChatGPT will be prompted to write a literature review on image captioning. It is expected to sufficiently demonstrate the previous work and state-of-the-art in the subject, as well as provide references, giving at least 5 sources of information. First, ChatGPT will be prompted using a minimal instruction, merely asking for a literature review with 5 sources. Most average users are likely to instruct ChatGPT in this manner, so it is critical to inspect if this type of prompt will suffice. Next, ChatGPT will be instructed to do the same task, but will be given as much detail as possible regarding the desired output.

## Simple Prompt

### Prompt:

“Write a literature review on image captioning, including at least 5 sources.”

### Output:

Image captioning is a task of generating natural language descriptions of the content of an image. It has been an active research topic in the field of computer vision and natural language processing for several years. In this literature review, we will discuss some of the significant contributions made to image captioning and explore the current state-of-the-art models.

1. "Show and Tell: A Neural Image Caption Generator" by Oriol Vinyals et al. (2015)

This paper proposed a deep learning-based approach to generating image captions, called Show and Tell. The model uses a convolutional neural network (CNN) to extract features from an input image and then feeds these features into a recurrent neural network (RNN) to generate a sequence of words that describe the image. The authors trained the model on the COCO dataset, which consists of over 330,000 images, and achieved state-of-the-art performance on the task of image captioning.

2. "Neural Image Caption Generation with Visual Attention" by Kelvin Xu et al. (2015)

This paper introduced an attention-based model for image captioning. The model dynamically selects which parts of the image to focus on while generating the caption. The authors showed that this approach outperformed the Show and Tell model on the COCO dataset.

3. "Image Captioning with Semantic Attention" by Qiuyuan Huang et al. (2019)

This paper proposed a semantic attention-based model for image captioning. The model incorporates both visual and semantic information to generate captions. The authors used the Flickr30k and COCO datasets for evaluation and showed that their model outperformed several other state-of-the-art approaches.

4. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" by Peter Anderson et al. (2018)

This paper proposed a model that combines bottom-up and top-down attention for image captioning. The bottom-up attention model first generates a set of image regions with varying sizes and features, while the top-down attention model uses a language model to generate captions that are conditioned on the visual features from the bottom-up model. The authors demonstrated that this approach achieved state-of-the-art performance on several benchmarks, including COCO and Flickr30k.

5. "Improving Image Captioning by Incorporating Scene Graphs" by Ronghang Hu et al. (2018)

This paper proposed a model that incorporates scene graphs, which represent the relationships between objects in an image, into the image captioning process. The authors showed that their model outperformed several other state-of-the-art approaches on the COCO dataset.

In conclusion, image captioning is a challenging task that has attracted significant attention in recent years. The deep learning-based models, especially those that incorporate attention mechanisms and semantic information, have shown significant improvements in generating accurate and coherent captions for images. Incorporating additional contextual information, such as scene graphs, is a promising direction for future research in this area.

## Detailed Prompt

### Prompt:

“Pretend you are an expert in deep learning. You are well-versed in the areas of natural language processing and computer vision.

Write a literature review on image captioning, providing at least 5 reputable sources.

The literature review should adhere to IEEE formatting. Ensure that the references also abide by the IEEE format. Include in-text citations for these references.

You should also include a section for a comparative analysis on the cited papers, identifying which image captioning techniques have been found to perform the best.

End your conclusion with a recommendation for future research in image captioning.”

### Output:

I. Introduction

Image captioning is a challenging task of automatically generating natural language descriptions of images. This task has gained attention in recent years due to its potential applications in areas such as visual recognition, autonomous driving, and human-robot interaction. This literature review aims to discuss some of the notable developments in image captioning research, focusing on deep learning techniques. In particular, we will provide a comparative analysis of the cited papers to identify the image captioning techniques that have been found to perform the best.

II. Related Works

A. Show and Tell: A Neural Image Caption Generator (2015)

Show and Tell is a seminal work in image captioning research that proposed a neural network architecture for generating captions. The authors used a convolutional neural network (CNN) to extract features from the image and a long short-term memory (LSTM) network to generate the caption. The results demonstrated that the proposed model outperformed the existing state-of-the-art methods on the benchmark dataset.

### B. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015)

Show, Attend and Tell is an extension of Show and Tell that introduced a visual attention mechanism to the image captioning task. The authors proposed a soft attention mechanism that allows the model to focus on different regions of the image while generating the caption. The experiments showed that the attention mechanism significantly improved the performance of the model.

