Journal of Ambient Intelligence and Humanized Computing (2023) 14:7719–7731
https://doi.org/10.1007/s12652-023-04584-y

ORIGINAL RESEARCH

Generative image captioning in Urdu using deep learning

Muhammad Kashif Afzal[1] · Matthew Shardlow[2] · Suppawong Tuarob[3] · Farooq Zaman[1] · Raheem Sarwar[2] · Mohsen Ali[1] · Naif Radi Aljohani[5] · Miltiades D. Lytras[5] · Raheel Nawaz[4] · Saeed-Ul Hassan[2]

Received: 22 April 2022 / Accepted: 20 February 2023 / Published online: 10 April 2023

Abstract
Urdu is morphologically rich language and lacks the resources available in English. While several studies on the image captioning task in English have been published, this is among the pioneer studies on Urdu generative image captioning. The study makes several key contributions: (i) it presents a new dataset for Urdu image captioning, and (ii) it presents different attention-based architectures for image captioning in the Urdu language. These attention mechanisms are new to the Urdu language, as those have never been used for the Urdu image captioning task (iii) Finally, it performs quantitative and qualitative analysis of the results by studying the impact of different model architectures on Urdu's image caption generation task. The extensive experiments on the Urdu image caption generation task show encouraging results such as a BLEU-1 score of 72.5, BLEU-2 of 56.9, BLEU-3 of 42.8, and BLEU-4 of 31.6. Finally, we present data and code used in the study for future research via GitHub (https://github.com/saeedhas/Urdu_cap_gen).

Keywords Image captioning · Information retrieval · Natural language processing · Urdu · Deeplearning

1 Introduction

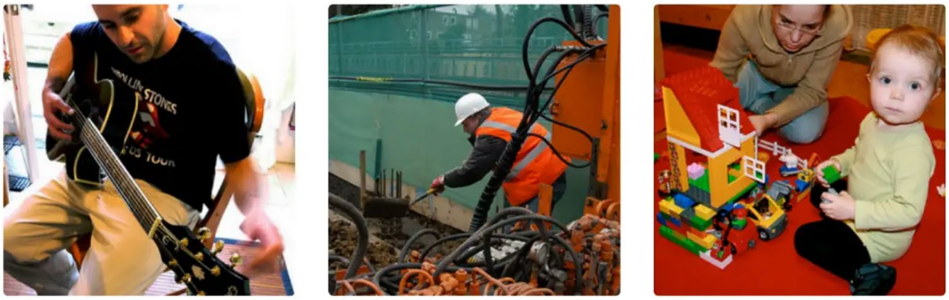
The image captioning task aims at describing the contents of an image in natural language (Mishra et al. 2021), which can be accomplished by combining Computer Vision techniques with Natural Language Processing methods. The general idea of image captioning system is encoding input image into a vector using computer vision techniques and then decoding that vector into words using any decoder from NLP language models. An Example of image caption is illustrated in Fig. 1. Figures are the input of the image captioning system and the captions are the output. Benchmark image captioning datasets for English include Flickr8K (Hodosh et al. 2013) , NOCAPS (Agrawal et al. 2019) and MSCOCO (Lin et al. 2014). Since natural language generation is key part of the captioning system, BLUE score is considered as the common evaluation metric (Papineni et al. 2002)

The applications of this task are wide and varied, including but not limited to: assisting visually impaired individuals to surf the web (Makav and Kılıç 2019; Fisch et al. 2020; Liu et al. 2020), enhancing image search with semantic information (Lindh et al. 2020), navigating video scenes (Wang et al. 2020; Zhou et al. 2020a), or even enabling AI driven cars to better understand their environment (Kim et al. 2018; Xu et al. 2015; Zhou et al. 2020b).

Inspired by prior work (Bahdanau et al. 2015), Xu et al. (2015) proposed a model based on visual attention, trained in a deterministic manner using standard back-propagation techniques and additionally learning to soft attend on objects as well as non-objects (semantics) while generating the corresponding tokens in the output sequence. Their model produced state-of-the-art performance on three benchmark datasets: Flickr8k, Flickr30k and MS COCO (Young et al. 2014). Later on, Aneja et al. (2018) achieved a similar score by using a purely convolutional architecture, replacing LSTM, with feed-forward masked convolutions to restrict the convolution operations to use only the past words' information. Vinyals et al. (2015) and Huang et al. (2019) proposed an "attention on attention" (AoA) module, which extends the conventional attention mechanisms to determine

✉ Raheem Sarwar
R.Sarwar@mmu.ac.uk

[1] Information Technology University, Lahore, Pakistan
[2] Manchester Metropolitan University, Manchester, UK
[3] Mahidol University, Nakhon Pathom, Thailand
[4] Staffordshire University, Stoke-on-Trent, UK
[5] King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia


"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Fig. 1 Introductory examples of image captioning. https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

the relevance between attention results and current context. Applying AoA to both the encoder and the decoder of the image captioning model achieved new state-of-the-art (SOTA) results (Wang et al. 2022).

1.1 Research objectives and our contributions

Urdu is an Indo-Aryan language that borrowed a large percentage of its vocabulary from other languages such as Arabic and Persian (Amjad et al. 2020). The Ethnologue, a well-known reference source that publishes statistics on living languages, has ranked Urdu as the 11^{th} most spoken language in the world in 2020. It is also widely acknowledged as a major South Asian language, with 490 million native speakers worldwide (Shaik and Venkatramaphanikumar 2021). It is the official language of five Indian states, including Bhiar, Uttar Pradesh, and Jharkhand. It is the national language of Pakistan, which has a population of about 220 million people. According to the 2011 census of linguistic statistics conducted by the Indian government, India had 50,772,631 Urdu speakers. Urdu speakers can also be found in the United Kingdom, the United States, Canada, Australia, the Middle East, and Europe.

