

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

Idris Abdulmumin^{1,6}, Satya Ranjan Dash², Musa Abdullahi Dawud², Shantipriya Parida³, Shamsuddeen Hassan Muhammad^{4,5}, Ibrahim Sa'id Ahmad⁶, Subhadarshi Panda⁷, Ondřej Bojar⁸, Bashir Shehu Galadanci⁶ and Bello Shehu Bello⁶

¹Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria; ²School of Computer Applications, KIIT University, Bhubaneswar, India; ³Silo AI, Helsinki, Finland; ⁴LIAAD - INESC TEC; ⁵Faculty of Sciences-University of Porto, Portugal; ⁶Faculty of Computer Science and Information Technology, Bayero University, Kano, Nigeria; ⁷Graduate Center, City University of New York, USA; ⁸Charles University, Faculty of Mathematics and Physics, ÚFAL, Prague, Czech Republic

Correspondence to: iabdulmumin@abu.edu.ng, sdashfca@kiit.ac.in, dawudmusa46@gmail.com, shantipriya.parida@siloi.ai, shmuhammad.csc@buk.edu.ng, isahmad.it@buk.edu.ng, bsgaladanci.se@buk.edu.ng, bsbello.cs@buk.edu.ng, spanda@gradcenter.cuny.edu, bojar@ufal.mff.cuni.cz

We present Hausa Visual Genome (HaVG), a multi-modal dataset suitable for English→Hausa machine translation, image captioning, and multimodal research.

Overview

- Neural Machine Translation (NMT) revolutionized automatic translation.
- Multi-modal Machine Translation (MMT) enables the use of visual information to enhance the quality of translations, supplementing the missing context and providing cues to the MT system for better disambiguation.
- Absence of sufficient training data in many languages limits the benefits of such systems.



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 court

Fig 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.

Data Collection

- Collect the English captions from Visual Genome.
- Translate (32,923) English sentences into Hausa using Google Translate.
- Post-edit the translation using annotation web page (as shown in Figure 2).

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

Fig 2: Annotation web page

Dataset

Set	Sentences	Tokens	
		English	Hausa
HaVG Train	28,930	1,47,219	1,44,864
D-Test	998	5,068	4,978
E-Test	1,595	8,079	7,952
C-Test	1,400	8,411	9,514

Table 1: Hausa Visual Genome Dataset Statistics

Text-Only Translation

- Used *Transformer* model as implemented in OpenNMT-py.
- Subword units were constructed using the word pieces algorithm.
- Vocabulary of 32k subword types jointly for both the source and target languages, sharing it between the encoder and decoder.
- Single GPU training followed the standard Noam learning rate decay.
- Starting learning rate was 0.2 and we used 8000 warm-up steps.

Multimodal Translation

- The list of object tags for a given image extracted using the pre-trained Faster R-CNN with ResNet101-C4.
- We pick the top 10 object tags based on their confidence scores.
- Object tags are appended to the English sentence which is to be translated to Hausa.
- The concatenation is done using the special token ## as the separator.
- The English sentences along with the object tags are fed to the encoder of a text-to-text transformer model.
- The decoder generates the Hausa translations autoregressively.

Text-Only Vs Multimodal

- The automatic evaluation suggests that text-only translation performs better on both the E-Test and C-Test compared to the multimodal translation.
- Manual verification shows that multimodal system was able to resolve ambiguity and generate a more appropriate translation of the given source sentence (see Figure 3 for an example).

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

Fig 3: Text-only Vs Multimodal Machine Translation

Image Captioning

- The model consists of three modules: an encoder, fusion, and decoder.
- **Encoder:** The features of the entire image, as well as features of the sub-region, are considered to train the model.
- **Fusion:** The final feature vector obtained by simple concatenation of features from the region and features from the entire image.
- **Decoder:** The decoder generates the tokens of the caption autoregressively using a greedy search approach.

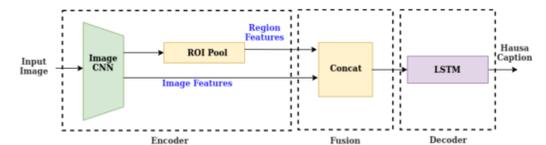


Fig 4: Architecture of the region-specific image caption generator

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 5 present result of manual evaluation of the sampled machine-generated captions and Figure 6, examples of each of these manual evaluation classes.

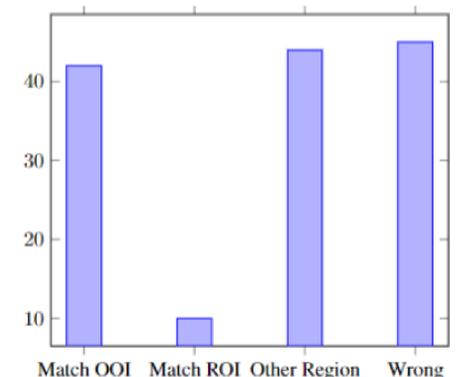


Fig 5: Manual Evaluation of Sampled Generated Captions.

Image	Text
	Match OOI Reference Wata yarinya a filin wasan tennis tana shirin biga kwallon Gloss A girl on the tennis court is preparing to hit the ball System output mutumin da ke wasan tennis Gloss the person playing tennis
	Match ROI/Other Region Reference TALABJIN a soyay. Gloss TV on the stand. System output mutum yana sanye da tabarau Gloss person wearing glasses
	Wrong Reference hahon siminti Gloss large cement block System output mutum yana kan kankara Gloss person is on snow

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

Availability

Hausa Visual Genome available for research and non-commercial usage at: <http://hdl.handle.net/11234/1-4749>.

Acknowledgement

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 financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

