# "Let's not Quote out of Context": Unified Vision-Language Pretraining for Context Assisted Image Captioning

**Abisek Rajakumar Kalarani** and **Pushpak Bhattacharyya**
Department of Computer Science and Engineering, IIT Bombay, India
{abisekrk, pb}@cse.iitb.ac.in

**Niyati Chhaya** and **Sumit Shekhar**
Adobe Research, India
{nchhaya, sushekha}@adobe.com

## Abstract

Well-formed context aware image captions and tags in enterprise content such as marketing material are critical to ensure their brand presence and content recall. Manual creation and updates to ensure the same is non trivial given the scale and the tedium towards this task. We propose a new unified Vision-Language (VL) model based on the One For All (OFA) model, with a focus on context-assisted image captioning where the caption is generated based on both the image and its context. Our approach aims to overcome the context-independent (image and text are treated independently) nature of the existing approaches. We exploit context by pretraining our model with datasets of three tasks- news image captioning where the news article is the context, contextual visual entailment, and keyword extraction from the context. The second pretraining task is a new VL task, and we construct and release two datasets for the task with 1.1M and 2.2K data instances. Our system achieves state-of-the-art results with an improvement of up to **8.34** CIDEr score on the benchmark news image captioning datasets. To the best of our knowledge, ours is the first effort at incorporating contextual information in pretraining the models for the VL tasks.

**Context:**

HAVANA: The Teatro Nacional a 2056 seat theater on the Plaza de la Revolucion was sold out. Two dozen photographers and videographers swarmed the aisles. The Minnesota Orchestras concert here Friday night was greeted not only as a rare chance to hear an orchestra from overseas but as a symbol of the rapprochement between the United States and Cuba. The concert, the first by a large United States orchestra here in more than 15 years was greeted with several standing ovations and huge cheers when the Minnesotans teamed up with the Cuban pianist Frank Fernandez and two local choirs to perform Beethovens Choral Fantasy. "They played beautifully, they send you to the clouds", Graciela Fonseca, 73 said after it ended adding that she viewed the concert as a sign of friendship between the two nations. "It was not your typical concert at Orchestra Hall in Minneapolis. Tickets here cost around 50 cents with students paying only half that part of an effort to make cultural events accessible in a country where salaries are low", said Rafael Vega the director of the theater which also presents ballet concerts plays and comedy.

**Image Caption:** A group of people in a large auditorium.

**Context Assisted Image Caption**: Audience members gave a standing ovation to the Minnesota Orchestra at a concert in Havana on Friday.

**Figure 1:** An example image with its context. Text in blue and green can be inferred from the image and the context respectively. Text in red requires both the image and the context.

## 1 Introduction

Large enterprises have several teams to create their content for the purpose of marketing, campaigning, or even maintaining a brand presence. Multimodal assets, particularly images are an integral part of this. The scale and the speed at which one needs to create and update content, especially to ensure personalization requires several resources, which in turn acts as a hindrance to the success of the enterprise. Opportunistic updates are critical for success in the current competitive marketing and advertising scenario. Ensuring that every multimodal asset associated with any piece of enterprise content has an appropriate caption and tag is not possible given the scale. While apparently a nuanced aspect of any image, the caption serves as the key information carrier of what the image is all about, in turn ensuring the right recall (ability to find) for the content that contains this image. If an image has a well-formed caption, that captures the context accurately – it also makes the document accessible. Making enterprise content accessible is an important metric for large organizations as they strive to be inclusive. The scale and the quick turn-around time demanded for the content creation cycle results in the lack of correct tagging and captions of images (multimodal assets), in turn an eventual lost of revenue and a target customer base. We propose a method towards automated context-aware captioning of images targeted to reduce the tedium and this critical gap in the enterprise authoring and content creation process.

Large-scale pretraining of language models (Devlin et al. 2019; Brown et al. 2020) has witnessed

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 5: Industry Track*, pages 695–706
July 10-12, 2023 ©2023 Association for Computational Linguistics

great success in many downstream NLP tasks. This success has inspired multi modal pretraining for image-text, image-only, and video-text tasks. Currently, building unified models that jointly learn multiple vision-language tasks is gaining a lot of attention and has shown promising results on many VL tasks (Wang et al. 2022, Lu et al. 2020, Cho et al. 2021, Wang et al. 2021).

The existing unified Vision-language models focus on tasks like image captioning (Stefanini et al., 2022), visual question answering (Wu et al., 2017), visual entailment (Xie et al., 2019), and image-text retrieval (Wang et al., 2019) that consider the image as a standalone entity. However, images are typically accompanied by text that adds additional meanings which are not utilized in these tasks as shown in Figure 1. Also, the same image can mean different things in different contexts. For example, a picture of a football player being emotional can mean they are celebrating a goal or are disappointed with their shot, depending on the context. Hence it is essential to consider the context of the image for understanding it completely.

