Hausa Visual Genome: A Dataset for Multi-Modal English to Hausa Machine Translation

Idris Abdulmumin[1,6], **Satya Ranjan Dash**[2], **Musa Abdullahi Dawud**[2], **Shantipriya Parida**[3], **Shamsuddeen Hassan Muhammad**[4,5], **Ibrahim Sa'id Ahmad**[6], **Subhadarshi Panda**[7], **Ondřej Bojar**[8], **Bashir Shehu Galadanci**[6], **Bello Shehu Bello**[6]

[1]Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria
[2]School of Computer Applications, KIIT University, Bhubaneswar, India
[3]Silo AI, Helsinki, Finland
[4] LIAAD - INESC TEC, [5]Faculty of Sciences-University of Porto, Portugal
[6]Faculty of Computer Science and Information Technology, Bayero University, Kano, Nigeria
[7]Graduate Center, City University of New York, USA
[8]Charles University, Faculty of Mathematics and Physics, ÚFAL, Prague, Czech Republic
iabdulmumin@abu.edu.ng, sdashfca@kiit.ac.in, dawudmusa46@gmail.com, shantipriya.parida@silo.ai, {shmuhammad.csc, isahmad.it, bsgaladanci.se, bsbello.cs}@buk.edu.ng, spanda@gradcenter.cuny.edu, bojar@ufal.mff.cuni.cz

Abstract

Multi-modal Machine Translation (MMT) enables the use of visual information to enhance the quality of translations. The visual information can serve as a valuable piece of context information to decrease the ambiguity of input sentences. Despite the increasing popularity of such a technique, good and sizeable datasets are scarce, limiting the full extent of their potential. Hausa, a Chadic language, is a member of the Afro-Asiatic language family. It is estimated that about 100 to 150 million people speak the language, with more than 80 million indigenous speakers. This is more than any of the other Chadic languages. Despite a large number of speakers, the Hausa language is considered low-resource in natural language processing (NLP). This is due to the absence of sufficient resources to implement most NLP tasks. While some datasets exist, they are either scarce, machine-generated, or in the religious domain. Therefore, there is a need to create training and evaluation data for implementing machine learning tasks and bridging the research gap in the language. This work presents the Hausa Visual Genome (HaVG), a dataset that contains the description of an image or a section within the image in Hausa and its equivalent in English. To prepare the dataset, we started by translating the English description of the images in the Hindi Visual Genome (HVG) into Hausa automatically. Afterward, the synthetic Hausa data was carefully post-edited considering the respective images. The dataset comprises 32,923 images and their descriptions that are divided into training, development, test, and challenge test set. The Hausa Visual Genome is the first dataset of its kind and can be used for Hausa-English machine translation, multi-modal research, and image description, among various other natural language processing and generation tasks.

Keywords: Natural Language Processing, Machine Translation, Hausa Language Resources, Low Resource Languages, Multi-modal dataset, Visual Genome

