computation

MDPI

Article

Multimodal Multisource Neural Machine Translation: Building Resources for Image Caption Translation from European Languages into Arabic

Roweida Mohammed [1], Inad Aljarrah [1], Mahmoud Al-Ayyoub [2,3,*] and Ali Fadel [3]

[1] Computer Engineering Department, Jordan University of Science and Technology, Irbid 22110, Jordan
[2] AI Research Center, College of Engingeering and IT, Ajman University, Ajman 346, United Arab Emirates
[3] Computer Science Department, Jordan University of Science and Technology, Irbid 22110, Jordan
* Correspondence: m.alshboul@ajman.ac.ae

Abstract

Neural machine translation (NMT) models combining textual and visual inputs generate more accurate translations compared with unimodal models. Moreover, translation models with an under-resourced target language benefit from multisource inputs (source sentences are provided in different languages). Building MultiModal MutliSource NMT (M^3S-NMT) systems require significant efforts to curate datasets suitable for such a multifaceted task. This work uses image caption translation as an example of multimodal translation and presents a novel public dataset for translating captions from multiple European languages (viz., English, German, French, and Czech) into the distant and under-resourced Arabic language. Moreover, it presents multitask learning models trained and tested on this dataset to serve as solid baselines to help further research in this area. These models involve two parts: one for learning the visual representations of the input images, and the other for translating the textual input based on these representations. The translations are produced from a framework of attention-based encoder–decoder architectures. The visual features are learned from a pretrained convolutional neural network (CNN). These features are then integrated with textual features learned through the very basic yet well-known recurrent neural networks (RNNs) with GloVe or BERT word embeddings. Despite the challenges associated with the task at hand, the results of these systems are very promising, reaching 34.57 and 42.52 METEOR scores.

Keywords: natural language processing; neural machine translation; multilingual multisource translation; image caption translation; attention-based NMT; convolutional neural networks; bidirectional recurrent neural networks; GloVe embeddings; BERT embeddings

check for updates

Academic Editor: Filippo Palombi

Received: 22 June 2025
Revised: 4 August 2025
Accepted: 5 August 2025
Published: 8 August 2025

Citation: Mohammed, R.; Aljarrah, I.; Al-Ayyoub, M.; Fadel, A. Multimodal Multisource Neural Machine Translation: Building Resources for Image Caption Translation from European Languages into Arabic. *Computation* **2025**, *13*, 194. https://doi.org/10.3390/computation13080194

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As its name implies, machine translation (MT) focuses on automatically converting text from a source language to a target language. This subfield of computational linguistics holds significant importance and offers diverse applications. Historically, numerous techniques and methods for MT have been proposed, primarily relying on parallel corpora—datasets where each source language sentence is paired with one or more translations in the target language. These early approaches were largely statistical, leading to the development of Statistical MT (SMT) [1]. More recently, the integration of neural networks into MT, giving rise to Neural MT (NMT), has proven exceptionally effective [2]. Distinct from earlier

methods, NMT techniques analyze the entire sentence within its context before translating it. This characteristic, among others, has dramatically improved translation quality, often approaching that of human experts.

The conventional approach in MT typically involves a single source language and a singular modality, namely text. However, several researchers, as highlighted by works like [3,4], have increasingly focused on expanding this paradigm to incorporate *multiple source languages* or *multiple modalities*, such as text combined with images. In the multisource scenario, the premise is straightforward: each source sentence is available in several linguistic variations. Conversely, the multimodal setting pairs each source sentence with an accompanying image, where the sentence itself serves as a caption for that image. The progress of *multimodal multisource neural machine translation (M^3S-NMT)* has been significantly propelled by an annual shared task at the renowned Workshop on Machine Translation (WMT) [4–6]. Since its introduction, this shared task has attracted the attention of many researchers [7]. https://www2.statmt.org/wmt25/translation-task.html (accessed on 21 June 2025) Throughout the years, it has witnessed significant jumps in translation quality, as measured by BLEU scores [8,9]. The consistent insight gleaned from this task and its participants over the years is clear: furnishing MT models with supplementary data, whether it is additional textual translations of the input or relevant visual information, can substantially enhance translation quality. Nevertheless, it is crucial to acknowledge that integrating such multimodal data can be challenging, as models might easily become "confused." Consequently, these models necessitate meticulous design and fine-tuning to effectively leverage the auxiliary information.

