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

Idris Abdulmumin1,6, Satya Ranjan Dash2, Musa Abdullahi Dawud2, Shanthi Priya Parida3, Shamsudeen Hassan Muhammad1,5, Ibrahim Sa’id Ahmad4, Subhadashi Panda2, Ondřej Bojar4, Bashir Shehu Galadanci2 and Bello Shehu Bello5

1Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria; 2School of Computer Applications, KIIT University, Bhubaneswar, India; 3SI Al, Helsinki, Finland; 4LIAD4 - 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, UFAL, Prague, Czech Republic

Correspondence to: 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

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.

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).

Dataset

<table>
<thead>
<tr>
<th>Set</th>
<th>Tokens</th>
<th>Sentences</th>
<th>English</th>
<th>Hausa</th>
</tr>
</thead>
<tbody>
<tr>
<td>HaVG Train</td>
<td>28,930</td>
<td>1,47,219</td>
<td>1,44,864</td>
<td></td>
</tr>
<tr>
<td>D-Test</td>
<td>998</td>
<td>5,068</td>
<td>4,978</td>
<td></td>
</tr>
<tr>
<td>E-Test</td>
<td>1,595</td>
<td>8,079</td>
<td>7,952</td>
<td></td>
</tr>
<tr>
<td>C-Test</td>
<td>1,400</td>
<td>8,411</td>
<td>9,514</td>
<td></td>
</tr>
</tbody>
</table>

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 English and Hausa.
• Starting learning rate was 0.2 and we used 8000 warm-up steps.

Multimodal Translation

• The list of object tags for a given image is used as features of the sub-region, are considered to be features of the entire image.
• Fusion: The final feature vector obtained by simple concatenation of features from the region and features from the entire image.

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 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 sentence autoregressively using a greedy search approach.