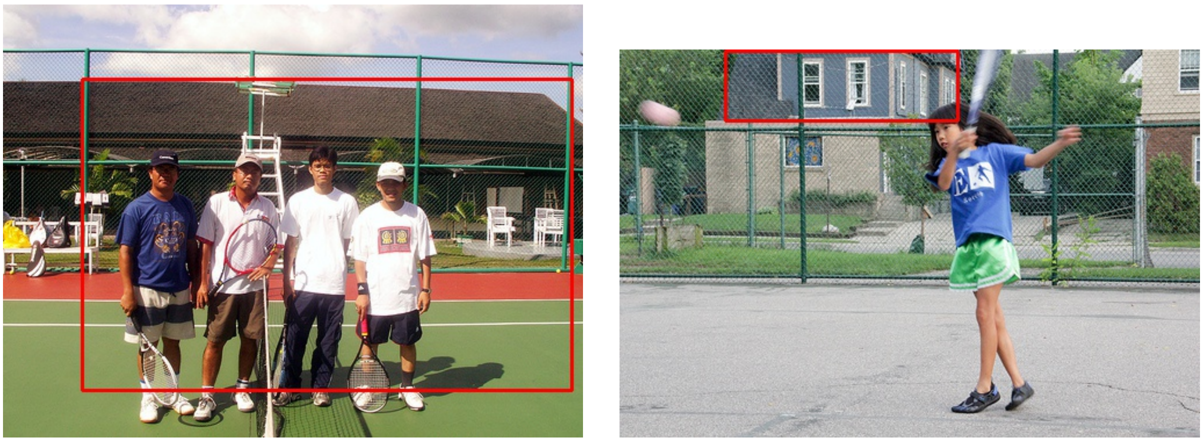
1. Introduction

Machine translation is the use of a computer to automatically generate the equivalent of a given source text in a language that is different from the original language. While Neural Machine Translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017; Gehring et al., 2017) has revolutionized automatic translation, the absence of sufficient training data in many languages has limited the benefits of such systems to a few languages rich in resources, although at least some treatment is possible even for low-resource languages (Sennrich and Zhang, 2019).

Multi-modal Machine Translation (MMT) enables the use of visual information to improve the translation quality, supplementing the missing context and providing cues to the machine translation system for better disambiguation. Despite the increasing popularity of multi-modal techniques, sufficiently large and clean datasets are scarce to fully benefit from the potential. For languages with such data, various approaches have been proposed, demonstrating their usability in improving translation quality, e.g., see (Krishna et al., 2017; Lin et al., 2020; Long et al., 2021; Liu et al., 2021).

The images in Figure 1 present some examples where the absence of context allows to consider two different translations, where each is correct in a different setting. In the first one, the English word "court" is translated as *kotu*, which is the Hausa word for a legal court. But the image illustrates that the men were standing on a [tennis] field. The absence of the word "tennis" misled the standard machine translation system and even many human translators into thinking that the former translation is required. The second example mentions a "story" of a two-story house. The MT system translated the description as *labarin*, meaning story (narrative), instead of the correct *bene* (house story/storey). Without the picture, even human translators may make the same error given the very short and not quite correct English source.

Hausa is a Chadic language and a member of the Afro-Asiatic language family. Hausa is the most-spoken language in this family, with an estimate of about 100 to



English: four men on court
Hausa: *maza hudu a* ***filin wasa***
Gloss: *four men on a* ***playing field***
MT: *maza hudu a* ***kotu***
Gloss: *four men on a* ***(legal) court***

English: second story of house
Hausa: ***bene na biyu*** *na gida*
Gloss: ***second storey*** *of a house*
MT: ***labarin*** *gida* ***na biyu***
Gloss: ***story of second*** *house*

Figure 1: Sample data from HaVG. The first translations (Hausa) are generated by Human Translators. The second translations (MT) are generated by a standard neural machine translation system, Google Translate. The wrong translations are in red font and bolded.

150 million first-language and second-language speakers.[1] The majority of these speakers are concentrated in the Northern part of Nigeria in cities such as Kano, Daura, Sokoto, Zaria, etc., and the Southern Niger Republic. The language is written in Arabic or Latin characters. The Arabic script is known as the *Ajami* and was mostly used in the pre-colonial era, dating back to the 17th century (Jaggar, 2006). The language is nowadays written in the Latin script known as *boko*.

Despite a large number of speakers and many written books, e.g., Jaggar (2006), Umar (2013), Turner (2021), Hausa is considered a low resource language in NLP. This is due to the absence of enough publicly available resources to implement most of the tasks in NLP. While some datasets exist, they are either scarce, machine-generated, or in the religious domain. This limits diversity, restricting the usage of trained models to very few domains. For tasks such as multi-modal translation, and image-to-text translation (image captioning), among others, there exist no training or evaluation data. For translation in the news domain, only an evaluation dataset exist(Goyal et al., 2021). Therefore, there is a need to create training and evaluation datasets for building machine learning models to help reduce the research gap between the low-resourced Hausa language and other languages.

This work, therefore, presents the Hausa Visual Genome (HaVG), a dataset that contains the description of an image or a section within the image in English and its equivalent in Hausa. The dataset was prepared by automatically translating the English description of the images in the Hindi Visual Genome (HVG) (Parida et al., 2019). The data is made of 32,923 images and their descriptions that are divided into training, development, test, and challenge test set. The machine-generated Hausa descriptions were then carefully post-edited taking into account the corresponding images. The HaVG is the first dataset of its kind in Hausa and can be used for Hausa-English machine translation, multi-modal research, and image description, among various other natural language processing and generation tasks.