It uses Arabic script in cursive format (Nastaliq style) with the segmental writing system. Specifically, the Urdu language is based on an "*abjad*" system where the long vowels and consonants are necessarily written while the short vowels (diacritics) are optional. It is a bidirectional language where the numerals are written from left-to-right, while the characters are written from right-to-left. When characters are joined to make the words, they develop different shapes based on the context. Specifically, a character can have a maximum four shape variants known as initial, medial, final and isolated. The characters that can develop all four shapes are known as joiners, while the characters that can only have two shapes (final and isolated) are known as non-joiners (Kanwal et al. 2020).

Unlike English, a white space character is not considered as a reliable word boundary indicator in Urdu. That is, Urdu does not have consistent word boundary markings. For example, a writer may insert a space within a word قابل.احترام (respectable) in oder to make it visually correct, where the character . represents the ASCII space character. If the writer omits the space it may lead to an incorrect visual form قابلاحترام of the same word. Contrarily, the writer may omit space between two words اردوزبان (Urdu language) because the shape of characters with or without space remains the same. That is, the Urdu words ending with non-joiner characters exhibit correct shape even without space. Consequently, a writer may omit space between words ending with non-joiner characters. Most existing studies on generative image captioning are focused on English. To the best of our knowledge, no such published work exists in the realm of neural image caption generation for Urdu. Urdu is a low-resource and more morphologically complex language than English (Mahmood et al. 2020; Malik et al. 2021).

Urdu is often regarded as a low-resource language due to the lack of or inadequacy of various critical resources, such as gold standard datasets and fundamental natural language processing (NLP) toolkits, such as reliable tokenizers and stemmers (Shaik and Venkatramaphanikumar 2021). Our discussion, however, is focused on the limitations of Urdu in the image captioning task, Some key limitations are as follows.

- **Lack of attention.** Image captioning task has been extensively investigated for resource-rich languages such as English. To the best of our knowledge, no such published work exists in the realm of neural image caption generation for Urdu. Urdu is a low-resource and more morphologically complex language than English (Mahmood et al. 2020; Malik et al. 2021).
- **Unavailability of resources.** Author gender identification is an important NLP task. However, as mentioned earlier, this is the first study on generative image captioning in Urdu and there is no existing corpus available to perform this task. Therefore in this paper we introduced a new corpus to perform this task.

Our contributions. The contributions of this work are as follows:

- We present a new dataset for Urdu image captioning which can be accessed via GitHub.[1]
- We also discuss different types of attention-based architectures for image captioning in the Urdu language. These attention mechanisms are new for the Urdu language, as those have never been used for the Urdu image captioning task.
- Further, we illustrate quantitative and qualitative analysis of the results - studying the impact of differing model architectures on the image caption generation task in Urdu.
- Finally, we show that the best model achieves a BLEU-1 score of 72.5, BLEU-2 of 56.9, BLEU-3 of 42.8, and BLEU-4 of 31.6 on the Urdu image caption generation task.

The rest of the paper is organized as follows. Section 2 reviews the existing image captioning techniques. Section 3 discusses methodology and experimental setup. Section 4 presents the experimental results. Section 5 presents the conclusions and future work directions.

2 Literature review

The image captioning techniques can be organized into extractive and generative techniques. More details on extractive and generative captioning is provided in the following paragraphs.

2.1 Extractive captioning

Earliest approaches rely on hand-engineered features for visual elements and rule-based systems for language models. Some progress was reported using human-engineered templates and piecing together the phrases containing detected objects. Hodosh et al. (2013) treated the sentence-based image annotation as a ranking problem mapped to a given pool of captions. Whereas, several studies formulated this task as a retrieval problem and proposed solutions which represent embedding of images and text in the same space (Gong et al. 2014; Li et al. 2020; Zhou et al. 2020a). Socher et al. (2014) used deep learning to co-embed image and sentences together and Karpathy et al. (2014) embedded image sub-regions and sub-sentences jointly. Regional attributes have been used in many image captioning methods to alleviate the issues with predetermined caption templates. Farhadi et al. (2010) proposed detections to infer a triplet of image regions to return the suitable text by filling in a textual template. Li et al. (2011) used object detections and then piece together a final description using phrases containing detected objects, modifiers and locations using web-scale n-grams. Yao et al. (2010) introduced the web-ontology-language based on semantic representation produced as a result of parsing images, which is converted to human readable text. Kulkarni et al. (2013) used detection beyond triplets but with template-based text generation. The advantage of using the template-based methods is that the resulting captions tend to be grammatically correct. However, they use hard-coded visual concepts and hence suffer to produce the required variety in the output. Kuznetsova et al. (2014) extracted similar images relevant to the query image, then extracted noun verb and prepositional phrases from captions of those images. Eventually they run an object detector on the query image and compose captions using detected objects by pairing them with relevant captions of previously fetched images.

2.2 Generative captioning via deep learning

In contrast to the aforementioned dual stage methods, the recent trend for image to text generation is to use deep learning based encoder-decoder architectures that connect a CNN to an RNN to learn the mapping from images to sentences without involving any rules or human engineered features. For example, Mao et al. (2014), proposed a multimodal RNN (m-RNN) to estimate the probability distribution of the next token given previous tokens and the deep CNN feature of an image at each time step. Similarly, Kiros et al. (2014) constructed a joint embedding space using a more effective approach i.e. deep CNN model to encode image and a long short-term memory (LSTM) model encodes the text. Karpathy et al. (2014) also proposed a multimodal RNN generative model, but in contrast to Mao et al. (2014), their RNN is conditioned on the image information only at the first time step. The first landmark paper that reported tangible results was by Vinyals et al. (2015) combined deep CNNs for image classification with an LSTM for sequence modelling, to create a single network that generates descriptions of images. Chen and Lawrence Zitnick (2015) learn a bi-directional mapping between images and their sentence-based descriptions, which additionally enables reconstruction of visual features when given a caption as input. Tanti et al. (2017, 2018) conjectured that in a CNN-RNN setting for image caption generation, the image information can be fed to the neural network either by directly incorporating it in the RNN i.e. conditioning the language model (LM) by 'injecting' or in a layer following the RNN i.e. conditioning the LM by 'merging' image features where the later allows the RNN's hidden state vector to shrink in size by up to four times.