Traditional image captioning models do not use contextual information. In news image captioning (Biten et al., 2019), the generated caption contains information extracted from both the news article and the image. The news image captioning task is a special subtask of context assisted image captioning task that uses the news article as the contextual information about the image. In our work, the task names- news image captioning and context assisted image captioning are hence used interchangeably. Existing pretrained VL models lack the ability to use contextual information as the pretraining tasks do not contain long text associated with image-text pairs. We propose a new unified VL model based on the One For All (OFA) model, with a focus on using the contextual information associated with the image for real-world problems like news image captioning.

As there are no existing VL classification task that uses contextual information, we introduce a new VL task called 'Contextual Visual Entailment'. Visual entailment (Xie et al., 2019) is a refined image-text matching task that checks for the entailment of the caption with the premise image. Visual entailment deals with only the descriptive characteristics of the image. In our contextual visual entailment task, both the image and the context of the image are treated as the premise, and the entailment of the caption is predicted with respect to both.

Our contributions are:

- A new unified VL model pretrained for keyword extraction, contextual visual entailment, and news image captioning with a focus on using contextual information which has not been explored before.
- State-of-the-art results on the GoodNews and NYTimes800k datasets with an improvement of $8.34$ CIDEr points on the GoodNews dataset.
- A novel VL classification task where the context information surrounding the image is utilized for detecting the entailment of the caption with the image.
- Release of two datasets[1]- a large synthetic dataset consisting of $1.1$M Image-Caption pairs with context and a more challenging dataset with manually annotated negative samples consisting of $2.2$K instances for the proposed contextual visual entailment task.

## 2 Related Work

**Image Captioning** was initially conceived as a caption retrieval or template filling task. It involved matching the query image with a predefined set of captions or identifying the objects in the image to place them in predefined templates (Farhadi et al. 2010; Li et al. 2011; Kulkarni et al. 2013). The advancements made with deep learning based techniques in machine translation inspired the community to adopt similar techniques for image captioning where images were fed to the encoder and the decoder generated caption as a sequence of words (Farhadi et al. 2010; Li et al. 2011; Kulkarni et al. 2013). Attention allows decoder to focus on different parts of the input and hence it was incorporated to generate words focused on important regions of the image in both sequential models (Xu et al. 2015; Lu et al. 2017; Anderson et al. 2018; Huang et al. 2019) and transformer based models (Cornia et al., 2020). In recent years, the models are trained on huge datasets with several millions of image-text pairs for image captioning (Li et al. 2020; Su et al. 2020; Radford et al. 2021). However, in all these works images are treated as standalone entities and their context is not taken into account.

**News Image Captioning** deals with the generation of captions for news images. The news articles

[1] Code and data are available at: https://github.com/abisekrk/context-assisted-image-captioning

contain the context of the image and they are taken into account during the caption generation process. Biten et al. (2019) propose an encoder-decoder model with attention over both image and news article encodings to generate news image captions. It generates captions with placeholders for named entities and fills those placeholders by choosing named entities from the news article. Chen and Zhuge (2019) model it as a query-based summarization problem where the news image acts as the query and the news article is the source text to be summarized. Tran et al. (2020) use transformers with separate encoders for extracting image features, object features and faces present in the image. The decoder receives the input from all three encoders to generate caption. Liu et al. (2021) use a visual selective layer that learns to align the image features with the text in the news article to generate captions. Yang et al. (2021a) discuss the journalistic guidelines followed while writing news image captions in journals and incorporate them in the generation process. Zhang et al. (2022) use prompt tuning to finetune pretrained models for news image captioning.

**Unified Vision-Language (VL) modeling** is a new paradigm that involves creating a unified framework for multiple vision-language tasks, allowing models to be trained on a range of datasets constructed for a range of tasks. ViLBERT (Lu et al., 2019) extends the BERT (Devlin et al., 2019) architecture to work with visual inputs. Lu et al. (2020) propose a multi-task training approach with 12 VL datasets on 4 broad tasks. VL-T5 (Cho et al., 2021) combines multiple VL tasks as text generation tasks using pretrained models for image features. UniT (Hu and Singh, 2021) unifies cross-modal tasks by using a modality specific encoder and a shared decoder. UFO (Wang et al., 2021) proposes to use the same transformer architecture as the encoder for both image and text in VL tasks. UniTAB (Yang et al., 2021b) supports VL tasks with bounding boxes by encoding the text and box output sequences to shared token sequences. OFA (Wang et al., 2022) abstracts all VL tasks into sequence-to-sequence problems.

Existing unified VL models do not consider the context of the image in their pretraining tasks. Our unified model is pretrained with tasks that include contextual information and hence it achieves state-of-the-art results on news image captioning datasets.