The objective of the paper is two-fold:

1. To describe the process of building the multimodal dataset for the Hausa language suitable for English-to-Hausa machine translation, image captioning, and multimodal research.

2. To demonstrate some sample use cases of the newly created multimodal dataset: HaVG.

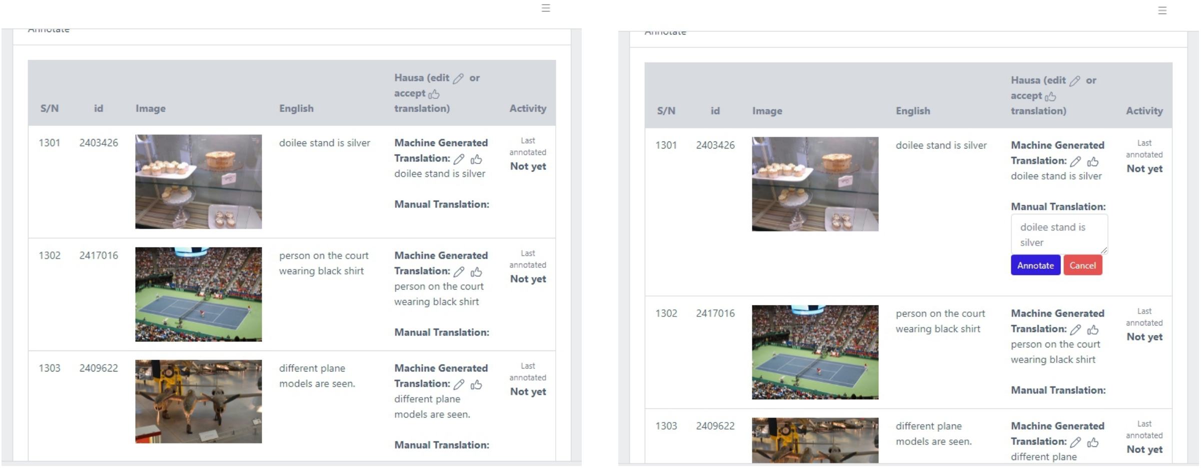
The rest of the paper is arranged as follows: Section 2 presents the available datasets for NLP in the Hausa language. Section 3 presents the processes of data collection and labeling. In Section 4, we present some experiments and results on the application of the HaVG data. Finally, we conclude the work and provide directions for the future in Section 5.

2. Related Work

While the Hausa language does not have any dataset for multimodal tasks, a few others have been created for other NLP tasks. Abubakar et al. (2021) produced sentiment annotations of tweets and used them in their work. Inuwa-Dutse (2021) provided a pseudo-parallel corpus for machine translation. The Tanzil dataset[2]

[1] https://www.herald.ng/full-list-hausa/

[2] https://opus.nlpl.eu/Tanzil.php



S/N	id	Image	English	Hausa (edit ✎ or accept 👍 translation)	Activity
1301	2403426		doilee stand is silver	Machine Generated Translation: ✎ 👍 doilee stand is silver Manual Translation:	Last annotated Not yet
1302	2417016		person on the court wearing black shirt	Machine Generated Translation: ✎ 👍 person on the court wearing black shirt Manual Translation:	Last annotated Not yet
1303	2409622		different plane models are seen.	Machine Generated Translation: ✎ 👍 different plane models are seen. Manual Translation:	Last annotated Not yet

a: Preview

S/N	id	Image	English	Hausa (edit ✎ or accept 👍 translation)	Activity
1301	2403426		doilee stand is silver	Machine Generated Translation: ✎ 👍 doilee stand is silver Manual Translation: doilee stand is silver Annotate Cancel	Last annotated Not yet
1302	2417016		person on the court wearing black shirt	Machine Generated Translation: ✎ 👍 person on the court wearing black shirt Manual Translation:	Last annotated Not yet
1303	2409622		different plane models are seen.	Machine Generated Translation: ✎ 👍 different plane	Last annotated Not yet

b: Annotate

Figure 2: Annotation web page showing images and their description. (a) To edit the machine translation, the pencil icon is clicked. To accept, the thumbs-up icon is clicked. (b) After clicking on the edit icon, a text area with the machine translation is displayed for post-editing.