[1] https://github.com/saeedhas/Urdu_cap_gen

Table 1 Summary of recent image captioning models for English

Model	BLEU-1	BLEU-4
VIN VL Zhang et al. (2021)	82.0	41.0
UNIFIED VLP Zhou et al. (2020a)	80.9	39.5
X-Transformer Pan et al. (2020)	80.9	39.7
Attention Model by (Xu et al. 2015)	67	21.3
Merge Model (Tanti et al. 2018)	60	17.8
SOTA Model (Wang et al. 2022)	77.4	37.2

Their results suggest that the visual and linguistic modalities for caption generation need not be jointly encoded by the RNN since it yields large, memory-intensive models with few tangible advantages in performance; rather, the multimodal integration should be delayed to a subsequent stage.

2.3 Attention driven generative captioning

Bahdanau et al. (2015) proposed the soft attention mechanism for machine translation that produced revolutionary results by generating the target language tokens conditioning the LM on previous prediction by learning to shift and pay attention to parts of the source sentence representation. Inspired by prior work (Bahdanau et al. 2015), Xu et al. (2015) proposed a model based on visual attention, trained in a deterministic manner using standard back-propagation techniques and additionally learning to soft attend on objects as well as non-objects (semantics) while generating the corresponding tokens in the output sequence. Their model produced state-of-the-art performance on three benchmark datasets: Flickr8k, Flickr30k and MS COCO (Young et al. 2014). Later on, Aneja et al. (2018) achieved a similar score by using a purely convolutional architecture, replacing LSTM, with feed-forward masked convolutions to restrict the convolution operations to use only the past words' information. Vinyals et al. (2015) and Huang et al. (2019) proposed an "attention on attention" (AoA) module, which extends the conventional attention mechanisms to determine the relevance between attention results and current context. Applying AoA to both the encoder and the decoder of the image captioning model achieved new state-of-the-art results (Table 1).

3 Methodology and experimental setup

We chose ResNet-101 (He et al. 2016) to act as an encoder and a LSTM as a decoder. We have used two encoder-decoder architectures; (i) The Merge Model (Tanti et al. 2018) as a baseline and (ii) the Attention driven Context based Model (Xu et al. 2015) as our main model as shown in Fig. 2.

3.1 Dataset

To prepare the image mapped Urdu dataset we make use of the Flickr8K (Hodosh et al. 2013) dataset for cross-reference which is a standard dataset and widely used by the research community to perform image caption generation tasks for English (Hodosh et al. 2013). The Flickr8K dataset comprises 8000 images where each image is presented with 5 English captions on average. We have selected a subset of data from the Flickr8K dataset consisting of five English captions per image; these were manually translated into Urdu by a native speaker followed by several rounds of quality control involving another native speaker of Urdu.

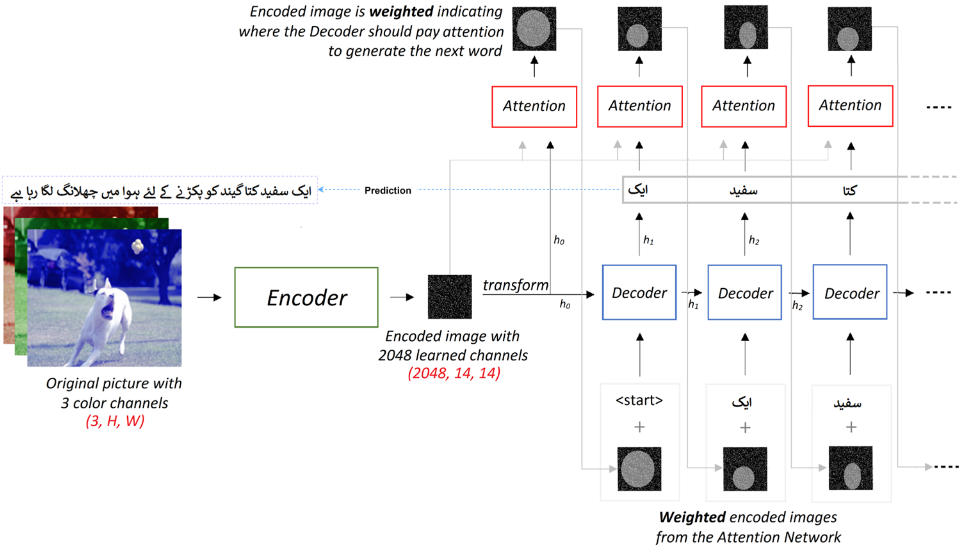

Encoded image is **weighted** indicating where the Decoder should pay attention to generate the next word
Attention
Attention
Attention
Attention
....
ایک سفید کتا گیند کو پکڑنے کے لئے ہوا میں چھلانگ لگا رہا ہے
Prediction
ایک
سفید
کتا
h_0
h_1
h_2
transform
h_0
Encoder
Decoder
Decoder
Decoder
h_1
h_2
....
Original picture with 3 color channels (3, H, W)
Encoded image with 2048 learned channels (2048, 14, 14)
<start>
ایک
سفید
+
+
+
....
Weighted encoded images from the Attention Network

Fig. 2 Caption prediction using attention driven Inject model

We select 1800 images from Flickr8K and translate 5 captions for each, thus producing 9000 Urdu captions. We call this dataset Dogs Flickr8K (see section Appendix for more details).

3.2 Model training

The data is randomized and split into 1440 images as train set, 180 as validation set and 180 as test set. Each image has five captions, such that it results in a corresponding split of 7200 train, 900 validation and 900 test captions.

For the encoder of our baseline model, we remove the last classification layer 'FC' to harness the image feature vector from the second last fully connected layer. However for our main model, based on attended annotation vectors, we make use of spatial context. We strip-off the trailing layers after convolutions i.e. pooling and fully connected (dense) layers to obtain the 3D tensor as an image feature set by adaptive average pooling the output of the last convolutional layer. This 3D feature set, 2048 layered 14x14 tensor, is flattened to a 2D representation of 196 annotation vectors each of size 2048 which is attended to by enhancing the relevant weight.