## 3 Dataset

We pretrain our model on a large pretraining dataset and evaluate its performance on benchmark datasets for news image captioning.

### 3.1 Pretraining Datasets

We use Visual News (Liu et al., 2021) and KPTimes (Gallina et al., 2019) datasets for constructing our pretraining datasets.

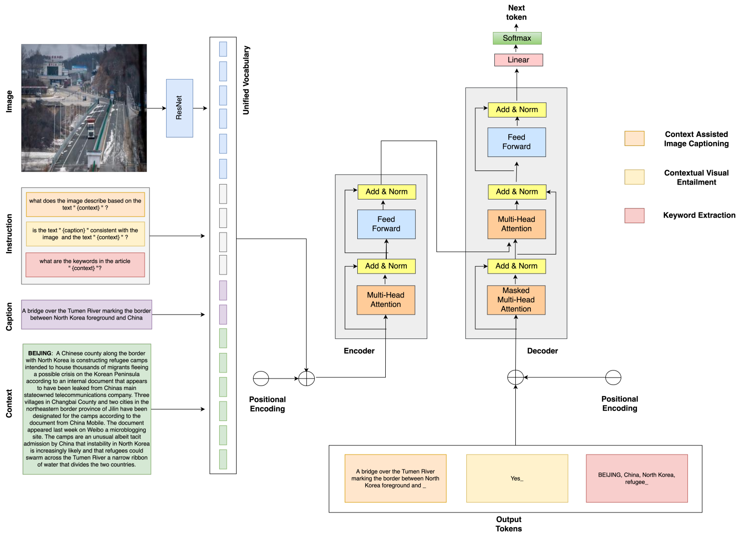
Visual News dataset was compiled by collecting news articles from four news agencies: The Guardian, BBC, USA Today, and The Washington Post. It only includes the articles with high resolution images and where the caption length is between 5 and 31 words. It is diverse with differences in properties like average caption length, article length, and distribution of named entities across news agencies.

KPTimes dataset was constructed by crawling over 0.5 million news articles, mainly from New York Times. The metadata associated with field- "news_keywords" and "keywords" form the gold standard keyphrases . The three pretraining datasets are constructed from these datasets.

**News Image Captioning:** We removed duplicate captions, and news articles without images, and captions from Visual News dataset. The cleaned dataset consists of $11,97,000$ data instances with image, news article, and the caption. These are split as $11,17,697$ for training, $40,000$ for validation, and $40,000$ for test sets respectively.

**Contextual Visual Entailment** is a binary classification problem, so it is required to construct both positive and negative pairing of the image, and caption with the context. For the data instances where the caption entails the image and the context, the original image, the caption, and the context from the training split of the above news image captioning dataset are used (P). We use the following operations to generate the inconsistent pairs in our dataset:

1. Choose a random caption different from the correct caption **(N-I)**.
2. Replace the named entities in the correct caption with named entities from randomly chosen caption **(N-II)**. For example, the caption *'John Garrison performing in Berlin, April 2015'* will be changed to *'Mark Pattinson performing in London, April 2015'*.
3. Keep the named entities of the original caption intact but replace the remaining content with a


what does the image describe based on the text " {context} " ?

is the text " {caption} " consistent with the image and the text " {context} " ?

what are the keywords in the article " {context} "?

A bridge over the Tumen River marking the border between North Korea foreground and China

**BEIJING**: A Chinese county along the border with North Korea is constructing refugee camps intended to house thousands of migrants fleeing a possible crisis on the Korean Peninsula according to an internal document that appears to have been leaked from Chinas main stateowned telecommunications company. Three villages in Changbai County and two cities in the northeastern border province of Jilin have been designated for the camps according to the document from China Mobile. The document appeared last week on Weibo a microblogging site. The camps are an unusual albeit tacit admission by China that instability in North Korea is increasingly likely and that refugees could swarm across the Tumen River a narrow ribbon of water that divides the two countries.

A bridge over the Tumen River marking the border between North Korea foreground and _

Yes_

BEIJING, China, North Korea, refugee_

**Figure 2:** An overview of our unified Vision-Language model pretrained for the three subtasks- context assisted image captioning, contextual visual entailment, and keyword extraction.

random caption that has the same type and the same number of named entities **(N-III)**. For example, the caption *'John Garrison performing in Berlin, April 2015'* will be changed to *'John Garrison waiting in queue for filing tax returns in Berlin, April 2015'*.

Named entity recognition is done with SpaCy (Honnibal and Montani, 2017) in our experiments. SpaCy allows the detection of 18 different named entities. We only use the named entities labeled as *'PERSON'*, *'FAC'*, *'ORG'*, *'GPE'*, *'LOC'*, and *'EVENT'* that represent a person, building/airport, organization, geopolitical entities, location, and event respectively, as they occur more frequently.