For the 2018 iteration, the WMT challenge on MMT featured four European languages: English, German, French, and Czech. In this work, we augment their dataset by incorporating a crucial non-European language: Arabic. This "first of its kind" dataset, along with the other resources we utilize in this work, are publicly available. https://github.com/Roweida-Mohammed/Multisource-Multimodal-Neural-MachineTranslation-to-Arabic (accessed on 21 June 2025) Furthermore, this research endeavors to unify the processing of linguistic and visual information through the application of deep learning. It provides a novel diagram illustrating our approach within MMT. The paper outlines strategies for integrating information from different modalities into self-attentive sequence-to-sequence models. This work culminates in a framework that accepts an image and its corresponding captions in the source languages (English, German, French, and Czech). It then generates a description in the target language (Arabic).

The rest of this paper is organized as follows. The following section presents the background details necessary to understand this work along with a brief discussion of the related work. Section 3 goes into the details of the dataset curation process while Section 4 goes into the details of the methodology we followed in this work. Section 5 presents the experimentation results along with a detailed discussions of them. The paper is concluded in Section 6 with a brief discussion of the future directions of this work.

2. Background and Related Work

MT is a challenging problem that processes a sequential input and generates a sequential output. Numerous neural approaches have been developed to manage such problems. The subsequent subsections describe these models, along with a concise survey of papers that have applied them to relevant applications. Given that the Arabic dataset constitutes a unique contribution of our work, we dedicate particular attention to studies that have considered Arabic in relevant settings.

2.1. Models

We now start our discussion of the different NMT models.

2.1.1. Recurrent Neural Networks

Recurrent neural networks (RNNs) are considered as one of the primary neural network architectures for many natural language processing (NLP) tasks. They are recognized for performing the same operation on each element within a sequence, which is the basis for their "recurrent" designation [10]. Furthermore, they possess a "memory" component, which enables them to retain information from previous computations.

Like other types of neural networks, RNNs employ a procedure that handles backpropagation. The authors in [11] explain how this procedure adjusts the network's weights to minimize the discrepancy between the true output and the network's generated output. The work in [12] utilized RNN trained via BackPropagation Through Time (BPTT) and contrasted their performance with statistical language modeling. They drew comparisons to standard feedforward-based language models and to backpropagation-trained RNN. Their experiments consistently demonstrated superior results for the backpropagation-trained RNN. In the study by [13], the authors applied RNN, optimized with an algorithm known as Hessian-Free, to the task of predicting the subsequent character in a text sequence. All experiments indicated that their methodology effectively identified easily learnable words.

Machine translation, like language modeling, involves processing sequences of words as input and generating sequences of words as output. The authors in [14] proposed a combination of recursive neural networks and recurrent neural networks, termed recursive recurrent neural network (R2NN). They applied this model to decode SMT and suggested three steps for a semi-supervised training method. Furthermore, they explored the embedding method of phrase pairs. They utilized a Chinese-to-English translation dataset to conduct their experiments, and their method outperformed the baseline by 1.5 points in BLEU score.

The work in [15] proposed a joint translation and language model using RNNs that predicts target words. These predictions are based on an unbounded history of source and target words. Their joint model is built upon an RNN architecture and augmented by a layer that processes information from the input source language. They utilized the dataset from WMT2012, and their model demonstrated improved results across several test datasets.

