Image Captioning Unified (ICU): A Resource-Efficient Approach to Image Captioning

G061 (s2447563, s2449019, s2449767)

Abstract

Image captioning is a critical task in multimodal research, with numerous real-world applications. Unfortunately, progress in image captioning has been limited to the English language primarily due to the scarcity of large image-caption paired datasets in many other languages. To address this limitation, we propose a method for training image captioning networks in languages other than English using pre-trained models and a transformer network. Specifically, we use the Contrastive Language-Image Pre-training (CLIP) framework as our image encoder and different fine-tuned versions of the GPT-2 language model as decoders for our target language. The transformer network is then trained to adapt the encoder's outputs to the decoder's inputs. We evaluate our approach on three languages, namely English, German, and Arabic, achieving results comparable to state-of-the-art models, even with smaller datasets. Moreover, our proposed method outperforms existing approaches on such datasets. To demonstrate the capabilities and limitations of our solution, we perform extensive qualitative assessments. Our proposed method has the potential to bridge the gap between English and other language image captioning, thus enabling more applications to benefit from this technology [1].

1. Introduction

The task of image captioning involves generating natural language descriptions from image inputs, and it can take various forms based on input and output formulation. This includes specialized image captioning for fine-grained and specific tasks(Cho et al., 2022), and multilingual image captioning for generating language descriptions in different languages for a single image(Elliott et al., 2017). This paper focuses on the classic formulation of image captioning, where a model takes in an image and produces a single caption. To accomplish this, the model must possess multimodal understanding, meaning it must be able to comprehend images by recognizing objects within them and the relationships between them, as well as have language understanding to generate descriptions that reflect the information extracted from the images.

However, image captioning is a challenging task that requires significant amounts of data and computation (Mokady et al., 2021). The state-of-the-art models in this field feature millions of parameters (Li et al., 2020) and are trained on web-scale datasets (Sharma et al., 2018), limiting their adoption in various real-world applications. Moreover, this intensive data requirement widens the gap between English image captioning and other languages due to the lack of datasets in most languages. Even for the languages that have datasets, they are much smaller and are primarily derivative of existing English datasets created through the translation of subsets of them (ElJundi. et al., 2020) This translation process results in artifacts that hinder the quality of the captions in other languages, as found by (Freitag et al., 2020). The ability to provide high-quality image captioning in different languages is crucial for various applications, such as native-language image search, and audio-described videos for visually impaired viewers. Therefore, our proposed solution has significant implications for bridging the gap between English and other language image captioning and enabling more applications to benefit from this technology.

To address this issue, we propose using pre-trained large models to train image captioning networks on languages other than English with relatively small datasets similar to (Mokady et al., 2021). We employ CLIP as our image encoder and different fine-tuned versions of the GPT-2 language model as decoders for our target language. Finally, we train a transformer network that adapts the generated embedding from CLIP to a latent space for GPT-2 to produce captions that reflect the information in that embedding. This setup only requires the training of the transformer network and is thus both data and computation efficient (Li & Liang, 2021).

We evaluate our proposed method on three target languages: English, German, and Arabic. For English and German, we use the Multi30k dataset, while for Arabic, we use the more challenging Flickr8k dataset. We primarily benchmark our approach based on performance in the English datasets due to the inadequacy of automated metrics for other languages, and the fact that existing works on other language image captioning are all set up in a translation context that uses English as a pivot language, which is not comparable to our case (Thapliyal & Soricut, 2020).

Our model achieves comparable results in most met-

[1] https://github.com/ammarnasr/Multi-Lingual-Image-Captioning

rics compared to (Mokady et al., 2021), even with smaller datasets. When compared to traditional methods on similar datasets such as (Xu et al., 2015), we achieve better performance on English in the Flickr30k datasets. These results support our hypothesis that we can adapt pre-trained models for smaller datasets in other languages to overcome the data scarcity problem. Finally, we perform a thorough quantitative and qualitative analysis on all three models. While quantitative evaluation may not necessarily have a high correlation with human evaluation, it is essential to have comparable results. Qualitative evaluation shows us the true capabilities and limitations of our approach.

In conclusion, our proposed approach to training image captioning networks on languages other than English using pre-trained large models and a transformer network has shown promising results on three target languages: English, German, and Arabic. However, there is still much work to be done in this field, particularly in the development of automated metrics for evaluating image captioning in languages other than English. Additionally, more research is needed to explore the use of different pre-trained encoder and decoder architectures.

2. Data set and task

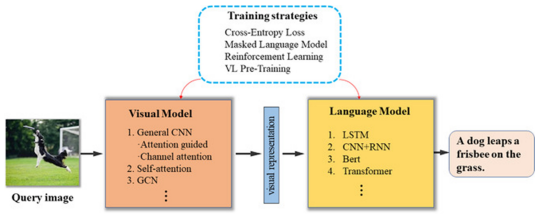

Figure 1. An overview of Image Captioning. Adapted from (Luo et al., 2022).

2.1. Resource-Efficient Image Captioning

As aforementioned, the task of image captioning is often resource-intensive, requiring large datasets and/or high computational power for the training process in order to obtain good results. In this paper, however, we explore image captioning in a resource-efficient way, using pre-trained models on small datasets for training.