The N-I class of negative captions will have different information and different named entities from the original caption. The N-II class will have same information as the original caption but will contain different named entities. The N-III class of captions will have same named entities but will convey different information. The final dataset has 1005925, 55884, and 55884 instances in the train, validation and test split respectively. We also create a separate manually annotated challenging dataset for evaluation.

In addition to synthetically creating a dataset for pretraining, we create and release a manually annotated challenging dataset for the task of contextual visual entailment consisting of $2.2K$ data instances. The negative captions in this dataset are created manually by changing a word or a small phrase from the original caption, such that its meaning changes significantly without much difference in the sentence structure. For example, *'Supporters marched peacefully during the protest'* will be changed to *'Supporters marched violently during the protest'*. The negative examples created in these ways will ensure that the models need to learn the relationship between image, caption, and context to identify the entailment correctly. This is used to test the model's knowledge of image-caption entailment at a more finer level.

**Annotation Details:** The annotations were performed by two annotators proficient in English. One is a master's student and the other is a bachelor's student in Computer Science and Engineering. They were provided with examples of negative captions before annotation. The image links, caption, and context were shared in Google Sheets for annotation. The annotators were only asked to select a word or phrase from the caption and replace it with a new word or phrase. The modified captions were exchanged and verified by each other.

### 3.1.1 Keyword Extraction

The dataset for the keyword extraction task is constructed from KPTimes dataset after removing duplicate news articles. The news article forms the input to the system and the sequence of keywords form the output. The final dataset has 259902, 10000, and 10000 data instances in the train, validation and test split respectively. The training data from these three datasets are combined to generate the pretraining dataset with 2.3M data instances.

| Dataset | Model | B-4 | MET. | ROUGE | CIDEr | Named Entities | |
|---|---|---|---|---|---|---|---|
| | | | | | | **P** | **R** |
| **GoodNews** | GoodNews | 1.86 | 13.75 | 20.46 | 17.57 | 8.23 | 6.06 |
| | Transform and Tell | 6.05 | 10.30 | 21.40 | 54.30 | 22.20 | 18.70 |
| | Visual News | 6.10 | 8.30 | 21.60 | 55.40 | 22.90 | 19.30 |
| | JoGANIC | 6.83 | 11.25 | 23.05 | 61.22 | 26.87 | 22.05 |
| | NewsMEP | 8.30 | 12.23 | 23.17 | 63.99 | 23.43 | 23.24 |
| | OFA | 6.41 | 10.63 | 23.59 | 67.19 | 23.06 | 19.04 |
| | **Ours** | 7.14 | 11.21 | **24.30** | **72.33** | 24.37 | 20.09 |
| **NYTimes800K** | Transform and Tell | 6.30 | 10.30 | 21.70 | 54.40 | 24.60 | 22.20 |
| | Visual News | 6.40 | 8.10 | 21.90 | 56.10 | 24.80 | 22.30 |
| | JoGANIC | 6.79 | 10.93 | 22.80 | 59.42 | 28.63 | 24.49 |
| | NewsMEP | 9.57 | 13.02 | 23.62 | 65.85 | 26.61 | 28.57 |
| | OFA | 6.91 | 10.77 | 22.70 | 61.81 | 27.14 | 22.51 |
| | **Ours** | 7.54 | 11.27 | 23.28 | **66.41** | 28.11 | 23.25 |

**Table 1:** Experimental results on the GoodNews and NYTimes800K datasets compared with other models. P and R denote the precision and recall of generating named entities. B-4 indicates BLEU-4 and MET. indicates METEOR.

### 3.2 Benchmark Datasets

The performance of models trained on the pretraining datasets is evaluated on two benchmark datasets- GoodNews (Biten et al., 2019) and NYTimes800K (Tran et al., 2020). We follow the train, validation, and test splits from the original work for both datasets. The GoodNews dataset has $424,000$, $18,000$, and $23,000$ in training, validation, and test split respectively. The NYTimes800K dataset has $763,000$ training, 8000 validation, and $22,000$ test instances in the dataset.

## 4 Our Model

Unified Vision-Language (VL) modeling has shown great promise in multiple VL tasks. Hence, we use a unified model for all three tasks- context assisted image captioning, contextual visual entailment, and keyword extraction. Figure 2 shows an overview of our unified VL pretraining strategy. We use the OFA$_{Large}$ (Wang et al., 2022) architecture. OFA is a task and modality agnostic model that unifies all vision-language, vision-only, and language-only tasks using a sequence-to-sequence learning framework. We use ResNet152 (He et al., 2016) and VQGAN (Esser et al., 2020) to obtain visual tokens for the given image. The text (context and caption) is tokenized by byte-pair encoding (BPE). A single unified vocabulary is used for both visual and linguistic tokens. Transformers are used as encoders and decoders and all vision-language tasks are abstracted to seq-to-seq conversion tasks with specific instructions created for each task, similar to the OFA pretraining.