2.1.2. Bidirectional Recurrent Neural Networks

The authors in [16] presented a straightforward extension to RNN known as Bidirectional RNN (Bi-RNN). This architecture facilitates simultaneous training of the network in both forward and backward directions. They also demonstrated Bi-RNN's improved performance compared to other artificial neural network (ANN) architectures.

The work in [17] presented an implementation for a neural language model (LM) utilizing Bi-RNN. The computational cost of training was substantial; therefore, the authors selected a subset of the training corpus using domain adaptation techniques. This partial corpus proved sufficient to train their network, achieving a perplexity equivalent to that of an n-gram LM trained with the entire corpus. They also demonstrated that Bi-RNN for LM performed superiorly to the unidirectional architecture in terms of both translation quality and perplexity.

2.1.3. Attention Mechanism

The encoder–decoder method discussed previously faces the challenge of encapsulating all necessary information from the source sequence into a fixed-length vector. Generally

speaking, managing long sequences is difficult, particularly when those sequences exceed the lengths encountered in the training corpus. The work in [18] demonstrated that a simple encoder–decoder can rapidly fail when the length of the source sequence increases. This problem was resolved by the attention mechanism proposed in [19]. Their method illustrates how the source sequence is not encoded into a static vector. Instead, the input sequence in the encoder maps to a sequence of vectors, and the decoder then selects from these vectors to output words in the target language.

The performance of the attention mechanism in NMT was remarkably good. For this reason, another architectural approach was proposed that replaced layers with self-attention mechanisms on both the encoder and decoder sides. This architecture is now known as Transformers [20].

2.2. Pretrained Word Embeddings

Machines are unable to comprehend raw text data in the same way humans do. They primarily operate with numerical data. Consequently, we need to create numerical representations of words that capture their meanings and semantic relationships. These representations are achieved through word embedding, and pretrained word embeddings are essential in NLP. They are learned on large datasets and then saved for deployment in diverse tasks. They function similarly to transfer learning (TL).

2.2.1. GloVe Word Embedding

Global Vector (GloVe) is a model for representing words. It constitutes an unsupervised learning method designed to create vectorized representations for words [21]. GloVe maps words into a continuous vector space where the distance between words is related to their semantic similarity. GloVe is utilized to identify connections between words, such as relationships between cities, or synonyms. It operates by analyzing the word co-occurrence matrix, which indicates how frequently two words appear together. For instance, the co-occurrence patterns of "cons" and "pros" can reveal their opposing semantic relationship.

In the work presented in [22], the authors utilized RNN to train an LM. To enhance the approach's effectiveness, they incorporated GloVe word embeddings. The results demonstrated how this approach surpassed other methods on numerous tasks. For the work in [23], pretrained word embeddings were employed to investigate their impact on multimodal NMT. They showed how GloVe word embeddings consistently improved the performance of all models in multimodal translation.

2.2.2. BERT Word Embedding

Bidirectional Encoder Representations from Transformers (BERT) is an NLP approach developed by Google [24]. It provides a deep, bidirectional, and unsupervised representation for languages. BERT was pretrained on raw text. While previous models, such as GloVe, generate a single word embedding for each word, BERT considers the context for each word. This allows it to provide distinct embeddings for words that may be orthographically similar but carry different meanings based on their contextual usage.

The authors in [25] proposed a new approach known as the BERT-fused model to merge NMT and BERT. They first utilized BERT to create word representations and then applied them to all layers in both the encoder and decoder. Their experiments demonstrate the effectiveness of this model in NMT across supervised, semi-supervised, and unsupervised settings. For the work in [26], the researchers investigated how BERT can be employed for supervised NMT. They compared various methods for integrating the BERT model into NMT. Their study was conducted using the WMT14 and IWSLT14 datasets, and their experiments indicate that BERT can contribute to improving NMT performance.