(Tiedemann, 2012), a translation of the Quran in many languages including Hausa, and the JW300[3] (Agić and Vulić, 2019) are available for machine translation tasks. All of these data, though, are either not natural or strictly in the religious domain, limiting the accuracy or general applicability of the translation models trained on them. Apart from the FLoRes evaluation dataset (Goyal et al., 2021) for machine translation tasks, which is not reflective of the domain of available training data, there exists no standard benchmark evaluation (test) sets that truly indicate the performance of natural language processing models to the best of our knowledge.

Resources for other NLP tasks in the language are also scarce. Abdulmumin and Galadanci (2019) provided two sets of word embeddings in Hausa for NLP. Schlippe et al. (2012) and Schultz (2002) built a collection of transcribed speech resources for automatic speech recognition (ASR) and similar tasks in the language. Tukur et al. (2019) trained a part-of-speech tagger for Hausa.

Initiatives such as Masakhane[4] and HausaNLP[5] have started creating these data for Hausa and other African languages, most of which are considered low-resource and these will help in future NLP research and application in such languages.

3. Training and Evaluation Data

3.1. Data Collection

The HaVG training and evaluation (development test and challenge test) data were produced by automatically translating the Hindi Visual Genome (HVG) and revising it as described below.

The HVG training, evaluation, and test dataset consist of randomly selected images and their descriptions from the Visual Genome (VG) corpus (Krishna et al., 2017). The HVG challenge test set was specifically sampled so that each sentence contains an English word that is lexically ambiguous when translated into Hindi. While the VG data contains multiple captions in English, with each caption representing a particular region in an image, the HVG data contains only a single random caption of a section in each image.

3.2. Annotation

To prepare the HaVG data, therefore, we implemented the following steps:

1. We use Google Translate[6] to translate all the available 32,923 HVG English captions into Hausa.

2. We developed a web-based annotation tool[7] and hosted it locally to help with the post-editing of these translations. The web interface enables the annotator to edit the generated translations by showing them the image and the original caption side-by-side. See the illustration in Figure 2.

3. We gave the machine translations of the captions to Hausa volunteers for post-editing. The translations of many of the unambiguous sentences were mostly found to be correct.

4. For a secondary check, we sampled 3,500 of the post-edited captions (representing about 10% of the whole dataset) for manual verification. It was found that a small number of the sentences were found unedited even though there were obvious

[3] https://opus.nlpl.eu/JW300.php
[4] https://www.masakhane.io/
[5] https://www.hausanlp.org/
[6] https://translate.google.com/
[7] https://github.com/abumafrim/visual-genome-dataset-creation-tool

Data	#Sentences	Language	Word Stat.			#Tokens	#Vocab
			max	min	avg		
Training	28,930	HA	36	1	5.01	144,864	6,636
		EN	29	1	5.09	147,219	7,046
Development Test	998	HA	14	1	4.99	4,978	1,167
		EN	13	1	5.08	5,068	1,092
Evaluation Test	1,595	HA	17	1	4.99	7,952	1,478
		EN	13	1	5.07	8,079	1,502
Challenge Test	1,400	HA	27	2	6.80	9,514	1,583
		EN	18	2	6.01	8,411	1,461
Total	32,923	–	–	–	–	–	–

Table 1: Statistics of the Hausa Visual Genome dataset

Method	D-Test BLEU	E-Test BLEU	C-Test BLEU
Text-to-text translation	31.3	46.7	17.7
Multimodal translation	15.7	22.6	8.2

Table 2: Results of text-only and multimodal translation on the HaVG dataset.

translation errors. The errors in these sentences were corrected by the verifiers.

Some statistics in the annotated HaVG dataset are provided in Table 1. We used the NLTK punkt tokenizer (Bird et al., 2009) to estimate the statistics. The Hausa sentences of the HaVG were found to have 36 and 1 word in the longest and shortest sentences, respectively. The average sentence length ranges from 4.99 to 6.80 words per sentence, with the challenge test set statistically having longer sentences. The training set has a low type-token ratio (TTR) – a measure of vocabulary variation or lexical richness of a text – of 0.05. This is reflective of the restricted domain of the data as most of the sentences are in the sports domain, mainly tennis.