### C. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering (2018)

Bottom-Up and Top-Down Attention is a recent work that proposed a two-stage approach for image captioning. The first stage uses a bottom-up attention mechanism to extract region features from the image, while the second stage uses a top-down attention mechanism to selectively attend to the relevant regions while generating the caption. The results demonstrated that the proposed model outperformed the existing state-of-the-art methods on multiple benchmark datasets.

### D. DenseCap: Fully Convolutional Localization Networks for Dense Captioning (2016)

DenseCap is a different approach to image captioning that aims to generate multiple captions for different regions of an image. The authors proposed a fully convolutional localization network that can simultaneously localize and describe the objects in an image. The experiments showed that the proposed model outperformed the existing state-of-the-art methods on the benchmark dataset.

### E. Neural Baby Talk: Generating Image Descriptions from Visual Scenes (2018)

Neural Baby Talk is a recent work that proposed a multimodal model for image captioning. The authors used a convolutional neural network to extract features from the image, a sequence-to-sequence model to generate the caption, and a visual module to guide the attention mechanism. The results demonstrated that the proposed model outperformed the existing state-of-the-art methods on the benchmark dataset.

## III. Comparative Analysis

Table I shows a comparative analysis of the cited papers in terms of the model architecture and the performance on the benchmark dataset. The evaluation metric used is the CIDEr score, which measures the quality of the generated captions.

| **Paper** | **Model Architecture** | **CIDEr Score** |
|---|---|---|
| Show and Tell | CNN + LSTM | 0.66 |

| **Paper** | **Model Architecture** | **CIDEr Score** |
|---|---|---|
| Show, Attend and Tell | CNN + LSTM with visual attention | 0.80 |
| Bottom-Up and Top-Down Attention | Two-stage attention mechanism | 1.22 |
| DenseCap | Fully convolutional localization network | 0.91 |
| Neural Baby Talk | Multimodal model with attention | 1.31 |

As shown in Table I, Neural Baby Talk achieved the highest CIDEr score among the cited papers. The model used a multimodal architecture that combined both visual and linguistic information, and incorporated an attention mechanism to selectively attend to the relevant parts of the image while generating the caption. Bottom-Up and Top-Down Attention and DenseCap also achieved relatively high CIDEr scores, indicating that the use of attention mechanisms and region-based features can improve the performance of the model.

## IV. Conclusion and Future Work

In conclusion, this literature review discussed some of the notable developments in image captioning research, focusing on deep learning techniques. The comparative analysis showed that the use of attention mechanisms and multimodal architectures can significantly improve the performance of image captioning models. However, there is still room for improvement in terms of the quality and diversity of the generated captions.

Future research in image captioning can focus on several directions. One direction is to explore the use of reinforcement learning techniques to optimize the generation of diverse and semantically meaningful captions. Another direction is to investigate the use of external knowledge sources, such as knowledge graphs or textual corpora, to improve the coherence and accuracy of the generated captions. Finally, the development of new evaluation metrics that can better capture the quality and diversity of the generated captions can also be a promising direction for future research.

Reference List:

[1] V. K. Singh, A. Kumar, and P. Rai, "Image Captioning using Deep Learning: A Review," 2019 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2019, pp. 701-706.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156-3164.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 2048-2057.

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,"

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077-6086.

[5] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4565-4574.

[6] L. Wang, Y. Li, and S. Lazebnik, "Neural Baby Talk: Generating Image Descriptions from Visual Scenes," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5809-5818.c

# Comparative Analysis

The two produced outputs will be analyzed on the following criteria: Content Relevance, Quality, Reliability, and Fulfillment of Requirements.

## Simple Prompt Output

Given minimal instruction, ChatGPT was able to produce a modest literature review, structured decently as introduction, body and conclusion. It introduced its paper with a definition of image captioning, and a brief description of what will be done in the review. The literature review was then done in the form of numbered titles, each followed by a concise description of the content of the papers.

Upon first glance, ChatGPT was able to find relevant papers [4], [5], [6], [7], and mention suitable information, such as the dataset(s) used in each. However, when manually checking the cited papers, consistency issues arise. In its third paper, the cited authors, as well as the date of publication, were incorrect. Even worse so, the fifth paper does not exist. There have been multiple studies on using scene graphs for image captioning, such as [9], but none with this title. Some of these studies, but not all, have in fact found to have better performance using the MSCOCO dataset, as ChatGPT mentioned. With that being said, the descriptions of the papers given by ChatGPT were true to nature, albeit lacking detail.