The pretraining of our model involves three tasks- News image captioning, contextual visual entailment, and keyword generation. For news image captioning, we convert the image, caption, and context into a sequence of input tokens and generate the caption as a sequence of tokens conditioned on these input tokens. For keyword generation, the news article is tokenized as the input sequence and the keywords are generated by the model as the output sequence. Contextual visual entailment is a classification task, so the input sequence to the model is the image, caption, and context tokens and the model is trained to generate 'Yes' or 'No' as the output indicating if the caption entails the image and context or not respectively.

The model is trained to reduce the cross entropy loss. For the input sequence x consisting of visual and text tokens and output y, the loss is given as:

$$\mathcal{L} = \sum_{i}^{|y|} logP_{\theta}(y_i|y_{<i}, x)$$

where $y_i$ is the text token to be predicted and $y_{i-1}$, $y_{i-2}$ are tokens predicted so far.

## 5 Experiments

The experimental details for pretraining and finetuning for context assisted image captioning are discussed here.

### 5.1 Pretraining

We used OFA$_{Large}$ architecture for pretraining our model. The model has 472M parameters with 12

| Model | BLEU-4 | METEOR | ROUGE | CIDEr |
|---|---|---|---|---|
| BLIP-2 + GPT-3 | 2.06 | 8.48 | 13.22 | 17.12 |
| OFA | 6.41 | 10.63 | 23.59 | 67.19 |
| OFA + Captioning + Contextual Visual Entailment | 6.90 | 11.01 | 23.83 | 69.97 |
| OFA + Captioning + Keyword Extraction | 6.69 | 10.81 | 23.44 | 67.83 |
| OFA + Captioning | 6.85 | 10.90 | 23.70 | 68.93 |
| Our Model (Without Context) | 2.24 | 5.34 | 14.45 | 18.04 |
| Our Model + NE | 6.95 | 11.06 | 23.98 | 70.03 |
| Our Model | **7.14** | **11.21** | **24.30** | **72.33** |

**Table 2:** Ablation study results on the GoodNews dataset. NE denotes fine-tuning done with named entities extracted separately. 'OFA + X' denotes pretraining of OFA done with X task. 'Our Model' refers to the model pretrained on the three tasks with context information.

encoder and 12 decoder layers. The weights were initialized with the publicly available OFA$_{Large}$ checkpoint to retain the knowledge from other VL tasks. The model was pretrained on the 2.3M data instances from the pretraining datasets.

All 3 tasks were abstracted into sequence-to-sequence task. For the instances of news image captioning dataset, the instruction was "*What does the image describe based on the text <context> ?*", where <context> holds the tokens from the news article. For contextual visual entailment, the instruction was "*Is the text <caption> consistent with the image and the text <context> ?*", where <caption> and <context> contain the text tokens from the caption and the context. For the keyword extraction task, the instruction given was "*What are the keywords in the article <context>?*".

### 5.2 Context Assisted Image Captioning

Our unified model pretrained model for the three tasks was finetuned on GoodNews and NYTimes800K datasets for the task of context assisted image captioning. The image resolution was fixed at $384*384$ and the news article was clipped to 512 tokens. The maximum caption length was fixed at 30. We use a batch size of 8 for training. We train the model with early stopping and choose the model that achieves the best CIDEr score on the validation set. The best-performing model is then tested on the unseen test data and the results are summarized in Table 1.

### 5.3 Training Details

The experiments were done with the OFA$_{Large}$ architecture. For both pretraining and finetuning, the image resolution was fixed at $384 * 384$. The input token length was restricted to 512 tokens while the output was restricted to 30 tokens. The dropout ratio was set to 0.1. We used Adam optimizer (Kingma and Ba, 2014) with 0.9 and 0.999 as the $\beta$ values with $\epsilon = 1e - 08$ and warm-up ratio was set as 0.06. We used an initial learning rate of $1e - 5$ with polynomial decay. We used a beam size of 10 during the test inference with temperature 0.98. We also used mixed precision training to speed up the training process.

### 5.4 Frozen Image Encoder + Frozen LLM

LLMs and large-scale pretrained VL models have shown great zero-shot performance in many downstream applications. We use BLIP-2 (Li et al., 2023) for getting zero-shot image captions for the GoodNews dataset. These captions are generated without contextual information and are descriptive in nature. These captions are passed to a LLM along with contextual information to generate context assisted image captions. We use `text-davinci-003` model in the GPT-3 family (Ouyang et al., 2022). The prompt for generating the caption was "Add contextual information to the caption. Caption: <sample caption> Context: <sample context>". We randomly sampled a caption, and context pair from the training dataset of the GoodNews dataset and used it as an example in the prompt. The contextual captions are predicted for caption and context pair in the test set. The results are discussed in Table 2.