4. Sample Applications of HaVG

4.1. Text-Only Translation

We used the Transformer model (Vaswani et al., 2018) as implemented in OpenNMT-py (Klein et al., 2017).[8] Subword units were constructed using the word pieces algorithm (Johnson et al., 2017). Tokenization is handled automatically as part of the pre-processing pipeline of word pieces.

We generated a vocabulary of 32k subword types jointly for both the source and target languages, sharing it between the encoder and decoder. We used the Transformer base model (Vaswani et al., 2018). We trained the model on a single GPU and followed the standard "Noam" learning rate decay,[9] see Vaswani et al. (2017) or Popel and Bojar (2018) for more details. Our starting learning rate was 0.2 and we used 8000 warm-up steps. The text-to-text translation results for the development (D-Test), dev (D-Test), test (E-Test), and challenge test (C-Test) are shown in Table 2.

In Table 3, we present some examples where the text-only translation system was able to generate correct translations, although not the exact wording of the reference translations. The system translated "**stand**" as "*tsayuwa*" whereas the most appropriate translation should have been "*mazauni*" (with *mazauni* meaning a place where something is kept while *tsayuwa* means something/someone is in a standing position). The system also translated "**block stone**" as "*dutse* (stone)", omitting "block".

4.2. Multimodal Translation

Multimodal translation involves utilizing the image modality in addition to the English text for translation to Hausa. We take the multimodal neural machine translation approach using object tags derived from the image (Parida et al., 2021a). We first extract the list of (English) object tags for a given image using the pre-trained Faster R-CNN (Ren et al., 2015) with ResNet101-C4 (He et al., 2016) backbone. We pick the top 10 object tags based on their confidence scores. In cases where less than 10 object tags are detected, we consider all tags.

Next, the object tags are concatenated to the English sentence which needs to be translated to Hausa. The concatenation is done using the special token '##' as the separator. The separator is followed by comma-separated object tags. Adding object labels enables the otherwise text-based model to utilize visual concepts which may not be readily available in the original sentence. The English sentences along with the object

[8] http://opennmt.net/OpenNMT-py/quickstart.html

[9] https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html

Image	Text	
	Source	Television in the tv stand.
	Reference	Talabijin a cikin mazaunin talabijin
	Object Tags:	person, potted plant, book, tv, vase
	Text-only	Talabijin a cikin tsayuwa.
	Gloss.	Television in the *standing*.
	Multi-modal	Talabijin a cikin teburin tv
	Gloss.	Television in the tv table.
	Source	woman sitting on a stone block
	Reference	mace zaune a kan bulon dutse
	Object Tags:	person, suitcase, bench, remote
	Text-only	mace zaune a kan dutse
	Gloss.	woman sitting on a stone
	Multi-modal	mace zaune akan bangon dutse
	Gloss.	woman sitting on a stone wall

Table 3: Text-only vs. Multi-modal Machine Translation.

Output	**140 samples (10%)**		
	Correct	**Partially correct**	**Incorrect**
Text-only translation	40	49	51
Multimodal translation	13	39	88
Multimodal translation resolves ambiguity	14		
Multimodal translation reasonable	74		

Table 4: Comparison of outputs from various systems through manual evaluation of the challenge test set.

tags are fed to the encoder of a text-to-text Transformer model. The decoder generates the Hausa translations auto-regressively. We generated a vocabulary of 50k subwords for both source and target languages. Then we trained the Transformer base model using the "Noam" learning rate decay. We used an initial learning rate of 2, dropout of 0.1, and 8000 warm-up steps. The results of the multimodal translation are shown in Table 2.

The automatic evaluation indicates that the text-only translation performs better on both the evaluation and challenge test sets when compared to the multimodal translation. However, upon manual inspection of the outputs, we observed instances where the multimodal system was able to resolve ambiguity and generate a more appropriate translation of the given source sentence, see Table 3 for some examples. The performance is strikingly lower on the challenge test set compared to the evaluation set in both setups. We performed a manual evaluation on a sample of this data to investigate the reason for this low performance.

About 10% of the translations of the challenge test set by both the text-only and multimodal systems were sampled and manually evaluated to assess the quality of the generated sentences. We categorized these sentences as either **correct**, **partially correct**, or **incorrect**. We also checked instances where the multimodal system is not only correct (or partially correct) but was also able to resolve ambiguity. Lastly, we checked whether the sentences generated by the multimodal system are reasonable or not, i.e. whether they generally capture the original meaning. The results of this evaluation are provided in Table 4.