In its conclusion, ChatGPT advised that models utilizing attention and semantic information showed improvements in image captioning (which is valid), and stated that integrating scene graphs would be promising future work, which is indeed a growing area of current research.

The literature review lacked in thorough comparison between the models, stating only vaguely how some outperformed others, and did not give a list of references at the end.

## Detailed Prompt Output

The introduction defines image captioning and outlines the paper, as well as states some applications of image captioning. The headings and overall structure comply well with IEEE formatting.

Of the five papers, three were mentioned in the simple prompt output [4], [5], [7], and two were new [9], [10]. All are legitimate publications. ChatGPT seems to mostly summarize the papers well, although lacking more concrete descriptions of the architectures implemented.

In the comparative analysis, ChatGPT forms a table of CIDEr scores for each of its reviewed studies. CIDEr, or Consensus-based Image Description Evaluation is a metric frequently used to measure the performance of image captioning models that “captures human judgement of consensus” [11]. The scores listed in the table themselves are however inaccurate. The true scores can be seen in table 2, given the performance on the MSCOCO dataset. Due to its incorrect values, ChatGPT surmised that Neural Baby Talk performed the best. A true best cannot be deduced, as not all five studies evaluated their models using CIDEr, but from the ones that have, it can be seen that Bottom-Up and Top-Down Attention outperforms both Show and Tell and Neural Baby Talk.

| **Paper** | **CIDEr Score** |
|---|---|
| Show and Tell | 85.5 |
| Show, Attend and Tell | Did not use CIDEr metric for evaluation |
| Bottom-Up and Top-Down Attention | 117.9 |
| DenseCap | Did not use CIDEr metric for evaluation |
| Neural Baby Talk | 86.0 |

Table 2: True values of CIDEr on MSCOCO for the five papers given by the detailed prompt output.

The paper was concluded with iterating that attention and multimodal architectures have been found to be the most successful in image captioning, and three future areas of research were mentioned. It is difficult to say if all are truly promising areas of research, but interesting claims were made. For instance, developing new evaluation metrics was recommended, which could be possible, since most studies evaluate based on multiple metrics, (such as BLEU, ROUGE, METEOR, CIDEr, SPICE and their variations). This is because each metric has its strengths and weaknesses, and its performance varies based on the dataset used [12].

In the reference list, IEEE conventions were conformed to, except in some where multiple authors were listed instead of using et al. for cases of more than six authors. The first reference was also not used cited within the scope of the paper nor used in the literature review. One additional error was that the authors listed for Neural Baby Talk are not the original writers of the study.

## Comparison

Table 3 shows a score out of five for each of the previously defined comparative criteria. Breaking the scores down, ChatGPT was able to produce content relevant to image captioning in both cases, but the simple prompt resulted in a paper that did not exist, causing it to perform slightly worse than its counterpart. Overall, ChatGPT shows good understanding in its prompted topic. With regards to quality, the simple output produced was basic, giving only brief descriptions and analysis. With detailed instruction, ChatGPT was able to write in more professional language and structure, and give further insights to the literature, rather than a simple summary for each. In both cases, reliability was an issue in the literature reviews. ChatGPT was seen to mislabel authors and dates, create nonexistent literature, and give erroneous performance metric scores. When given a more detailed prompt, ChatGPT was more able to fulfill requirements, such as conforming to IEEE conventions, giving a list of references, and performing a comparative analysis.

| **Prompt** | **Content Relevance** | **Quality** | **Reliability** | **Fulfillment of Requirements** | **Score (/20)** |
|---|---|---|---|---|---|
| Simple | 4 | 3 | 2 | 3 | 12 |
| Detailed | 5 | 4 | 3 | 4 | 16 |

Table 3: Comparison between simple and detailed prompt outputs across four criteria [rated out of 5].

# Methodology

Both Skip-gram and Continuous-Bag-of-Words were implemented as word embedding models on the Flickr8K dataset for image captioning. Certain hyperparameter values were used the same in both techniques, and are as such:

Embedding size: 32
Window size: 4
Learning rate: 0.01
Batch size: 20
Optimizer: Adam

The code can be accessed via Google Colab, following the steps described in the 'G5 – NLP for Image Captioning – Colab Steps.docx' file

## Skip-gram

Skip-gram is a natural language processing word embedding technique that learns vector representations of words by predicting the context of a target word within a given window of text [13]. Each word is represented as a high-dimensional vector in this approach, with the distance between vectors indicating the semantic similarity between words. Then, a neural network is used in the skip-gram model to predict the likelihood of occurrence of context words given a target word. This model learns a dense representation of each word in the vocabulary by optimizing the model to minimize the difference between predicted and actual probabilities.