To initialize the language model (LSTM), annotation vectors are first averaged to produce a single vector of size 196 that is projected using two independent fully connected layers of neurons to the cell state size (512) and hidden state size (512). Soft attention is deterministic and a differentiable function comprising MLPs. This dense neural network is learnt as part of the training process to conditionally decide the amount of soft attention to be applied to each annotation vector a_i based on the decoder's last hidden state h_{t-1}. This warrants for two inputs to this attention network i.e. the flattened image feature annotations and the latest hidden state of the LSTM. The image feature vectors are projected to a 512-dimensional feature space by a fully-connected layer while another separate fully connected layer does the same for h_{t-1}. The projected hidden state is amalgamated with each of the projected annotation vectors using the add operation which further produces a ReLU activated output of shape (196, 512). The tensor is passed to a Softmax layer that converts it to a probabilistic attention vector of dimension (196, 1). This vector is used to attend the (2048, 196) shaped annotation vectors to finally give the context vector representation of image features.

RNNs require fixed length sequences but we have sentences which are intrinsically of varied lengths. To make them uniform sized, we fixed the maximum size of the caption to be of a suitable length i.e. 39. This does not correspond to the longest sentence size in the dataset but was chosen by doing a percentile analysis discarding outliers to cover 95% of the captions. Longer captions are clipped to comply with the maximum allowed length. To compensate for shorter lengths $< pad >$ tokens are appended to make each caption the same length. We substituted words with frequency of occurrence less than 3 with an $< unk >$ token. This models the probability of unknown words that might appear in validation and test sets captions but are not present in the train set.

We introduced a custom embedding layer of size 512 which learns a fixed length continuous domain representation during the training process. This is the final representation of words that is consumed by the LSTM decoder. The LSTM is used with a hidden state size of 512. To predict the next word, we use the updated hidden state which is upsampled by a fully-connected layer projecting the 512 vector to the vocabulary space. This is connected with Softmax for word prediction. Cross entropy loss (multi class) is used for back-propagation of gradients.

For the baseline model, we use only the last prediction S_{t-1}'s word embedding (512) as input to the next time step. The hidden state h_t incurs a cyclic update in the LSTM. For the attention driven main model, the context vector is combined with the previous prediction's word embedding S_{t-1} to constitute the input. The vectors are combined using concatenation and fed together to the LSTM decoder to generate the next word.

The Adam optimizer is used with a learning rate of $4e^{-4}$. BLEU-4 metric is tracked on the validation set throughout the training process. Adaptive learning rate is used with a decay of 20%, if there is no improvement in BLEU for 8 consecutive epochs. Drop Out of 0.5 has been employed with teacher forcing for 50% of the training epochs chosen randomly. A maximum of 100 Epochs was used, each having mini-batches of 32 while leveraging early stopping based on BLEU score if there are 20 epochs of no improvement.

Cross entropy loss, top 5 accuracy and BLEU scores were tracked. It is observed that the improvement in BLEU score does not always correspond to a reduction in loss so we stopped the training process early using BLEU-4. The resulting improvement in the language scoring metric BLEU-4 is evident as the stabilized img2seq model is tuned further to enhance the Encoder's adaptability. This is done by image encoder retraining. Initially, transfer learning was leveraged on the encoder by keeping its weights frozen and only the decoder was trained. The training phase lasted for 31 epochs with the BLEU-4 score peaking at about 21.56 on the 11th epoch. We fine-tuned the encoder, restarting the training with parameters of the 11th checkpoint using a reduced batch size and reduced learning rate. This is because the trainable model size is now larger, additionally incorporating the computation and backpropagation of the encoder's gradients. For ResNet, we only fine-tune convolutional blocks 2 through 4 while keeping the initial block intact, because the first convolutional block would have usually learned low level features that are fundamental to image processing, such as detecting lines, edges, curves, etc. Consequently we don't

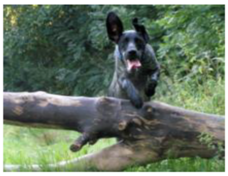
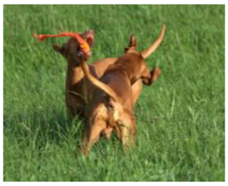
image	Early stopping via BLEU	Early stopping via LASER
	ایک سیاہ کتا گھاس میں چھلانگ لگا رہا ہے [A black dog is jumping on the grass]	ایک سیا سفید کتا اپنی زبان باہر نکالے درخت پر چھلانگ مارتا ہے [A black white dog with his tongue out is jumping over a tree]
	ایک بھورا کتا گھاس میں بھاگ رہا ہے [A brown dog s running the grass]	دو بھورے کتے گھاس میں کھیلتے ہیں [Two brown dogs are playing on the grass]

Fig. 3 Impact of early stopping via BLUE versus LASER

Table 2 Language model trained

Model - Inference Strategy	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Encoder Tuned Attention - Greedy	70.35	53.65	39.3	28.32
Encoder Tuned Attention - Beam 2	72.2	56.04	41.78	30.71
Encoder Tuned Attention - Beam 3	71.75	55.66	41.36	30.14

Table 3 Performance of our model and state-of-the-art

Paper	dataset	Lang.	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Proposed Model	Dogs Flickr8k	Ur	72.5	56.9	42.8	31.6

change foundations. This resulted in improving the BLEU-4 score to a new high of 23.05 after 4 epochs.