### 5.5 Ablation Study

In order to analayse the importance of the three pretraining tasks we used, we pretrained the OFA model using three different subsets of the pretraining tasks. We pretrained the model with only "Captioning and Contextual Visual Entailment" tasks, with only "Captioning and Keyword Extraction" task and with only "Captioning" task and compared their performance with the model trained on all the three tasks.

Previous works in news image captioning (Liu et al. 2021; Yang et al. 2021a; Zhang et al. 2022) have shown that extracting named entities from the context and feeding them to the decoder helps generate correct named entities in the caption. Hence, we also try injecting named entities into the prompt while finetuning the model. We use SpaCy for identifying and extracting named entities. We update the prompt as "what does the image describe about the names <named entities> based on the text <context>?" during finetuning. We clipped the named entity tokens to 64 and restricted the context to 512 tokens as done in previous experiments. We also perform experiments without using the contextual information with our pretrained model in the traditional image captioning setting, to analyze the usefulness of the contextual information. The results are summarized in Table 2.

## 6 Results and Analysis

Our model achieves state-of-the-art results on both GoodNews and NYTimes800K datasets. The OFA model finetuned on benchmark datasets also shows good performance. This shows the ability of OFA to adapt to new tasks and the correctness of our instructions for finetuning. However, it can be seen that due to the lack context information in the pretraining tasks used in OFA, the model doesn't produce substantially better results compared to the current SOTA models. Our model pretrained on the three tasks shows a 5.14 CIDEr score improvement over the OFA model on the GoodNews dataset which is an 8.34 CIDEr score improvement over the current SOTA model. The model also achieves a SOTA result of 66.41 CIDEr score on the NYTimes800K dataset.

The average length of news articles in GoodNews and NYTimes800K dataset are 451 and 974 words respectively. The larger article length in NYTimes800K dataset is the reason for the CIDEr scores being closer to the current SOTA as the context length in our experiments is restricted to 512 tokens. We also obtain comparable performance in precision and recall of named entity generation despite not feeding the named entities directly to the model like in the previous works.

The BLIP-2 model has shown great promise in zero-shot image caption generation. We use BLIP-2 to generate descriptive captions and feed those captions as input to the GPT-3 model along with the context to generate the final context assisted caption. The BLIP-2 + GPT-3 model generates fluent captions but it does not contain the relevant information based on the image features as indicated by the poor performance on evaluation metrics in Table 2. This indicates that it is essential to train with both image and context together.

Our pretraining tasks indirectly direct the model to capture named entity information from the context. However, earlier works on news image captioning show that extracting named entities and feeding them directly to the model can help it generate better captions with correct named entities. Our pretrained model showed a slight decrease in performance when named entity information is presented to it in the prompt. This is because the named entity tokens take up valuable space in the 512 tokens allowed for the context, leading to information loss gained from the context. Also, since only 64 tokens are allowed for named entities, not all the important entities in the news article are presented in the prompt and it disadvantages the model in both ways.

We also pretrained the OFA model with a subset of the three pretraining tasks to identify the importance of each task in pretraining. The models pretrained with a combination of two tasks and with only the captioning task performed poorly compared to the model pretrained on the three tasks. This shows the importance of training with all three tasks. Between the two two-task pretrained models, the model that used contextual visual entail task performed better, indicating the usefulness of the task we introduced.

## 7 Summary and Conclusion

In this work, we proposed a new unified VL model that uses contextual information of images that has not been utilized in pretraining before. We introduce a new VL classification task called contextual visual entailment and pretrain a model with three subtasks that uses long text along with image and caption. Our model achieves new state-of-the-results on benchmark datasets for news image captioning and highlights the importance of using contextual information in pretraining.

In the future, we aim to deploy our model to allow context-aware caption generation which could be used in enterprise authoring and many other content creation processes.

## References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jingqiang Chen and Hai Zhuge. 2019. News image captioning based on text summarization using image as query. In *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 123–126.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. *ArXiv*, abs/2102.02779.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, J. Hockenmaier, and David Alexander Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2019. Kptimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1419–1429.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4642.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, USA. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

J. Lu, C. Xiong, D. Parikh, and R. Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250, Los Alamitos, CA, USA. IEEE Computer Society.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10434–10443.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, PP.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. *ArXiv*, abs/1908.08530.

Alasdair Tran, A. Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13032–13042.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Ufo: A unified transformer for vision-language representation learning. *ArXiv*, abs/2111.10023.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2019. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:394–407.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.*, 163:21–40.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2048–2057. JMLR.org.