2.3. Work Focusing on the Multi30k Dataset

The authors in [27] utilized the Multi30K dataset in the task of multimodal machine translation (MMT). Their submission achieved first place for English to French and second place for English to German. Their approach focused on enhancing the attention mechanism over the visual input. They employed a filtering technique to enable the network to disregard unrelated parts of the image that would not be relevant during decoding. This was accomplished by implementing an attention mechanism between the source sentence and the CNN feature maps.

The work in [28] employed Transformer NMT within MMT. They utilized both the Multi30k and MSCOCO datasets. Their baseline NMT model was the OpenNMT implementation of the Transformer for both the encoder and decoder. For visual features, they used Detectron mask surface and domain labeling. They conducted numerous experiments and compared their results with text-only baselines, which demonstrated an improvement in performance when visual features were incorporated.

The research in [29] investigated a combination of SMT and NMT systems to achieve a merged translation. They submitted two distinct types of models. The first model comprised a combination of the VMT system and the Marian NMT model. The second model involved integrating the VMT system with the Marian, OpenNMT, and Moses models.

The work in [30] employed two Transformer-based architectures. The datasets utilized were Multi30K and EU Bookshop. The first model solely used parallel data, while the second model incorporated an Imagination model adapted to the Transformer architecture. During the training phase, their model leveraged the encoder to predict the image representation.

The authors in [31] participated in two distinct tasks, employing different approaches for each. For the first task, they utilized two methods: an ensemble of attentive NMT models constructed using the NMTPY toolkit, and word sense disambiguation (WSD). For the second task, they adopted three approaches. The first involved taking the 10 best translations from German–Czech, French–Czech, and English–Czech. These were then re-ranked using a WSD model, similar to their approach in task one. The second method involved searching for agreement among dissimilar 10-best lists. The final method was data augmentation.

The work in [32] employed ensemble models for both textual and visual information. Both types of models utilized the global attention mechanism, and ResNet-101 was used for extracting image features. The models were trained using reinforcement learning and scheduled sampling. For task two, they trained the multisource models using the same methodology as in task one.

2.4. Work Focusing on the Arabic Language

The authors in [33] utilized the BERT model and trained it specifically on Arabic text. Their pretrained model, named AraBERT, was employed to compare multilingual BERT with other state-of-the-art approaches. They detailed the training process for their Arabic-specific model using the BERT architecture and evaluated its performance on three distinct tasks: Question Answering (QA), Sentiment Analysis (SA), and Named Entity Recognition (NER). They used both Dialectal Arabic (DA) and Modern Standard Arabic (MSA) datasets. The results indicated that their model achieved superior performance for NLP tasks in Arabic across both datasets.

The authors in [34] addressed the task of NMT to translate Amharic as the source language into Arabic as the target language. They constructed a small corpus containing these two languages to train their model. They utilized both Long Short-Term Memory

(LSTM) and Gated Recurrent Unit (GRU) models and compared their performance with the output of the Google Translate service.

The work in [35] proposed an approach for translating from Arabic as the source language to Chinese as the target language. They established a baseline model for their experiments. The results demonstrated that employing tokenization during Arabic preprocessing significantly improved performance.

The work in [36] employed various neural networks for the classification of Arabic texts, drawing upon recent studies for this task. Their model was constructed using several neural network architectures, including LSTM, Feedforward Neural Network (FFNN), CNN, and Backpropagation Neural Network (BPNN). They utilized diverse corpora, some developed by the authors and others publicly available. They did not definitively identify the optimal neural network for Arabic text classification. This is because each neural network model operates differently and can be fine-tuned to enhance its results.

The work in [37] presented a new multimodal dataset for the Persian language. The dataset comprises statements extracted from YouTube videos. Additionally, a framework for multimodal sentiment analysis was proposed, designed to merge audio, visual, and textual features. Their experiments demonstrated that this framework outperformed all other unimodal classifiers.