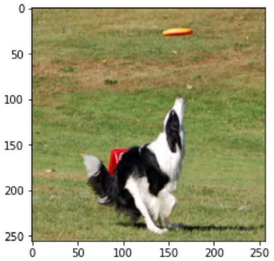
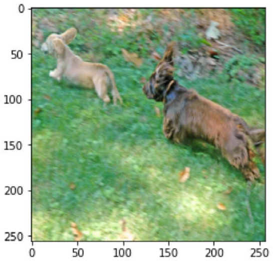
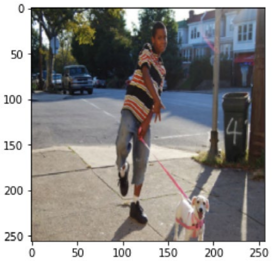
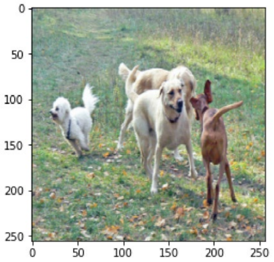
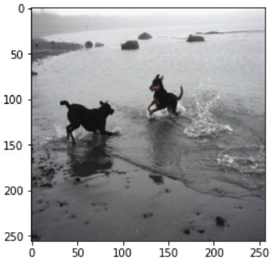
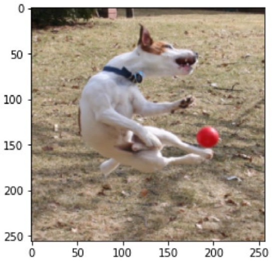
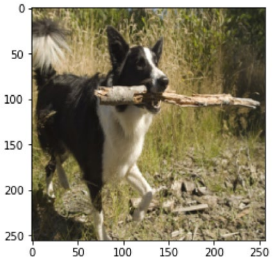
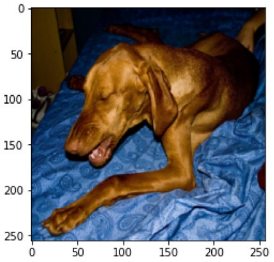
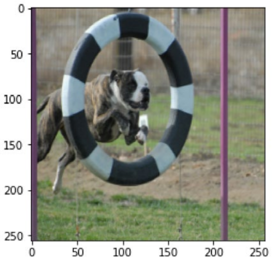
4 Experimental results and discussions

The image to natural language connection jointly tunes the encoder on top of the trained decoder to bridge the contextual gap between visual and linguistic components. This allows the loss feedback to flow to the image encoder improving the visual component compatibility with the language model. Gains in all BLEU 1-4 scores are recorded in Table 2. Table 3 shows the results on Urdu and those of relevant papers and state-of-the-art for English. We decided to test a multilingual BERT model that covers Urdu as well as being implemented in Hugging Face. The model consists of 110M parameters and is sized at 0.7 GB. We configured the main model to integrate with the BERT encoder. The embedding layer was frozen and the LSTM cells were configured to a layer size of 768, matching the dimensionality of the word embedding extracted from BERT. The BERT model uses sentence context in its entirety to generate the embedding and is very effective at encoding semantics. For Urdu, the best strategy was to learn the embeddings from scratch as part of the training process, rather than relying on pre-trained embeddings. This study reports the results using BLEU score as a quantitative metric to evaluate the goodness of fit as well as maximising BLEU score during the training process. BLEU score is based on the sequential conformance of N-Grams whereas natural language involves much more flexible constructs where alternate words or their combinations may constitute the same semantic sense. METEOR and CIDEr metrics are also used by the latest papers but they lack the necessary resources for Urdu. In the pursuit of better metrics for Urdu, we leveraged two additional candidates for sentence semantics (i) BERT-F1 Score (Zhang et al. 2019) which uses the BERT transformer model extracting word features from multiple layers to form semantic representation pools using the words from each of the reference and hypothesis sentences. It then computes Precision and Recall to give F1 for the hypothesis. (ii) LASER is introduced by Facebook Research (Artetxe and Schwenk 2019) to generate multi-modal sentence embeddings for zero-shot cross-lingual transfer. For the languages used for its training, LASER can transform the sentence into a joint space which produces language-independent vectors. To use them as a qualitative measure, there are multiple options such as L1, L2 norms and

Table 4 Early stopping, BLEU verses LASER

Stopper	BLEU-4	BLEU-3	BLEU-2	BLEU-1	BERT-F1	LASER–macro	LASER-macro
BLEU	27.23	38.71	53.29	69.28	85.07	73.99	74.44
LASER	26.57	36.79	50.01	67	85.29	74.63	74.82

Table 5 Samples of good predictions

Good predictions		
 ایک سیاہ سفید کتا ایک فریسبی پکڑنے کے لئے ہوا میں کود رہا ہے A black and white dog is jumping in the air to catch a Fresbi	 ایک بڑا بھورا کتا گھاس میں ایک دوسرے کتے کا پیچھا کرتا ہے A big brown dog is chasing another dog on the grass	 ایک شخص اپنے کتے کے ساتھ سڑک پر کھڑا ہے A person is standing with his dog on the road
 چار کتے ایک دوسرے کے ساتھ کھیل رہے ہیں Four dogs are playing together	 دو کتے ساحل سمندر پر پانی میں بھاگ رہے ہیں Two dogs are running on the beach in water	 ایک سفید کتا سرخ گیند کے ساتھ کھیل رہا ہے A white dog is playing with a red ball
 ایک سیاہ سفید کتا اپنے منہ میں چھری پکڑے ہوئے ہے A black and white dog is holding a knife in his mouth	 ایک بھورا کتا نیلے رنگ کے کپڑے کے ساتھ کھیل رہا ہے A brown dog is playing with a blue cloth	 ایک سیاہ سفید کتا ایک رکاوٹ پر چھلانگ لگا رہا ہے A black and white dog is jumping over a barrier.

cosine similarity. The initial two being subject to certain biases across dimensions, we have used the cosine similarity of each hypothesis against 5 reference captions and computed macro and micro averages as measures to cover the whole evaluation set. We leveraged LASER and BERT F1 scores to govern the model training via early stopping. It was observed that they do not always correlate with BLEU score and the training process stops at a different junction which offers lower BLEU metric but maximizes LASER see Table 4 and Fig. 3. Final results on the evaluation set are listed in Tables 5,6, and 7 for reference and organized into good, average and bad predictions, respectively.