Image captioning is the automated process of generating natural language descriptions of images based on their content/features. In other words, the task of image captioning takes as input an image and generates as output a text caption that describes that image, as shown in Figure 1. Intrinsically, image captioning incorporates both the fields of Natural Language Processing (NLP) and Computer Vision (CV) (Herdade et al., 2019). More specifically, in order to 'understand' and analyse the content of an image, image captioning employs a visual encoder (written as Visual Model in Figure 1) to extract important features from that image. Visual Encoders are neural networks (usually, CNNs or Transformer models) that work by transforming relevant features in an image into vectors of numerical values (or visual representations), which are then fed into a decoder, to produce the output caption (Suresh et al., 2022).

Decoders are also neural networks (e.g., RNNs or Transformer models) that take in the output vectors of the encoders containing the important features of an image and transform them into a natural language caption of that image. In order to generate these captions, decoder networks integrate language models to produce word-by-word captions of an image based on the vectors fed by the encoder (Suresh et al., 2022). Both encoders and decoders are trained using supervised techniques that involve minimizing the loss (e.g., cross-entropy loss) which measures the discrepancy between generated captions and reference or ground truth captions for a data set of images (Suresh et al., 2022).

In this paper, we perform the above-described image captioning process using multiple languages, including English, Arabic, and German. Arabic and German were specifically chosen as they are considered to be low-resource languages, i.e., there do not exist many verified data sets available in these languages that can be used for image captioning.

2.2. Data sets

Since we apply image captioning to multiple languages, we had to find appropriate data sets in these respective languages.

2.2.1. Multi30k

Multi30K (Elliott et al., 2016) is a dataset that builds on the original Flickr30k data set by extending it with verified German translations and descriptions of the original dataset. Flickr30k is a collection of 31783 images obtained from Flickr.com with each image's caption in English obtained by 5 independent annotators, resulting in 158915 captions per image (Young et al., 2014). Multi30k further extends this by providing 31014 German professional translations of English captions in Flickr30k in addition to 155070 independent German image captions obtained through crowd-sourcing. Since this data set contains both English and German captions, we use it for training and evaluating our English model and our German model (i.e., the English model is trained/evaluated only using English data, and the German model uses only German data).

2.2.2. Flickr8k-Arabic

Flickr8k-Arabic (ElJundi. et al., 2020) is a translation of the English Flickr8k (Hodosh et al., 2013) data set, a subset of Flickr30k which, as its name suggests, only contains 8000 images from Flickr, each captioned by 5 independent annotators. It is important to note that in Flickr8k-Arabic, the captions are translated using Google Translate and then edited and validated by a professional Arabic translator. However, as the only validated image captioning data set we could find in Arabic, combined with the fact that it is noticeably smaller than Flickr30k/Multi30k, proves to show the sparsity of high-quality data in languages such as

Arabic compared to English. Our Arabic image captioning model was trained and evaluated on this data set.

2.2.3. Crossmodal-3600

In order to further test our models, we used the Crossmodal-3600 (XM3600) (Thapliyal et al., 2022) image captioning data set that was developed by Google specifically for evaluating and benchmarking image captioning models. This data set includes 3600 culturally diverse images from around the world each image with high-quality, human-annotated captions in 36 languages. More interestingly, for each of the 36 languages, there are 100 images in the data set that correspond to regions these languages are spoken. This data set comes with both raw and tokenized captions.

2.3. Evaluation Metrics

There are several qualitative evaluation metrics used in the task of image captioning. In this subsection we will discuss the qualitative metrics we will use in our paper namely, BLEU, ROUGE, METEOR, and CIDEr. Moreover, due to the complexity of the task of image captioning, performing qualitative feedback in addition to quantitative evaluation is strongly recommended. As such, we will also discuss how we are planning to apply qualitative evaluation in our paper.

2.3.1. BLEU

One of the most commonly used quantitative evaluation metrics for image captioning is the Bilingual Evaluation Understudy (BLEU) score. BLEU is originally a metric that was created to score machine-generated translations based on reference, however, it is also used in image captioning (Luo et al., 2022). In image captioning, BLEU evaluates a generated caption by using n-gram overlap between the generated and reference caption. BLEU is calculated as follows (Luo et al., 2022):

$$BLEU = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log\left(\text{precision}_n\right)\right)$$

where:

w_n is the weight of the n-gram.

BP (brevity penalty) is a factor that penalizes captions that are shorter than the reference captions. It is defined as $\exp\left(1 - \frac{\text{reference\_length}}{\text{translation\_length}}\right)$ if $translation_length > reference_length$, and 1 otherwise.

precision_n is the precision of n-gram n.

2.3.2. ROUGE

Another common evaluation metric that was initially developed for machine translation is the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004). Unlike BLEU, ROUGE is a recall-based evaluation metric (Luo et al., 2022). This means that for image captioning, ROUGE evaluates a generated caption by its ability to capture the most important information (co-occurrence of information) from a reference caption. The most common variation of ROUGE, ROUGE-L is calculated as follows (Lin, 2004):

$$ROUGE - L = \frac{\text{LCS}(C,\text{ref})}{\max(\text{len}(C),\text{len}(\text{ref}))}$$

where:

LCS is the longest common sequence between the generated caption S and reference caption ref.