The authors in [38] investigated methods to identify a robust set of features for Arabic MT. They combined three models (SkipGram, CBOW, and FastText) using the UN dataset. Additionally, a deep learning architecture was employed for the task of MT between English and Arabic texts. They primarily utilized Bidirectional LSTM (Bi-LSTM) and CNN. The results indicate that their work achieved a higher BLEU score compared to other studies.

The author of [39] addressed the scarcity of large Arabic-labeled image datasets by creating an Arabic ImageNet. The researcher successfully mapped Arabic synsets from Arabic WordNet (AWN) to ImageNet's existing synsets. Initially, he found direct Arabic equivalents for 1219 ImageNet synsets. By leveraging the hierarchical structure of synsets and recursively searching for Arabic hypernyms in AWN, he significantly expanded the dataset. Ultimately, Arabic labels were found for over 99.9% of ImageNet's synsets, effectively labeling 11 million images. This effort provides a crucial resource for advancing Arabic computer vision tasks.

The authors of [40] introduced AraTraditions10k, a culturally rich image–caption dataset featuring 10,000 images of Arabic traditions, each annotated with five Arabic captions and professional English translations. It fills a critical gap in cross-lingual vision–language research by offering high-quality, balanced content across diverse cultural themes. The authors trained and evaluated models for tag and sentence recommendation using this dataset, showing strong performance and improved cultural relevance.

3. Dataset

This section delves into the details of the dataset preparation and validation. It can benefit any researcher who is interested in expanding this dataset (or the original Multi30K dataset) into new languages. We start by discussing the source of our dataset, which is the Multi30K dataset.

3.1. Multi30K Dataset

The prominent dataset utilized for MMT research is known as Multi30K [41]. This dataset is an extension of the Flickr30K dataset [42], comprising image descriptions where five English descriptions were collected for each of 31,014 images. One of these five English descriptions was professionally translated into German by human translators (without the translators being shown the images) to construct a corpus with corresponding images [41].

Subsequently, the authors in [43] gathered additional images and translations to create development and test datasets. The dataset was then further expanded to include French [5] and Czech [4] translations. This expansion resulted in 31,014 translated entries and three distinct MMT subtasks in 2018: English→French, English→German, and English→Czech pairs, as illustrated in Figure 1.

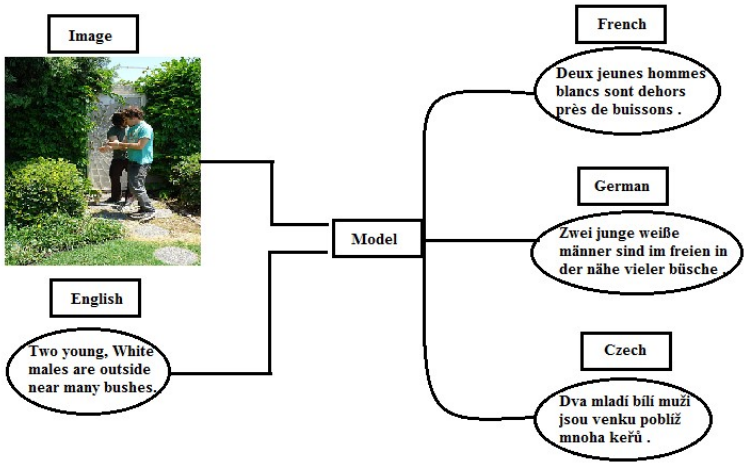


Figure 1. The three separate MMT bilingual subtasks in WMT 2018.

In 2018, a subtask of multimodal multisource translation was proposed. Its goal was to construct an MT system that performs many-to-one translation, leveraging images and descriptions in English, German, and French to produce a translation into Czech. The utilization of additional resources, such as image caption datasets and MT systems, is typically supported, and submissions that employ them are designated as "unconstrained." The training dataset for this task consists of 29,000 images, the development dataset comprises 1000 images, and the test dataset contains 1041 images. The data split for Multi30K aligns with that of Flickr30K. For every image, one description is translated into French, German, and Czech. For the 2017 and 2018 competitions, new test datasets of 1000 images each were created, following the same procedure as the Flickr30K test dataset creation. Furthermore, a smaller dataset, specifically focused on ambiguous words, was derived from the MSCOCO dataset. Information pertaining to the dataset is presented in Table 1. This table highlights the corpus-level differences between the languages used. In terms of length, English and French are nearly identical in their token counts, whereas Czech and German are notably shorter. Moreover, French and German largely exhibit similar character counts. To generate results for our experiments, we utilized the 2016 test dataset.