Table 6 Samples of average predictions

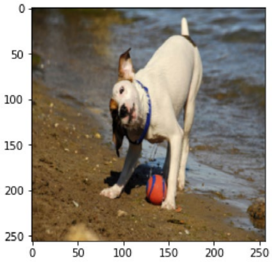
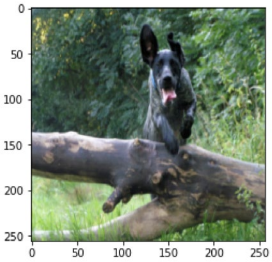
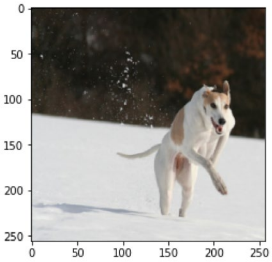
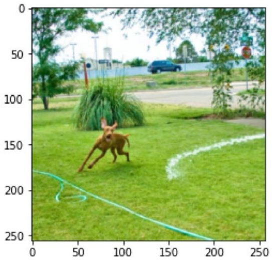
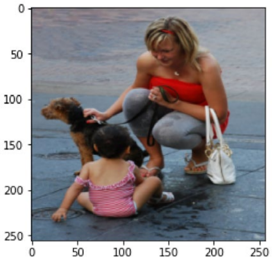
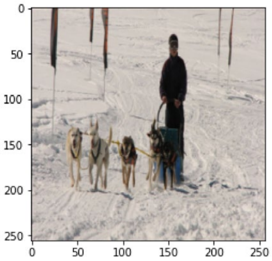
Average Predictions		
		
ایک کتا ساحل سمندر پر چل رہا ہے A dog is walking on the beach	ایک سیاہ کتا گھاس میں چھلانگ لگا رہا ہے A black dog is jumping on the grass	ایک سفید کتا برف میں چل رہا ہے A white dog is walking on the snow
		
ایک بھورا کتا گھاس میں ایک لکڑی کو پکڑتا ہے A brown dog is catching wood on the grass	ایک عورت اپنے کتے کے ساتھ کھیل رہا ہے A women is playing with his dog	ایک آدمی برف میں دو کتوں کو چلا رہا ہے ا A man is walking his dogs n the snow

Table 7 Samples of bad predictions

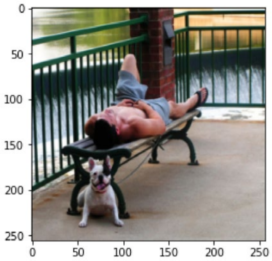
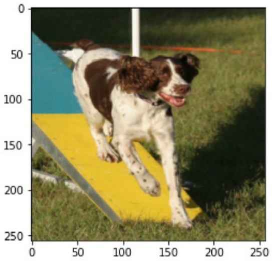
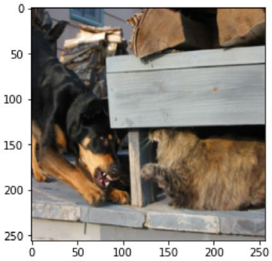
Bad Predictions		
		
ایک سیاہ سفید کتا ایک چھوٹے کتے کے ساتھ کھیل رہا ہے A black and white dog is playing with a small dog	ایک بھورا سفید کتا ایک پیلے رنگ کے کتے کے ساتھ کھیلتا ہے A brown and black dog is playing with a yellow dog	ایک سیاہ کتا ایک سیڑھیوں پر چھلانگ مارتا ہے A black dog is jumping on a stairs

Table 8 Annotator disagreements along with the examples

Disagreement	Example
Color Translations	کالا ، سیاہ ، سرمئی (Black), لال ، سرخ ، نارنجی (Red)
Generic Translations	گرے ھوئے، گرے پڑے، نیچے پڑے ، گرے ہوئے (Fallen tree)
Named Entities	پائپ ، نالی (Pipe)
Action descriptions	برف میں دوڑ، برف پر دوڑ
Confusing Counts	A black and a brown dog hold a stick versus A black and brown dog hold a stick

5 Conclusions and future work

This is the first study on generative image captioning in Urdu. We present a new dataset for Urdu image captioning, annotation treatment and generalization guidelines to make visio-lingual deep learning models effective and applicable to modest sized dataset. We highlight the hindrances of standard evaluation metrics in Urdu and show the use of semantics driven techniques such as Bert-F1 and LASER may be appropriate for evaluating this task in Urdu. One can use transformer for decoder part to enhance the language model ability in the captioning which is left as future work at this movement.

Appendix: A image captions dataset creation for Urdu

To prepare the image mapped Urdu dataset we make use of Flickr8K dataset for cross-reference which is a standard dataset and widely used by the research community to perform image caption generation tasks for English. The Flickr8K dataset comprises 8000 images where each image is presented with 5 English captions on average. Our dataset was created in three phases. There are three high-level approaches and all were exploited in turn one after the non-viability of the other. These approaches are explained in the following subsections.

A.1 Automatic translation

To translate English captions to Urdu, we subscribed to the Google cloud hosted neural machine translation (NMT) model (v2). Once the translation was completed, a preliminary baseline model was trained as a trial. However, it was noted that even though the evaluation scores were acceptable (i.e., BLEU=13), but several generated captions were absurd and un-related to the image. We also found that the translation API lagged in producing quality Urdu translations. These findings of erroneous instances enforced the consideration of human translation as the reliable option to prepare captions.

A.2 Human translation

The human translators consisted of a few colleagues, who are proficient in English while having Urdu as their native language. As translation was progressing, a parallel task of analysing the Urdu annotations was initiated and plethora of issues were faced such as: (i) Urdu Words are not essentially space separated and since they do not always form invalid or different words unlike English. This makes such typing anomalies hard to spot while causing high variability in data (اور آخر میں ، اورآخرمیں) (ii) Typos pertaining to missing or extra spaces are hard to correct, as it causes the confusion to consider such occurrence as a named entity or result of a missing space (پانی میں ، پانیمیں) (iii) Numerous instances where missing spaces are considered syntactically correct and semantically identical Urdu words (كي ساتھ (کیساتھ) کی ساتھ (کیساتھ) ، ہوئے ، (ھوئے) (iv) Many words don’t have corresponding Urdu translation and were typed by same or multiple typists differently while each being correct e.g.: Frisbee (فرسبی ، فریسبی ، فرسبي ، فريسبي) (v) Typists variably used Phonetic Characters (Ayraab). (vi) Inter and Intra annotator disagreements were observed while translating the same English phases at different instances (see Table 8). All of these observations established the source of high textual variability of captions potentially causing the model learning process to suffer. Such findings paved the way for Phase 3 to apply corrections and standardization.