2.3.3. METEOR

Metric for Evaluation of Translation with Explicit Ordering (METEOR) is another evaluation used in for image captioning. METEOR calculates the harmonic mean between the precision and recall between a generated caption and a reference caption using word alignment(Banerjee & Lavie, 2005). METEOR also takes into account synonyms, stemmed words, and WordNet-based word senses. It is calculated as follows (Banerjee & Lavie, 2005):

$$\text{METEOR} = \frac{(1-\alpha)\cdot \text{precision} \cdot \text{recall}}{\alpha \cdot \text{precision} + (1-\alpha)\cdot \text{recall}} \cdot (1 - \gamma \cdot \text{penalty})$$

where:

α controls the weighting of the precision and recall.

penalty is a penalty designed to take into account word order between reference and generated captions.

2.3.4. CIDEr

Of all of our quantitative evaluation metrics, Consensus-based Image Description Evaluation (CIDEr) is the only one that was specifically developed for evaluating image captioning models (Vedantam et al., 2015). Unlike the previous evaluation metrics which rely on some form of n-gram matching, CIDEr calculates the similarity between generated captions and reference captions by treating each sentence as a document and calculates the cosine angle of the word frequency-inverse document frequency (TF-IDF) vector (Vedantam et al., 2015). It is calculated as follows (Vedantam et al., 2015):

$$\text{CIDEr}n(c,S) = \frac{1}{M}\sum i = 1^{M} \frac{\left(\sum_{c' \in S_i} g_n(c) \cdot g_n(c')\right)}{(\|g_n(c)\|_2 \cdot \|g_n(S_i)\|_2)}$$

where:

c represents a generated caption.

S represents a set of reference captions.

n represents an n-gram to be evaluated.

M represents the number of reference captions.

g^n represents an n-gram-based TF-IDF vector.

2.3.5. Qualitative Evaluation

Although there are many different quantitative metrics for evaluating an image captioning model's output, it is often the case that these metrics do not accurately indicate the quality of some generated captions. As such, it is important

to also perform a qualitative evaluation of your model to ensure it is of good quality. For image captioning, this entails that humans (who are proficient in the language the output is in) manually inspect the output of the image captioning model and ensure that it accurately describes the input image and/or truthfully represents the reference caption.

3. Methodology

As mentioned in the previous sections, training the classical image captioning models usually requires a huge amount of data, which is not possible with languages other than English, to address this issue we decided to go with an approach similar to (Mokady et al., 2021), rather than training the model from scratch we use pretrained models for both the vision and language part, and a small neural network between those models to adapt the output from the vision part producing representations that the language model could use to output meaningful captions.

The final model consists of an Encoder in which we use a pretrained vision model called CLIP, a Decoder where an autoregressive model known as GPT-2 is used, and a transformer Adapter between those two models. The Encoder stays the same for different languages (same image representations for all languages) but the Decoder is loaded with the GPT-2 model pretrained with that specific language, and the Adapter is always trained from scratch.

3.1. Encoder

CLIP (Contrastive Language-Image pretraining) is a neural network introduced by (Mokady et al., 2021) that is trained on a large variety of image-text pairs instead of the usual predetermined Object categories task, meaning that the final model has a shared representation between image and text facilitating the captioning process.

Vision models trained on image-text pairs are not new but the contrastive approach taken to train CLIP makes it capable of "Zero-Shot" learning, here rather than trying to predict the words in the description of images the model tries to learn a multimodal (text-image)l embedding space by trying to increase the cosine similarity between an image encoder and a text encoder

Since our encoder is used only to produce the image representations, we only use the image encoder in CLIP (throwing out the textual Encoder) to produce the semantic-rich embeddings of the image which is guaranteed to contain all the essential visual data needed for captioning (no need for further training).

3.2. Decoder

To generate the captions we need to employ a textual decoder in the form of a language model. Language modeling is the process of determining the probability of a sequence of words occurring in a sentence (or the probability of the next word in a sentence) this is usually done using probabilistic methods or neural network models, probabilistic methods include the famous n-gram, but those methods became obsolete with the introduction of neural network language models, Neural Network architectures like LSTM (Long Short Term Memory) (Hochreiter & Schmidhuber, 1997) has been successfully used since the nineties for this purpose but due to their scaling limitations they were also replaced by the state-of-the-art Transformer based models.

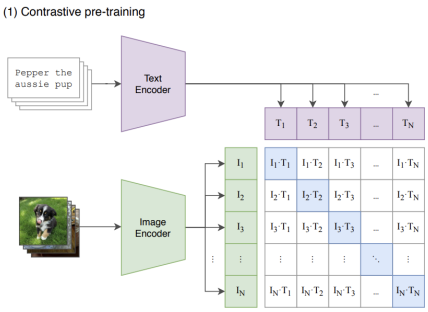

Figure 2. CLIP Contrastive training: CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples

A Transformer model is a Neural Network architecture that was introduced in the famous paper "Attention is all you need" (Vaswani et al., 2017), this architecture is based on the self-attention mechanism that allows the model to weigh the importance of different parts of the input text allowing for long-range dependencies between elements in a sequence, it also allows the input to be processed in parallel (no step-by-step processing).