While the multimodal system was found to be half as accurate compared to the text-only model, it was able to resolve ambiguity in about 10% of the sampled data. Finally, we observe that the annotation for "reasonable" translations (i.e. whether the meaning is "generally captured" is apparently much more permissive that the annotation for correctness: a substantial amount of the generated text (74 items, i.e. 53%) was found to be reasonable even though only about 37% of the sentences are either correct or partially correct translations of the source sentences. This detailed analysis nevertheless confirmed that the multi-modal system produces overall worse translations, perhaps confused by the automatic object captions.

4.3. Image Caption Generation

To generate the Hausa captions, we followed Parida et al. (2021b) who proposed a region-specific image captioning method through the fusion of the encoded features of the region and the complete image. The model consists of three modules – an encoder, fusion, and decoder – as shown in Figure 3.

Image encoder In the proposed approach, the features of the entire image, as well as features of the sub-region, are considered to train the model. The

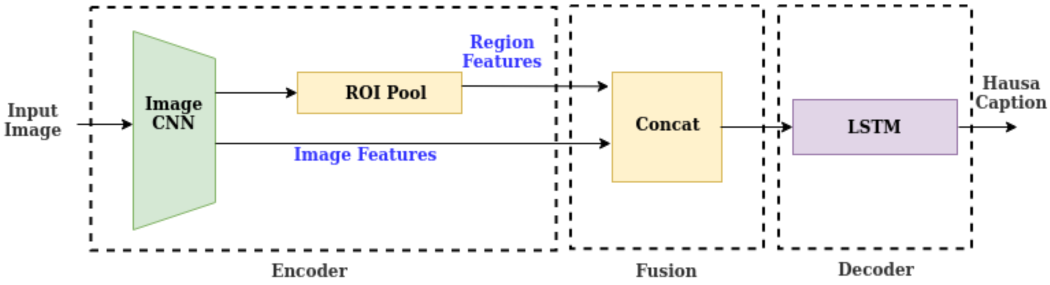


Figure 3: Architecture of the region-specific image caption generator.

Method	D-Test BLEU	E-Test BLEU	C-Test BLEU
Image captioning	2.6	3.1	0.7

Table 5: Results of image caption generation on the HaVG dataset.

features from the corresponding regions are extracted through Region of Interest (RoI) pooling (Girshick, 2015). Specifically, the feature vector is the output of the fourth block of ResNet-50 in our experiments. It is a 2048-dimensional vector for both the image and the sub-region. We keep the image encoder module non-trainable. In other words, it is used as a feature extractor.

Fusion module While the region-level features capture details of the region (objects) to be described, the image-level features provide an overall context. To generate a meaningful caption, both need to be fused appropriately. We obtain the final feature vector by simple concatenation of features from the region and features from the entire image. The concatenation resulted in a 4096-dimensional vector.

LSTM decoder The concatenated feature vector is passed through a linear layer to project it into a 128-dimensional vector which is then fed as input to an LSTM decoder as the first time step. The decoder generates the tokens of the caption autoregressively using a greedy search approach. A single-layer LSTM is used and its hidden size is set to 256. The dropout is set to 0.3. While the image encoder module is non-trainable, the LSTM decoder module is trainable. During training, the cross-entropy loss is minimized, which is computed using the output logits and the tokens in the gold caption. Weights are optimized using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0001. Training is halted when the validation loss does not improve for 10 consecutive epochs.

The results of the image captioning in terms of BLEU scores are shown in Table 5. We observe that the BLEU scores of the generated image captions are much lower than the translation-based captions.

This is not very surprising because automatic captioning is free to choose a very different aspect of the image or use wording very different from the reference caption. BLEU only checks for n-gram overlap between the caption and the reference. Therefore, we perform a manual evaluation to further analyze the performance of the image caption generation model.