A.3 Compliment with human annotation

We combined phase 2 with manual re-annotation, validating each English captions for correctness and relevancy to each corresponding image. Upon verification, we translate English to Urdu, otherwise the human annotators shall self-generate 5 grammatical descriptions in Urdu and type them. The annotations were periodically analyzed using basic NLP techniques while keeping a check on vocabulary size and instances per word that shall be available to learn the Urdu language model later. Keeping in view these issues, below were the high-level aspects that were identified to be fixed: (i) Preprocessing techniques applicable to English text cannot be used directly for Urdu e.g. *string. punctuation* in Python that works effectively for English did not detect Urdu punctuations being limited to ASCII only. Urdu does not have upper/lower case, English punctuations and their representations are different altogether from those of Urdu. For instance, Urdu full-stop comma etc. is different from that we have in English (. vs ۔ , vs ،) and unfortunately a mix was used by the typists. (ii) Urdu had instances of Ayraab: Zabr, Zayr, Pesh, Shad, Mad. (iii) After human annotation, we have digits versus Urdu Counting versus Word based counting (8, ۸, آٹھ) (iv) Mechanism to detect and correct the words with space missing compounds versus true compounds. (v) The need to standardize multiple correct versions of the same word. (vi) Pre-trained effective NER Models are not readily available that could be leveraged for text standardization. (vii) Urdu WordNet, Normalizer, Stemmer were non-existent in standard libraries NLTK. In addition to vocabulary analysis, baseline model training was also carried out periodically to ascertain that the Urdu dataset creation effort is fruitful. It was learnt that Language Model learning becomes hard and comparatively needs more data for Urdu as compared to English because of the inherent variety and depth of natural Urdu text. Apart from typing and translation issues , there are such genuine constructs in Urdu that introduce high variability in data and hinder Model learning thus requiring much more data to accommodate such variations. Some of the aspects are noted below:

- **Gender sense**
  Urdu characterising words exhibit context driven Feminine Masculine variations:

English	Urdu
A blue ball	ایک نیلی گیند
A blue bat	ایک نیلا بلا

- **Plural sense**
  Counts and context govern the variability in reporting same property e.g. Color:

English	Urdu
A brown dog is sitting	ایک بھورا کتا بیٹھا ہوا ہے
Two brown dogs are sitting	دو بھورے کتے بیٹھے ہیں
A brown dog has a ball in his mouth	ایک بھورے کتے کے منہ میں ایک گیند ہے

- **Dynamic versus static usage** Same property when translated to Urdu has two sources of variation (i) Annotator (ii) non-uniform rules i.e. certain Urdu translations have a static usage while others exhibit context adherence:

Static	Green: سبز , Black: سیاہ
Context Driven	Green: ہرا , Black: کالا , Blue:نیلا

- **Model puzzling count variations**
  Urdu words observe complex numerical correspondence:

English	Urdu
Two dogs are sitting	دو کتے بیٹھے ہیں
A dog has a ball in his mouth	ایک کتے کے منہ میں ایک گیند ہے

A.4 Dataset creation and standardization principles

The overall exercise of preparing Urdu dataset became exceedingly laborious and demanding. Considering the time constraints, to limit the stretch, a potential way out was to prepare a quality Urdu dataset with correctness as the focus but at the expense of data size. A set of principals were formulated to finalize the dataset:

- **Two options** for reducing the annotation dataset size: (i) Reduce the Caption volume, keep complete 8K Images set in scope (1 caption/image: 8K captions). (ii) Reduce the scope by sampling the images systematically while leveraging multiple captions per image (9K Captions: 5 captions/image)
- **Prioritize Quality:** Garbage In , Garbage Out
  Produce high quality dataset, having multiple review cycles for corrections covering all the troubled scenarios enlisted earlier.
- **Standardization:** Have a single annotator to standardize the corpus after correction rounds. Preference is to avoid high annotation variability through standardized usage of vocabulary
- **Learnable Volume:** Volume should be enough to generalize well while being able to effectively train decent generative deep learning models for the chosen scope.

To provide the model with enough examples per image to learn the language model without loss of generality, **Option 2** was chosen:

- Analyzed the frequent subject in Flickr8K: **Dogs**
- Flickr8K has **1800 images** related to Dogs
- Annotators to use English captions as support to translate or preferably annotate where required yielding 5 captions per image : **9000 Urdu Captions**
- Set of rules were applied to the post annotation Urdu text :

  1. **Preprocessing of Urdu punctuations:**
     We leveraged the Unicode character set as it demonstrated the property that fancy characters had a Unicode of pattern ‘P*’. Urdu punctuation characters were effectively covered under this category.
  2. **Compound word normalization and split corrections:**
     The vocabulary was sorted by token size descending and ascending, selecting top 500 for each. Each of these tokens were manually analysed to identify the missing space or extra space resulting sub-word issues and fixed by replacing each of such instance in the corpus with the appropriate substitute.
  3. **Typing mistakes identification and correction:**
     We used Urdu-to-English word translation look-up on the corpus vocabulary to flag the typing mistakes. This effectively supported with NER and typing issues. Typos were systematically searched and replaced in the annotations corpus.
  4. **Standardization of colour expressions:**
     As enlisted earlier, there are multiple colour expressions in Urdu that correspond to the same colour in English, with the noticeable property that a subset demonstrates static usage per color while others posit a plural or gender sense. All such instances were changed in the favour of static equivalents to reduce variety e.g. Green (ہرا، ہری، ہرے ← سبز)
  5. **Named entity normalization**
     Named entities were identified using earlier stated systematic analysis as well as a manual round of proof reading. The improper nouns were not touched to avoid loss of generality. However, the most common scenarios where normalization was applied pertained to Dog breeds e.g. (Labrador, German Shepherd, Mastiff) → تعداد + حجم + رنگ + کتا/ ے
  6. **Relevance assurance, count and multi-word correction rounds**
     Corrective iterations were done for each of caption relevance, number standardization to Urdu words and normalization of multi-token representation of the same word.