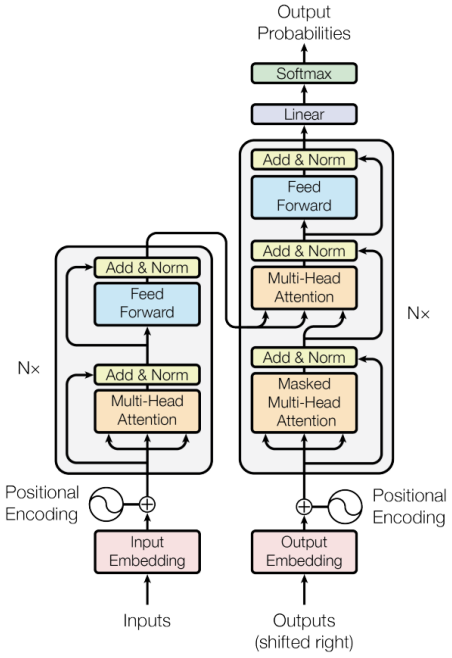

Figure 3. The Transformer - model architecture

With the help of the transformer architecture and the introduction of self-supervised learning as a transfer learning

technique, a new breed of language models was born called the pretrained language models, those models are trained in two stages the pretraining stage where the model is trained on a huge unsupervised corpus of sentences (in the order of hundreds of millions) using a self-supervised task, the output of this stage is a model that can produce sentence representations that reflect the syntax, context and, semantics of that sentence, this model is then finetuned (using a labeled dataset) in downstream tasks.

The model that we are using as our Decoder is a member of this family and it is called GPT-2 (Generative Pre-trained Transformer), introduced in the paper (Radford et al., 2019), This model has a decoder-only architecture Figure 4 (only the decoder is used in the Transformer architecture), a future mask is also applied to sequence to make sure that the model does not peak ahead (in training time) while generating the words. GPT-2 models are usually pretrained using a generic Langauge modeling task where the goal is to predict the next word in the sequence, after pretraining the model could be fine-tuned in downstream tasks or even used with "zero-shot" learning (Xian et al., 2018) and prompting (Han et al., 2021).

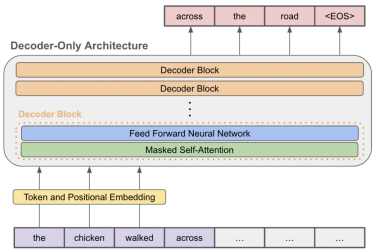

Figure 4. Decoder-only transformer architecture

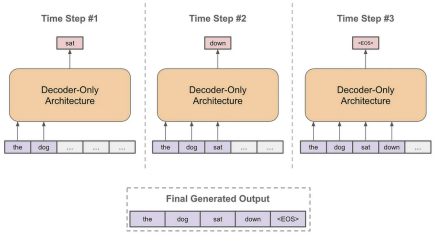

Figure 5. Future Mask in a Decoder Only Architecture

As mentioned before we will be using a different GPT-2 model for each language trained on a big unsupervised corpus in the language modeling task.

3.3. The Adapter

As mentioned before, both our Encoder and Decoder are pretrained and capable of producing semantic-rich image embeddings (Encoder) or a cohesive sequence of words (Decoder), The natural next step is to fine-tune both models in the image-captioning task, but this process is usually computationally expensive and requires a huge amount of data, instead, a Parameter-Efficient Transfer Learning method is used proposing a different kind of architecture modification to repurpose both the pretrained models for the new task, this modification includes injecting a small number of layers (called the Adapter modules) into the network and training the model where only those newly injected layers are updated and the rest of the network is frozen, this is different from the original fine-tunning method where the new top-layer and the original weights are co-trained (Houlsby et al., 2019). Adapter tunning could be done sequentially for each new task adding only a small number of parameters for each one.

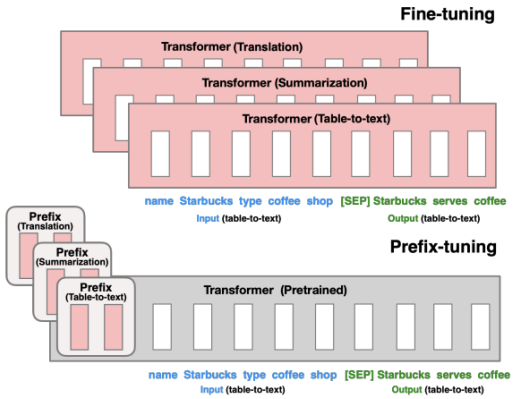

Figure 6. finetuning (top) Prefix tuning (bottom), the red blocks are the only layers that are updated during training

Since we are doing Natural Language Generation (NLG) we use an Adapter architecture specifically tailored for it called "Prefix-tuning" proposed by (Li & Liang, 2021). This method is optimized for tasks where text is generated sequentially in many steps Figre 6 which makes it perfect to be used with GPT-2, Prefix-tuning prepends a sequence of continuous task-specific vectors to the input, which is called a prefix, depicted by red blocks in Figre 6, when generating a new sequence the model looks at the prefix first which contains the trained parameters, this means that we don't need to train the whole network for each task (only the prefix) which significantly reduces the number of task-specific parameters that need to be stored.

3.4. How it all fits together

Our training dataset consists of image-text pairs, where the text is a descriptive caption of the image, in the first stage we encode each image using CLIP encoder, which will produce semantic-rich embeddings of these images, those embeddings are then adapted by the trained prefix Adapter to produce task-specific image captioning representations, those representations are then used as prefixes to the GPT-2 model sequence that generates the final captions (see figure 7).