The procedure implemented was as follows: Preprocessing was done on the dataset, resulting in 'trainImageCaptions' and 'testImageCaptions.' These were further preprocessed to remove 'startseq' (which marks the start of a caption), and use 'endseq' (which marks the end of each caption) to determine the appropriate window size of the context. This was all melted into a corpus of unique words, which is fed into the model.

Next, the model's embedding size was set to 32 (deduced by a function that returns the longest caption), and the words were one-hot encoded into vectors. A window size of 4 was chosen, 2 words to the left and 2 to the right. A larger window size could not be used as some captions consisted of only 5 words. A smaller size of 1 would also not be adequate context. To cover corner cases, such as retrieving the context of the first word (i.e. there do not exist 2 words to its left), the context on its right is taken twice (to retrieve 4 words). The same is done for the last words, taking the 2 words to their left twice. If there is only one word on the left or right, it is taken twice, with the 2 to the opposite side.

Finally, the model, which consists of 1 layer, was then fed _ batches of words (as one-hot encoded vectors).

## Continuous-Bag-of-Words (CBOW)

Continuous Bags of Words (CBOW) is a neural network model commonly used in NLP for word embedding generation. In CBOW, the goal is to predict a target word given its context words. The context words are represented as a bag of words and are combined to generate a dense vector representation for the target word. This approach has been shown to be effective in various NLP tasks, such as sentiment analysis and text classification. Mikolov et al. [13] demonstrated the effectiveness of CBOW in generating word embeddings and showed that it outperformed other state-of-the-art methods in multifarious word similarity tasks.

The same procedure described in skip-gram was followed to implement CBOW, noting some crucial differences. The input fed into the model was instead the context words, and the center word was used as the label (rather than the reverse done in skip-gram). Additionally, the model was trained on only 1000 (rather than 5000), as it was computationally expensive.

## Results

Skip-gram:
Epoch 1000: Loss = 5.379305
Epoch 2000: Loss = 6.373714
Epoch 3000: Loss = 5.329025
Epoch 4000: Loss = 5.506539
Epoch 5000: Loss = 5.520831

CBOW:
Epoch 1000: Loss = 82.9637

Although CBOW was unable to be trained for the same number of epochs, within the first 1000 epochs a large discrepancy can be seen. Using skip-gram resulted in a significantly smaller loss than its counterpart. Another influential factor was how long each took to compile. Skip-gram was able to run 5000 epochs in 30 minutes, whereas CBOW spent 196 minutes on just 1000 epochs.

Furthermore, Principal Component Analysis (PCA) was performed and the first 2 principal components were plotted for a sample of words, for both skip-gram and CBOW shown in Fig. 2 and Fig. 3 respectively. These show the similarity between the words. The closer words are to one another in the plot, the more similar they are. Although the same sample of words was used for both plots, each word has vastly different weights in skip-gram versus CBOW, showing the extent to which they produce different embeddings and understandings. However, the plot for skip-gram, Fig. 2, seems to be the more dependable of the two, due to its lower loss.

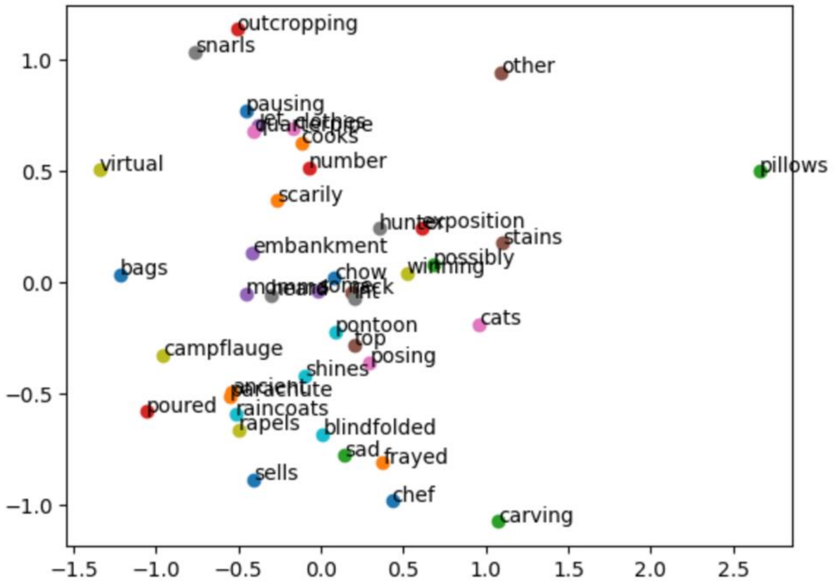

Fig. 2: First two principle components on a sample of words, from skip-gram.