References

Agrawal H, Desai K, Wang Y, Chen X, Jain R, Johnson M, Batra D, Parikh D, Lee S, Anderson P (2019) Nocaps: novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 8948–8957

Amjad M, Sidorov G, Zhila A (2020) Data augmentation using machine translation for fake news detection in the urdu language. In: Proceedings of the 12th language resources and evaluation conference, LREC 2020, Marseille, France, May 11-16, 2020. European Language Resources Association, pp. 2537–2542

Aneja J, Deshpande A, Schwing AG (2018) Convolutional image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5561–5570

Artetxe M, Schwenk H (2019) Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transac Assoc Comput Linguist 7:597–610

Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate." In: 3rd International Conference on Learning Representations

Chen X, Lawrence Zitnick C (2015) Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2422–2431

Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D (2010) Every picture tells a story: generating sentences from images." In: European conference on computer vision. Springer, pp. 15–29

Fisch A, Lee K, Chang M, Clark JH, Barzilay R (2020) Capwap: image captioning with a purpose. In: Webber B, Cohn T, He Y, Liu Y (Eds.) Proceedings of the 2020 Conference on empirical methods in natural language processing. Association for computational linguistics, pp 8755–8768

Gong Y, Wang L, Hodosh M, Hockenmaier J, Lazebnik S (2014) Improving image-sentence embeddings using large weakly annotated photo collections. In: European conference on computer vision. Springer, pp 529–545

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778

Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res 47:853–899

Huang L, Wang W, Chen J, Wei X-Y (2019) Attention on attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4634–4643

Kanwal S, Malik K, Shahzad K, Aslam F, Nawaz Z (2020) Urdu named entity recognition corpus generation and deep learning applications. ACM Trans Asian Low Resour Lang Inf Process 19(1):8:1-8:13. https://doi.org/10.1145/3329710

Karpathy A, Joulin A, Fei-Fei L (2014) Deep fragment embeddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679

Kim J, Rohrbach A, Darrell T, Canny J, Akata Z (2018) Textual explanations for self-driving vehicles. In: Proceedings of the European conference on computer vision, pp 563–578

Kiros R, Salakhutdinov R, Zemel RS (2014) Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539

Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2013) Babytalk: understanding and generating simple image descriptions. IEEE Transac Pattern Anal Mach Intell 35(12):2891–2903

Kuznetsova P, Ordonez V, Berg TL, Choi Y (2014) Treetalk: composition and compression of trees for image descriptions. Transac Assoc Comput Linguist 2:351–362

Li S, Kulkarni G, Berg T, Berg A, Choi Y (2011) Composing simple image descriptions using web-scale n-grams. In: Proceedings of the fifteenth conference on computational natural language learning, pp. 220–228

Li C, Bao Z, Li L, Zhao Z (2020) Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNS for multimodal emotion recognition. Inf Process Manag 57(3):102185

Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp. 740–755

Lindh A, Ross RJ, Kelleher JD (2020) Language-driven region pointer advancement for controllable image captioning. In: Scott D, Bel N, Zong C (Eds.) Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics, pp 1922–1935

Liu J, Wang K, Xu C, Zhao Z, Xu R, Shen Y, Yang M (2020) Interactive dual generative adversarial networks for image captioning. In: The thirty-fourth AAAI conference on artificial intelligence, AAAI. AAAI Press, pp 11 588–11 595

Mahmood Z, Safder I, Nawab RMA, Bukhari F, Nawaz R, Alfakeeh AS, Aljohani NR, Hassan S-U (2020) Deep sentiments in roman Urdu text using recurrent convolutional neural network model. Inf Process Manag 57(4):102233

Makav B, Kılıç V (2019) A new image captioning approach for visually impaired people. In: 2019 11th International conference on electrical and electronics engineering (ELECO). IEEE, pp 945–949

Malik MI, Sindhu MA, Khattak AS, Abbasi RA, Saleem K (2021) Automating test oracles from restricted natural language agile requirements. Expert Syst J Knowl Eng 38(1):e12608

Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2014) Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632

Mishra SK, Dhir R, Saha S, Bhattacharyya P (2021) A Hindi image caption generation framework using deep learning. ACM Trans Asian Low Resour Lang Inf Process 20(2):32:1-32:19

Pan Y, Yao T, Li Y, Mei T (2020) X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10971–10980

Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp. 311–318

Shaik R, Venkatramaphanikumar S (2021) Sentiment analysis with word-based Urdu speech recognition. J Ambient Intell Humanized Comput 25:1–21

Socher R, Karpathy A, Le QV, Manning CD, Ng AY (2014) Grounded compositional semantics for finding and describing images with sentences. Transac Assoc Comput Linguist 2:207–218

Tanti M, Gatt A, Camilleri K (2017) What is the role of recurrent neural networks (rnns) in an image caption generator? In: Proceedings of the 10th international conference on natural language generation, pp. 51–60

Tanti M, Gatt A, Camilleri KP (2018) Where to put the image in an image caption generator. Nat Lang Eng 24(3):467

Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164

Wang J, Wang W, Wang L, Wang Z, Feng DD, Tan T (2020) Learning visual relationship and context-aware attention for image captioning. Pattern Recognit 98:107075

Wang C, Shen Y, Ji L (2022) Geometry attention transformer with position-aware ISTMS for image captioning. Expert systems with applications. Elsevier, p 117174

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057

Yao BZ, Yang X, Lin L, Lee MW, Zhu S-C (2010) I2t: image parsing to text description. Proc IEEE 98(8):1485–1508

Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transac Assoc Comput Linguist 2:67–78

Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2019) "Bertscore: Evaluating text generation with bert." In: International conference on learning representations

Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J (2021) Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5579–5588

Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J (2020a) Unified vision-language pre-training for image captioning and VGA. Proc AAAI Conf Artif Intell 34(07):13041–13049

Zhou Y, Wang M, Liu D, Hu Z, Zhang H (2020b) More grounded image captioning by distilling image-text matching model. In: 2020 IEEE/CVF conference on computer vision and pattern recognition. IEEE, pp. 4776–4785

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.