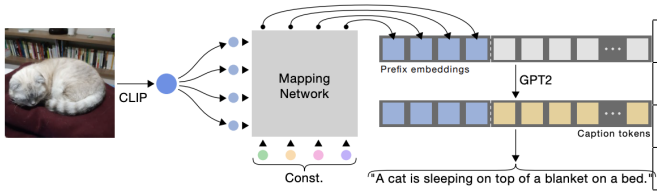

Figure 7. the Encoder-prefix Adapter-Decoder Architecture

Model	Dataset	GPT-2 Version	Tokenizer Vocab Size
English	Multi30K English	"gpt2"	50257 Tokens
German	Multi30K English	"anonymous-german-nlp/german-gpt2"	52000 Tokens
Arabic	Flickr8K-Arabic	"elgeish/gpt2-medium-arabic-poetry"	64000 Tokens

Table 1. Datasets, Tokenizers and GPT-2 fine-tuned versions details used for the training each of our models

4. Experiments

4.1. Motivation

Our main goal is to verify that the method described in the previous section can produce good results (captions) when trained on smaller datasets (order of tens of thousands rather than millions), where those smaller datasets are representative of the available captioning corpora for all languages other than English. We also want to examine the quality of the captions produced when the model is trained in Deutsch and Arabic and compare it to the English-trained model with a similar dataset size.

4.2. Baselines

We compare our model against two different baselines from the literature, covering both old methods trained on small datasets and newer models that were only trained on a huge corpus, we expect our model to produce results better than the first baseline (small data old) and comparable to the second baseline (large data new) across all languages.

First baseline: we decided to compare our solution against the models reported by (Xu et al., 2015), as it constitutes the best results (that we could find) produced using a relatively small captioning dataset (Flicker8K and Flicker30K), moreover we are using the same corpus split (Flicker30K) in training our English model and a human-translated version of it in Deutsch (Multi30K) and Arabic (Flicker8K-Arabic). The method used by (Xu et al., 2015) was SOTA at that time for image captioning, deploying a CNN Encoder and an RNN Decoder adding an attention layer between the encoder and decoder. They also used BLEU scores and METEOR which we also benchmark our model against.

Second baseline: Is taken from the original paper (Mokady et al., 2021) that we adopted our solution from. They use the same architecture but only train on huge (English only) corpora(in the order of millions), they benchmark their model against three different datasets: Conceptual Captions(Sharma et al., 2018), nocaps (Agrawal et al., 2019), and COCO (Lin et al., 2014). They also used several evaluation metrics including ROUGE, CIDEr, BLEU, and METEOR.

It's also worth mentioning that we tried to train a CNN-RNN model (without attention) on our small dataset, but the model did not have enough capacity to learn the image-caption alignment producing random words for the first few epochs and then memorize a caption (for more details please see Appendix).

4.3. Experimental Setup

Prior to model training, we pre-processed our datasets by generating embeddings for all images and saving them separately to optimize training time and memory usage. Additionally, we tokenized all captions and stored the tokenized versions. All three models utilized the same CLIP model for embedding images, while each had its own tokenizer tailored to its respective language 1.

To speed up the training process of our Transformer network and minimize the effect of weight deviation, we have employed specific parameters, as suggested by (Mokady et al., 2021). Specifically, we have utilized a learning rate of 0.01, Adam optimizer with decay, and a linear scheduler featuring 1000 warmup steps. We trained our models for 30 epochs, at which point the training loss stabilized across all models.

During each training batch, our adaptor networks converted the CLIP embeddings to GPT-2 word embeddings with a prefix length of 10, which has been demonstrated to be advantageous for LLM adaptation by (Li & Liang, 2021). Following prefix generation, the GPT-2 model generated the target caption token by token based on the prefix, and we computed loss based on cross-entropy between the reference caption tokens and the generated caption tokens. This setup remained consistent across all models, with variations only in the GPT-2 model versions, tokenizers, and datasets as detailed in table 1.

4.4. Results & Discussion

In this section, we will present, highlight, and interpret the results we obtained for our models compared to our baselines. Additionally, we will discuss how dataset size affects the performance of our Arabic model. Finally, we will perform a qualitative evaluation of our English model using the XM3600 dataset.

4.4.1. Interpretation

Table 2 depicts the results of our models (formatted as Ours (Language)) while additionally showing the performance of the baselines we mentioned before. It can be observed from the table that within the context of the Multi30K/Flickr30k dataset, our English model outscores the large, computationally intensive baseline models which utilize soft and

Dataset	Model	BLEU	ROUGE-1	ROUGE-L	CIDEr	METEOR
Flickr30k	Soft Attention (Xu et al., 2015)	19.1	-	-	-	18.49
	Hard Attention (Xu et al., 2015)	19.9	-	-	-	18.46
	Ours (English)	16.35	**31.29**	**28.4**	33.09	19.41
	Ours (German)	12.83	26.02	24.19	23.46	16.0
Flickr8k	Soft Attention (Xu et al., 2015)	19.5	-	-	-	18.93
	Hard Attention (Xu et al., 2015)	21.3	-	-	-	20.30
	Ours (Arabic)	14.17	18.48	18.06	26.35	14.67
COCO	Transformer (Mokady et al., 2021)	**33.53**	-	-	**113.08**	**27.45**
Conceptual Captions	Transformer (Mokady et al., 2021)	-	-	25.12	71.82	-