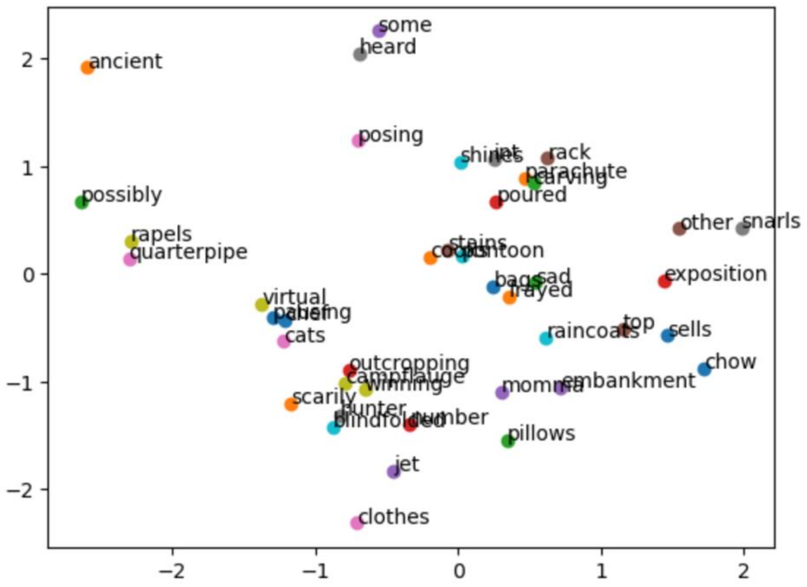

Fig. 3: First two principle components on a sample of words, from CBOW.

## Conclusion

In summation, ChatGPT has generated promising outputs and shows impressive understanding of topics. It has also been proved that feeding the LLM more detailed instructions is more likely to produce a higher quality output that meets the needs of the user. Useful prompting includes defining a ‘role’ (such as ‘pretend you are an expert in deep learning’), and describing specific instructions in concise sentences. Nonetheless, the reliability of the output is questionable. For this reason, as it stands, ChatGPT cannot accurately and competently replace human work, and should rather be used as an assistive tool (for example, ChatGPT was able to list some relevant papers, which should then be checked and read manually). It is also important to note that ChatGPT 3 was trained up until 2021, so will not be well-versed in the most recent advancements in image captioning, or any other field of research.

Pertaining to the findings of the implemented word embeddings, although Skip-gram outperformed CBOW, its results were still not satisfactory. Neither will be implemented in the upcoming phase. Since both ChatGPT literature reviews had no mentioned what was used in the papers, they were manually checked. Of the seven various papers given by ChatGPT, only three had explicitly mentioned which technique they implemented. In [6], [7], and [8], GloVe word embeddings were applied. Thus, we will also be utilizing GloVe in Phase II on our image captioning problem.

# References

[1] J. White et al., “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,” *arXiv*, 2023. doi: 10.48550/ARXIV.2302.11382.

[2] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, “From show to tell: A survey on deep learning-based image captioning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 539–559, 2023.

[3] adityajn105, “Flickr 8K dataset,” *Kaggle*, 27-Apr-2020. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k. [Accessed: 16-Mar-2023].

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] K. Xu et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” *arXiv*, 2015. doi: 10.48550/ARXIV.1502.03044.

[6] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[8] M. Geng and Q. Zhao, “Improve image captioning by modeling dynamic scene graph extension,” *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022.

[9] J. Johnson, A. Karpathy, and L. Fei-Fei, “DENSECAP: Fully convolutional localization networks for dense captioning,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] J. Lu, J. Yang, D. Batra, and D. Parikh, “Neural baby talk,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[11] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based Image Description Evaluation,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] G. Luo, L. Cheng, C. Jing, C. Zhao, and G. Song, “A thorough review of models, evaluation metrics, and datasets on image captioning,” *IET Image Processing*, vol. 16, no. 2, pp. 311–332, 2021.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *arXiv*, 2013. doi: 10.48550/ARXIV.1301.3781.