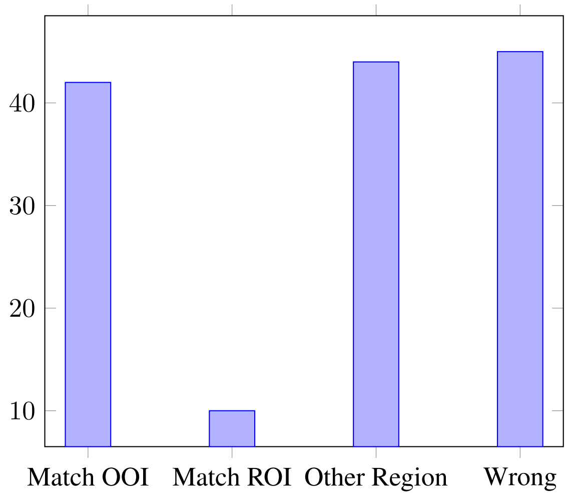


Figure 4: Manual Evaluation of Sampled Generated Captions.

4.3.1. Manual Evaluation

A sample of about 10% of the generated captions was manually evaluated and categorized into the following classes:

Match OOI for captions that describe the object of interest provided in the reference caption, exactly or closely.

Match ROI for captions that describe a different object within the region of interest.

Other Region for captions that describe an object in the image that is outside the region of interest.

Wrong for captions that do not describe any object in the associated image.

Figure 4 presents the result of the manual evaluation of the sampled machine-generated captions. From the evaluated sample, it was observed that about 68% of the generated captions correctly describe an object in

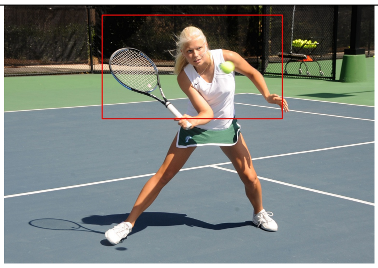
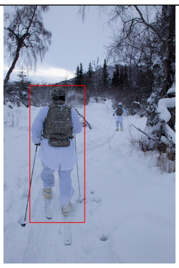
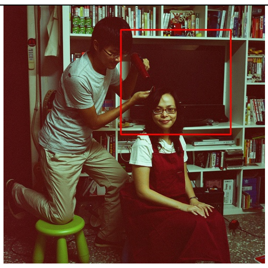
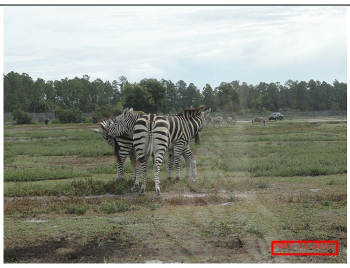
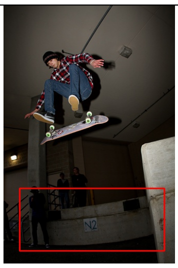
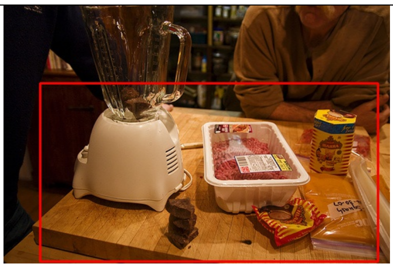
Match OOI			
			
Reference	Wata yarinya a filin wasan tanis tana shirin buga kwallon	**Reference**	mutum na biyu yana gudun kan dusar kankara
Gloss	A girl on the tennis court is preparing to hit the ball	**Gloss**	second man skiing in snow
Model	mutumin da ke wasan tennis	**Model**	mutum yana kan kankara
Gloss	the person playing tennis	**Gloss**	person is on snow
Match ROI		**Other Region**	
			
Reference	TALABIJIN a tsaye.	**Reference**	hatimin kwanan wata a kusurwar hoton
Gloss	TV on the stand.	**Gloss**	the date stamp in the corner of the pic
Model	mutum yana sanye da tabarau	**Model**	alfadari a cikin ciyawa
Gloss	person wearing glasses	**Gloss**	zebra in the grass
Wrong			
			
Reference	babban siminti	**Reference**	wani bulon katako da ke zaune a kan tebur
Gloss	large cement block	**Gloss**	a wooden block sitting on the table
Model	mutum yana kan kankara	**Model**	wani mutum yana cin abinci
Gloss	person is on snow	**Gloss**	a person eating food

Table 6: Manual classification of the qualities of sampled region of interest captions taken from the challenge dataset.

the image. Of this number, about 54% of the captions describe an object in the region of interest. However, most of the descriptions, although correct, do not match the description given in the reference caption (our evaluation does not quantify this aspect.)