Table 2. Our models' performance compared to the baseline models based on the discussed quantitative evaluation metrics.

hard attention in terms of the METEOR score (19.41 vs. 18.46 & 18.49). However, our model still achieves lower BLEU scores (16.35 vs 19.9 & 19.1) than the baseline. Considering the way METEOR and BLEU are computed, this suggests that our model may have better paraphrasing capabilities than the baseline models and is able to generate more diverse captions. On the other hand, the lower BLEU score indicates that our English model is worse at generating captions that utilize exact n-gram matches with the reference captions. It is also important to note that our German language model performs worse than the baseline models in terms of both BLEU and METEOR. However, since the baseline models are all in English directly comparing them to the German model using the same-sized dataset may not be very insightful.

Furthermore, table 2 shows that our Arabic language model which is trained on the even smaller translated Flickr8k also achieves a lower score (in both BLEU and METEOR) compared to the soft/hard attention model trained on the English counterpart of that dataset. Again, this comparison between these models may not be very accurate because of the language difference, e.g., METEOR takes into account vocabulary and grammar rules, which are very different in Arabic and English.

Moving on to our second set of baselines: computationally efficient models trained on large datasets, namely the transformer models, there are several observations. First, the transformer model trained on the COCO dataset outperforms all other models, including ours, on all the metrics it was evaluated on: BLEU, CIDEr, and METEOR. This is expected since the COCO dataset is much larger than the Flickr30k/Flickr8k datasets we trained our models on. Interestingly enough, our English language model outscores the transformer model trained on the much larger Conceptual Captions dataset on the ROUGE-L metric (28.4 vs 25.12), however, the transformer model achieves a much higher CIDEr score (33.09 vs. 71.82). With the way both metrics are computed, this may suggest that our model maybe have a better understanding of the overall structure and meaning of the reference captions but is not able to produce captions that contain similar n-grams to the ones found in the reference captions.

All in all, our results suggest that using resource-efficient models, in terms of both computation and data size, can still produce results that can match or even outperform models that are either computationally or data-intensive in some aspects. However, what the results also highlight, is the unreliability and inconsistency of qualitative evaluation metrics for image captioning, which is why conducting qualitative evaluation is rather important in this domain.

4.4.2. Dataset Size Analysis

Figure 8. A comparison of how the loss changes through training steps as we increase the size of the dataeset

Figure 9. A comparison of how the final mean loss per epoch changes as we increase the size of the dataest

In order to investigate the impact of dataset size on the performance of our Arabic captioning model and its ability to learn from limited samples, we conducted a series of experiments. We randomly sampled varying percentages of the training data, ranging from 10% to 100%, and trained the model on each of these subsets. Figure 8 illustrates the relationship between dataset size and model performance, as measured by the loss function. We observe that the model trained on only 10% of the data was unable to achieve good mapping, as evidenced by its consistently high loss throughout the training process. However, increasing the dataset size to 30% resulted in a similar performance to that of the model trained on 80% of the data, indicating the model's robustness against data sparsity.

Figure 9 provides further support for this conclusion, showing that the mean loss of the models trained on 30% to 80% of the data was comparable, with only slight improvements seen when using the full training dataset. These results highlight the importance of dataset size in the training of AI models and suggest that our image captioning models can effectively learn from limited samples, making it a promising solution for real-world applications with limited data availability.

4.4.3. Qualitative Evaluation

The qualitative analysis section investigates a group of sample captions produced by our English models from images in the XM3600 dataset. The choice of the English model was made to ensure better readability, and the XM3600 dataset was selected because of its high-quality and diverse images. The captions were evaluated based on their F1 RougeL score to determine good and bad examples as shown in figure 10.

The first example serves as an outstanding demonstra-

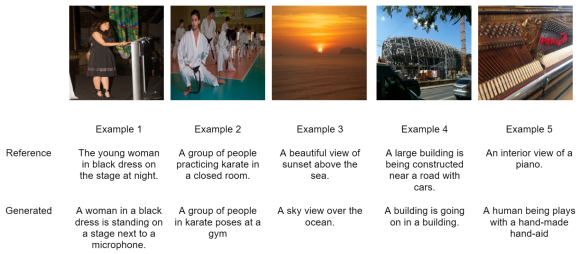

Figure 10. Sample images from the XM3600 along with their reference English captions and captions generated by our model

tion of the model's captioning capabilities. The generated caption accurately describes all critical information in the image, showcasing the adaptor's ability to translate information from CLIP to GPT accurately. The second example is also an excellent illustration of the model's ability to describe common objects and their relationships in a simple manner.

However, the third example highlights some of the challenges inherent in image captioning. Although the generated caption is acceptable to a human evaluator, it received a low F1 score because it differed from the reference caption. This reflects the intrinsic ambiguity in image captioning and the limitations of automatic evaluation metrics (Kasai et al., 2021).