Xuewen Yang, Svebor Karaman, Joel Tetreault, and Alejandro Jaimes. 2021a. Journalistic guidelines aware news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5162–5175, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021b. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *ArXiv*, abs/2111.12085.

Jingjing Zhang, Shancheng Fang, Zhendong Mao, Zhiwei Zhang, and Yongdong Zhang. 2022. Fine-tuning with multi-modal entity prompts for news image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4365–4373, New York, NY, USA. Association for Computing Machinery.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.

# A Appendix

## A.1 Dataset Details

Table 3 provides the summary of the three datasets used for pretraining.

| | | Train | Val | Test |
|---|---|---|---|---|
| **Pretraining** | Keyword Extraction | 259902 | 10000 | 10000 |
| | Contextual Visual Entailment | 1005925 | 55884 | 55884 |
| | News Image Captioning | 1117697 | 40000 | 40000 |
| **Benchmark** | GoodNews | 424000 | 18000 | 23000 |
| | NyTimes800K | 763000 | 8000 | 22000 |

**Table 3:** Summary of the statistics of the datasets used for pretraining and benchmarking.

| **Model** | | **Overall** | | | |
|---|---|---|---|---|---|
| | | **Acc.** | **Pre.** | **Rec.** | **F1** |
| **w/o context** | 1) CLIP + FNN | 61.30 | 61.61 | 60.00 | 60.79 |
| | 2) CLIP + Transformer | 60.00 | 59.83 | 60.87 | 60.34 |
| | 3) Ours | 66.96 | 69.70 | 60.00 | 64.49 |
| **w/o image** | 4) CLIP + FNN | 59.13 | 61.54 | 48.70 | 54.37 |
| | 5) CLIP + Transformer | 56.96 | 56.25 | 62.61 | 59.26 |
| | 6) Ours | 65.22 | 66.67 | 60.87 | 63.64 |
| **w/ context** | 7) CLIP + FNN | 64.35 | 63.64 | 66.96 | 65.25 |
| | 8) CLIP + Transformer | 65.65 | 63.64 | **73.04** | 68.02 |
| | 9) Ours | **73.04** | **79.78** | 61.74 | **69.61** |

**Table 4:** Experimental results on the manually annotated contextual visual entailment dataset, where w/o context, w/o image and w/ context indicate experiments done without context (Image + Caption), without image (Caption + Context), and with context (Image + Caption + Context).

## A.2 Additional Experiments

Our pretrained model achieves state of the results on news image captioning task. In addition, it performs very well on the other two pretraining tasks. The contextual visual entailment is a new task introuduced by our work and hence we propose baselines for comparing the results of our model. We compare our model's performance on keyword extraction against standard works.

### A.2.1 Contextual Visual Entailment

We propose a two baselines for contextual visual entailment, where the image and text features are extracted from pretrained networks. The features are obtained from a pretrained CLIP (Contrastive Language–Image Pre-training) model. CLIP (Radford et al., 2021) was trained on large scale image-text corpus to minimize contrastive loss such that the text embedding and the image embedding will have higher cosine similarity if the text describes the image perfectly and low when the text incorrectly describes the image.

**CLIP and FNN model**

In our CLIP embedding based models, the representation for image, caption, and context is obtained from a pretrained CLIP model.

The CLIP and FNN model, uses a simple early fusion strategy in which the image, caption, and context embeddings from CLIP are concatenated and fused with feed-forward neural networks. The three input embeddings are concatenated and passed to a two-layer feedforward neural network for combining the information. An output layer predicts the entailment label. The experiments are repeated without the context information and then again without the image features as input to study their impact on entailment detection.

**CLIP and Transformer model**

The fusion of image-text information using transformers has helped achieve good performance on many standard vision-language tasks (Zhou et al. 2020; Chen et al. 2020; Wang et al. 2022 ). The CLIP and Transformer model uses transformer Vaswani et al. (2017) encoders with multi-head attention to combine these multimodal information. A transformer layer receives the input features from the image, caption, and context and generates the combined representation for the information. The outputs from the transformer layers are pooled to a single dense layer followed by a classification layer to perform the final binary classification.

We summarize the results of our experiments

| **Model** | **F@10** |
|---|---|
| FirstPhrases | 9.2 |
| MultipartiteRank | 11.2 |
| CopySci | 11 |
| CopyNews | 39.3 |
| **Ours** | **40.6** |

**Table 5:** Results of Keyword extraction task on KPTimes dataset. F@10 represents the F1 score at the top N = 10 keyphrases

on the manually annotated contextual visual entailment data in Table 4 and compare it with the results produced by our pretrained model.

### A.2.2 Keyword Extraction

We use our pretrained model to perform keyword extraction as a sequence to sequence task where the output is the set of keyword tokens. We use similar hyperparameters to caption generation for keyword generation. We finetune the pretrained model for 5 epochs and report the results on KPTimes dataset in Table 5.