This explains the low BLEU scores reported in Table 2. A more appropriate metric may be needed, therefore, to correctly measure the performance of such systems.

In Table 6, we provide examples of each of these manual evaluation classes.

5. Conclusion and Future Work

We present the HaVG, the multimodal dataset suitable for English→Hausa machine translation, image captioning, and multimodal research.

The dataset is freely available for research and non-commercial usage under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License[10] at: http://hdl.handle.net/11234/1-4749.

[10] https://creativecommons.org/licenses/by-nc-sa/4.0/

In future versions of the HaVG, we plan to create the dataset from scratch without relying on an initial MT system and post-editing. Other future works include *i)* organizing a shared task using the HaVG, *ii)* extending the HaVG corpus for Visual Question Answering (VQA).

6. Acknowledgements

This work has received funding from the grant 19-26934X (NEUREM3) of the Czech Science Foundation, and has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ. This work is also partially funded by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

7. Bibliographical References

Abdulmumin, I. and Galadanci, B. S. (2019). hauWE: Hausa Words Embedding for Natural Language Processing. In *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*, pages 1–6. IEEE.

Abubakar, A. I., Roko, A., Bui, A. M., and Saidu, I. (2021). An Enhanced Feature Acquisition for Sentiment Analysis of English and Hausa Tweets. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(9):102–110.

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Gehring, J., Michael, A., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional Sequence to Sequence Learning. In Doina Precup et al., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1243–1252, Sydney, Australia. PMLR.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. (2021). The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

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

Inuwa-Dutse, I. (2021). The first large scale collection of diverse Hausa language datasets. *arXiv e-prints*.

Jaggar, P. J. (2006). Hausa. *Elsevier Ltd*, pages 222–225.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73.

Lin, H., Meng, F., Su, J., Yin, Y., Yang, Z., Ge, Y., Zhou, J., and Luo, J. (2020). Dynamic Context-guided Capsule Network for Multimodal Machine Translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329, New York, NY, USA. Association for Computational Linguistics.

Liu, P., Cao, H., and Zhao, T. (2021). Gumbel-Attention for Multi-modal Machine Translation. *CoRR*.

Long, Q., Wang, M., and Li, L. (2021). Generative Imagination Elevates Machine Translation. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748. Association for Computational Linguistics.

Parida, S., Bojar, O., and Dash, S. R. (2019). Hindi Visual Genome: A Dataset for Multi-Modal English to Hindi Machine Translation. *Computación y Sistemas*, 23(4).

Parida, S., Panda, S., Biswal, S. P., Kotwal, K., Sen, A., Dash, S. R., and Motlicek, P. (2021a). Multimodal neural machine translation system for English to Bengali. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 31–39, Online (Virtual Mode), September. INCOMA Ltd.

Parida, S., Panda, S., Kotwal, K., Dash, A. R., Dash, S. R., Sharma, Y., Motlicek, P., and Bojar, O. (2021b).

Nlphut's participation at wat2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 146–154.

Popel, M. and Bojar, O. (2018). Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA. MIT Press.

Schlippe, T., Djomgang, E. G. K., Vu, N. T., Ochs, S., and Schultz, T. (2012). Hausa Large Vocabulary Continuous Speech Recognition. In *Spoken Language Technologies for Under-Resourced Languages*.

Schultz, T. (2002). GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University. In *Seventh International Conference on Spoken Language Processing*, pages 345—348.

Sennrich, R. and Zhang, B. (2019). Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Tukur, A., Umar, K., and Muhammad, A. S. (2019). Tagging Part of Speech in Hausa Sentences. In *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, pages 1–6.

Turner, T. D. (2021). Hausa Songs in Algeria: sounds of trans-Saharan continuity and rupture. *The Journal of North African Studies*, pages 1–29, may.

Umar, M. S. (2013). Hausa Traditional Political Culture, Islam, and Democracy: Historical Perspectives on Three Political Traditions. In *Democracy and Prebendalism in Nigeria*, pages 177–200. Palgrave Macmillan US, New York.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., and Uszkoreit, J. (2018). Tensor2tensor for neural machine translation. In *Proc. of AMTA (Volume 1: Research Papers)*, pages 193–199.