Table 1. Statistics about the translation data.

Translations (31,014 Sentences)	Tokens	Word Types	Characters	Avg. Length	Singletons
English	357,172	11,420	1,472,251	11.9	5073
French	374,680	13,861	1,751,894	12.5	6246
Czech	271,414	24,991	1,300,308	9.0	13,320
German	333,833	19,397	1,774,234	11.1	11,285

3.2. ArMulti30K: The Arabic Version of Multi30K

In 2016, the MMT task for the WMT workshop initially focused solely on English to German translation. Subsequently, French was incorporated as a target language in 2017. By 2018, the Czech language was added to the dataset. For our current research, we created and extended the Multi30K dataset to include an Arabic version for the target language.

Our experiments specifically addressed the multimodal multisource MT problem involving English, German, French, and Czech as source languages translating into Arabic. The Arabic language could present a challenging target due to its distinct stylistic characteristics compared to the other languages. Specifically, English, German, French, and Czech are written from left to right, whereas Arabic is written from right to left.

The translation was performed by two fluent English–Arabic translators. They primarily focused on translating the English captions of images into Arabic, consulting the images only when an ambiguous sentence or word required visual context for resolution. To further enhance translation quality, a professional translator subsequently reviewed the translations, making modifications where deemed necessary. Figure 2 displays samples from the Multi30K dataset in the various languages utilized in this research.



1. A man in a blue shirt is standing on a ladder cleaning a window. (En)
2. Ein mann in einem blauen hemd steht auf einer leiter und putzt ein fenster. (Gr)
3. Un homme dans une chemise bleue se tient sur une échelle pour nettoyer une fenêtre. (Fr)
4. Muž v modrém tričku stojí na žebříku a myje okno. (Cz)
5. رجل يرتدي قميصًا أزرق يقف على سلم ينظف النافذة. (Ar)



1. Two men are at the stove preparing food. (En)
2. Zwei männer stehen am herd und bereiten essen zu. (Gr)
3. Deux hommes aux fourneaux préparent à manger. (Fr)
4. Dva muži připravují u sporáku jídlo. (Cz)
5. رجلان عند الموقد يجهزان الطعام. (Ar)

Figure 2. Sample examples in the Multi30K dataset.

4. Methodology

In this section, we present the details of the methodology we followed to build our models.

4.1. Software

The system presented in this paper was constructed using OpenNMT, a Python toolkit developed by Klein et al. [44]. OpenNMT is an open-source framework focused on training models for vision-and-language tasks, as well as other multimodal tasks. It is built upon the PyTorch framework, developed by Paszke et al. [45]. Although the default model in OpenNMT features a 2-layer LSTM on both the encoder and decoder, the toolkit's flexible API readily permits the construction and training of diverse types of end-to-end deep neural networks.

Beyond our own work, OpenNMT has been successfully employed by other researchers primarily for MT [32,46–48]. It has also found application in various other tasks. These include speech-to-text (Zenkel et al. [49]), summarization (Parida et al. [50] and Dixit [51]), and visual-to-text applications (Gwinnup et al. [29]).

4.2. Preprocessing

As with many machine learning models, several preprocessing steps are necessary for each modality. We start by discussing the preprocessing steps for the textual part of the input.