## A.3 Examples

Figure 3 shows examples from the GoodNews dataset and compares the caption generated by each model.

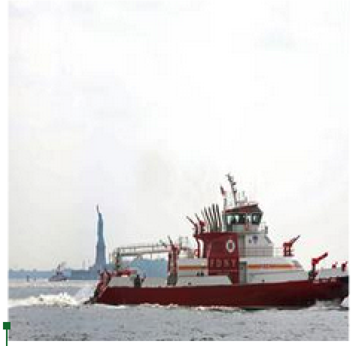
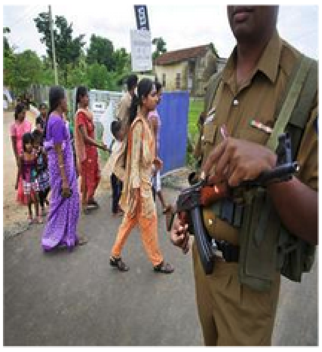
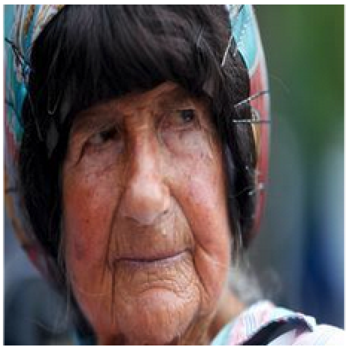
| Image | Context | Captions |
|---|---|---|
|  | Fears that a fire had erupted inside the Statue of Liberty on Wednesday led several hundred visitors to flee its crown observation deck pedestal and base officials said. In the end, it was a trickle of smoke emanating from an overheated elevator motor that set off the statues fire alarm system. This was the second such problem in recent days a spokesman for the Fire Department said….The alarms caused people in the statue, part of a throng of visitors to Liberty Island that swells to 15000 a day in the summer to descend stairways to get out, Ms Rambo said. For visitors who were in the crown that meant a scramble down 354 steps. | **True Caption**: A fire boat made a trip to the Statue of Liberty on Wednesday after an overheated elevator motor set off the fire alarm system<br>**BLIP-2**: a red and white tug boat in the middle of a body of water<br>**BLIP-2 + GPT-3**: The tug boat was sent to the island to help evacuate the visitors.<br>**Our model**: A boat escorted visitors to the Statue of Liberty on Wednesday after a fire broke out inside the statue |
|  | "The Tamils will go through the motions to try to use the council system but they will be frustrated because the system doesnt work", said Paikiasothy Saravanamuttu executive director of the Centre for Policy Alternatives in Colombo. So the Tamils will ask for reforms and the government will be very stupid if they frustrate them. The elections were hardfought and a variety of apparent dirty tactics were used to try to defeat the Tamil alliance…A fake version of a respected newspaper in Jaffna Uthayan was also distributed…The fake newspaper may explain why turnout in the first hours of the election was relatively light in Jaffna with many polling places reporting only a trickle of voters… The army confiscated more than 6000 acres. Now that the war is over however the government has decided to develop that land rather than return it to its owners… | **True Caption**: A police officer stood guard as voters left a polling station in Jaffna<br>**BLIP-2**: a man holding a gun in front of a group of people<br>**BLIP-2 + GPT-3**: The man holding a gun was a member of the Sri Lankan military, sent to intimidate Tamil candidates and voters in the country's first post-war<br>**Our Model**: A Sri Lankan soldier guarded a polling station in Jaffna on Thursday |
|  | WASHINGTON Concepcion Picciotto held a vigil for 32 years across the street from the White House protesting war and nuclear proliferation. Her small encampment never left the edge of Lafayette Square, her tent surrounded by signs with slogans like, "Read My Lips No New Wars". "I've been beaten up arrested many times and gassed", said Ms Picciotto, who is 77 and maybe five feet tall her skin worn and tanned. Early Thursday morning the United States Park Police confiscated Ms Picciottos tent and signs when she left for the night leaving a man in her place They left only a pale footprint on the red brick sidewalk Office workers passing by were caught off guard by the vigils conspicuous absenceIts been there since Ive been here even longer John Tomasetti 26 said on Thursday as he walked back to his office carrying lunch. | **True Caption**: Ms Picciotto 77 holds her vigil in Lafayette Square.<br>**BLIP-2**: a woman wearing a head scarf with pins on it<br>**BLIP-2 + GPT-3**: Concepcion Picciotto was a passionate activist who had been protesting for 32 years, wearing a head scarf with pins on it as a symbol of<br>**Our Model**: Concepcion Picciotto 77 at a vigil in Lafayette Square on Thursday |

**Figure 3:** Some examples from our dataset along with the captions generated by each model.