Example 4 illustrates a failure mode of our model. The generated caption is grammatically incorrect, and our analysis indicates that this occurs for two reasons. Firstly, the adapter fails to convey all the information from the image to the language model, as in this case. Some works in image captioning involve using graph-based image encoders to address this limitation. Secondly, when the image contains uncommon objects, known as long-tail objects (Changpinyo et al., 2021), the generated caption may be inadequate due to limitations in world knowledge throughout the image captioning system, not just the adapter.

5. Related work

Visual Encoding in image captioning can be achieved through various methods (Stefanini et al., 2021), including non-attentive methods based on global CNN features(Vinyals et al., 2014), additive attentive methods that embed the visual content using either grids or regions (Dai et al., 2018), graph-based methods adding visual relationships between visual regions (Yao et al., 2018), and self-attentive methods that employ Transformer-based paradigms (Yang et al., 2019). Global CNN features are used to extract high-level representations, and they can be employed as conditions or at each time step. However, they lack granularity and may compress too much information, making it difficult to produce specific and fine-grained descriptions. To address this issue, most recent approaches have increased the granularity level of visual encoding by using attention over grids of CNN features. In some studies, graphs built over image regions are used to enrich the representation by including semantic and spatial connections between objects. Self-attention encoding has also been proposed as a method for visual encoding. This approach uses a pre-trained Vision Transformer network as an encoder and a standard Transformer decoder to generate captions. This method has been used by subsequent captioning approaches, such as CLIP-based features.

Language models for image captioning are primarily categorized as LSTM-based and Transformer-based approaches. LSTM-based models, proposed by (Vinyals et al., 2014), use a single-layer LSTM with visual encoding as the initial hidden state to generate output captions. Transformer-based models (Wang et al., 2021; Herdade et al., 2019) use a standard decoder with masked self-attention and cross-attention operations, followed by a feed-forward network. During training, a masking mechanism is applied to the previous words to ensure a unidirectional generation process.

Multilingual image captioning involves generating image captions in a target language given a source language information which is normally English. (Elliott et al., 2017) studied this task using the Multi30K dataset, supplemented with out-of-domain data due to limited samples. (Jaffe, 2017) and (Thapliyal & Soricut, 2020) both used a multi-stage approach, with (Jaffe, 2017) using 2 LSTM decoders to generate English and German captions sequentially, while (Thapliyal & Soricut, 2020) used a multistage Transformer decoder conditioned on both the image and English caption to generate captions in multiple languages.

Pretrained large vision-language models use vision-and-language pre-training to create a shared latent space for both vision and text, and can be used as encoder-decoder models. (Yang et al., 2021) use object detector visual tokens as a prefix to caption tokens to pre-train a BERT architecture for prediction. CLIP is used by (Mokady et al., 2021) as the image encoder and GPT-2 prefix to address the lack of joint representation of language and vision. Both image and text embeddings are used in the work of (Li et al., 2023) from CLIP to train a decoder from scratch, instead of adapting to different domains.

6. Conclusions

In conclusion, our approach of using pre-trained large models and a transformer network to train image captioning networks in languages other than English has shown promising results. We have demonstrated that our method can overcome the data scarcity problem and achieve good quantitative and qualitative performance.

In the future, we plan to extend our approach to other languages and datasets and evaluate our method on more diverse benchmarks. We also hope to investigate the use of other pre-trained models and explore ways to incorporate external knowledge and context to improve the quality of the generated captions.

References

Agrawal, Harsh, Desai, Karan, Wang, Yufei, Chen, Xinlei, Jain, Rishabh, Johnson, Mark, Batra, Dhruv, Parikh, Devi, Lee, Stefan, and Anderson, Peter. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Banerjee, Satanjeev and Lavie, Alon. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Changpinyo, Soravit, Sharma, Piyush Kumar, Ding, Nan, and Soricut, Radu. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567, 2021.

Cho, Jaemin, Yoon, Seunghyun, Kale, Ajinkya, Dernoncourt, Franck, Bui, Trung, and Bansal, Mohit. Fine-grained image captioning with clip reward. *Findings of the Association for Computational Linguistics: NAACL 2022 - Findings*, pp. 517–527, 5 2022. doi: 10.18653/v1/2022.findings-naacl.39. URL https://arxiv.org/abs/2205.13115v1.

Dai, Bo, Ye, Deming, and Lin, Dahua. Rethinking the form of latent states in image captioning. In *European Conference on Computer Vision*, 2018.

ElJundi., Obeida, Dhaybi., Mohamad, Mokadam., Kotaiba, Hajj., Hazem, and Asmar., Daniel. Resources and end-to-end neural network models for arabic image captioning. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pp. 233–241. INSTICC, SciTePress, 2020. ISBN 978-989-758-402-2. doi: 10.5220/0008881202330241.

Elliott, Desmond, Frank, Stella, Sima'an, Khalil, and Specia, Lucia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL https://aclanthology.org/W16-3210.

Elliott, Desmond, Frank, Stella, Barrault, Loïc, Bougares, Fethi, and Specia, Lucia. Findings of the second shared task on multimodal machine translation and multilingual image description. *WMT 2017 - 2nd Conference on Machine Translation, Proceedings*, pp. 215–233, 10 2017. doi: 10.18653/v1/w17-4718. URL https://arxiv.org/abs/1710.07177v1.

Freitag, Markus, Grangier, David, and Caswell, Isaac. Bleu might be guilty but references are not innocent. *ArXiv*, abs/2004.06063, 2020.

Han, Xu, Zhang, Zhengyan, Ding, Ning, Gu, Yuxian, Liu, Xiao, Huo, Yuqi, Qiu, Jiezhong, Yao, Yuan, Zhang, Ao, Zhang, Liang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

Herdade, Simao, Kappeler, Armin, Boakye, Kofi, and Soares, Joao. Image captioning: Transforming objects into words. In *Neural Information Processing Systems*, 2019.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 05 2013. doi: 10.1613/jair.3994.

Houlsby, Neil, Giurgiu, Andrei, Jastrzebski, Stanislaw, Morrone, Bruna, De Laroussilhe, Quentin, Gesmundo, Andrea, Attariyan, Mona, and Gelly, Sylvain. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Jaffe, Alan. Generating image descriptions using multilingual data. In *Conference on Machine Translation*, 2017.

Kasai, Jungo, Sakaguchi, Keisuke, Dunagan, Lavinia, Morrison, Jacob Daniel, Bras, Ronan Le, Choi, Yejin, and Smith, Noah A. Transparent human evaluation for image captioning. In *North American Chapter of the Association for Computational Linguistics*, 2021.

Li, Wei, Zhu, Linchao, Wen, Longyin, and Yang, Yi. Decap: Decoding clip latents for zero-shot captioning via text-only training. *ArXiv*, abs/2303.03032, 2023.

Li, Xiang Lisa and Liang, Percy. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

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

Lin, Chin-Yew. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Luo, Gaifang, Cheng, Lijun, Jing, Chao, Zhao, Can, and Song, Guozhu. A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Processing*, 16(2):311–332, 2022. doi: https://doi.org/10.1049/ipr2.12367. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12367.

Mokady, Ron, Hertz, Amir, and Bermano, Amit H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, Sutskever, Ilya, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9, 2019.

Sharma, Piyush, Ding, Nan, Goodman, Sebastian, and Soricut, Radu. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Stefanini, Matteo, Cornia, Marcella, Baraldi, Lorenzo, Cascianelli, Silvia, Fiameni, Giuseppe, and Cucchiara, Rita. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:539–559, 2021.

Suresh, K. Revati, Jarapala, Arun, and Sudeep, P. V. Image captioning encoder–decoder models using cnn-rnn architectures: A comparative study. *Circuits, Systems, and Signal Processing*, 41(10):5719–5742, Oct 2022. ISSN 1531-5878. doi: 10.1007/s00034-022-02050-2. URL https://doi.org/10.1007/s00034-022-02050-2.

Thapliyal, Ashish V. and Soricut, Radu. Cross-modal language generation using pivot stabilization for web-scale language coverage. *ArXiv*, abs/2005.00246, 2020.

Thapliyal, Ashish V., Pont-Tuset, Jordi, Chen, Xi, and Soricut, Radu. Crossmodal-3600: A massively multilingual multimodal evaluation dataset, 2022.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vedantam, Ramakrishna, Zitnick, C. Lawrence, and Parikh, Devi. Cider: Consensus-based image description evaluation, 2015.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, D. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2014.

Wang, Zirui, Yu, Jiahui, Yu, Adams Wei, Dai, Zihang, Tsvetkov, Yulia, and Cao, Yuan. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2021.

Xian, Yongqin, Lampert, Christoph H, Schiele, Bernt, and Akata, Zeynep. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41 (9):2251–2265, 2018.

Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron, Salakhudinov, Ruslan, Zemel, Rich, and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Yang, Xu, Zhang, Hanwang, and Cai, Jianfei. Learning to collocate neural modules for image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4249–4259, 2019.

Yang, Zhengyuan, Gan, Zhe, Wang, Jianfeng, Hu, Xiaowei, Ahmed, Faisal, Liu, Zicheng, Lu, Yumao, and Wang, Lijuan. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, 2021.

Yao, Ting, Pan, Yingwei, Li, Yehao, and Mei, Tao. Exploring visual relationship for image captioning. In *European Conference on Computer Vision*, 2018.

Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 02 2014. ISSN 2307-387X. doi: 10.1162/tacl_a_00166. URL https://doi.org/10.1162/tacl_a_00166.

7. Appendix

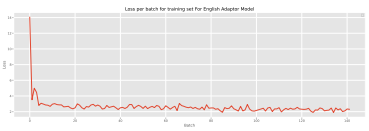

Figure 11. Healthy Train Loss patterns when training our English Adaptor Model

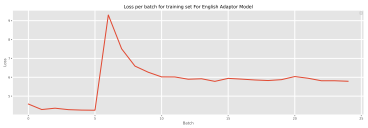

Figure 12. Unstable divergent Loss when training the CNN-RNN baseline on the small datasets

Model	BLEU	ROUGE-1	ROUGE-L	CIDEr	METEOR
English	**15.11**	**27.93**	**30.4**	**17.27**	**26.19**
German	11.78	11.78	12.98	12.29	9.86
Arabic	-	2.47	2.54	3.58	1.31

Table 3. Results of our models in the three different languages obtained on the CrossModal-3600 evaluation dataset.

Figure 13. Sample Captions from out German Models

Figure 14. Sample Captions from out Arabic Models