4.2.1. Text Processing

For preprocessing the dataset, we combined the four source languages (English, German, French, and Czech) into a single text file. Each sentence from these four languages was appended to this final text file in a fixed order: first English, then German, French, and finally Czech. This process yielded a consolidated text file containing all sentences from the four languages. Subsequently, all sentences were converted to lower case and tokenized. We then applied the byte pair encoding (BPE) method to create a subword level vocabulary, resulting in a vocabulary of 40 K sub-word units [52]. BPE segments words in the dataset based on their corpus occurrence frequency: the more frequently a word appears, the less likely it is to be broken down into smaller sub-words. Table 2 illustrates the vocabulary size before and after the application of BPE to the dataset.

Table 2. Statistics of the dataset.

	Train Dataset (Sentences)	Development Dataset (Sentences)	Vocabulary	Vocabulary After BPE
English, French, German, and Czech	116,000	4056	50,002	38,988

Following text preprocessing, we utilized GloVe word embeddings (Pennington et al. [21]) and applied them to both the encoder and decoder. Additionally, to enhance the results, we employed the pretrained BERT word embedding (bert-base-multilingual-cased) and applied it to both the encoder and decoder (Devlin et al. [24]).

4.2.2. Image Features

For image feature extraction, we utilized the pretrained ResNet50 (He et al. [53]) and VGG-19 (Simonyan and Andrew [54]) models, both provided by OpenNMT (Klein et al. [44]), to generate feature vectors for each image. We did not resize the images; their original dimensions were maintained for feature vector extraction. We employed two distinct procedures to integrate these feature vectors into the MMT pipeline:

- Initializing the encoder hidden state using the image (IMG-e).
- Initializing the decoder hidden state using the image (IMG-d).

4.3. Experimental Design

We now discuss our experimentation starting with describing the experimental design.

4.3.1. Experiments with Multisource Text-Only MT

The objective of this experiment is to produce an Arabic sentence from a set of semantically equivalent sentences in English, German, French, and Czech.

The experiments were performed utilizing the extended Multi30K (Elliott et al. [4,5,41]) dataset. The source languages comprised English, German, French, and Czech, while Arabic served as the target language. Our dataset contained 145,000 sentences for training, 5070 for validation, and a distinct set of 5000 for testing. For this multisource text-only experiment, we employed RNN and Bi-RNN models as baselines (Schuster and Paliwal [16]). The encoder and decoder each consisted of 2 layers of Bi-LSTM with an embedding size of 500. Due to memory constraints, the batch size was reduced to 25. Figure 3 illustrates the MT setup for multisource text-only.

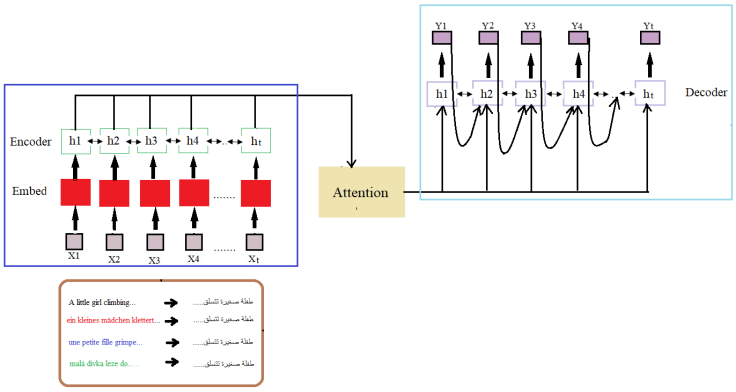


Figure 3. Machine translation of multisource text-only.

4.3.2. Experiments with Multimodal Multisource Machine Translation

This experiment applied images along with the source sentences to produce the target sentences. We divided this experiment into two phases. In the first phase, image feature vectors extracted by ResNet50 (He et al. [53]) were applied to the RNN and Bi-RNN models. In the second phase, the ResNet50 network was replaced with the VGG-19 (Simonyan and Andrew [54]) network, and the same experiments were carried out.

We utilized the same model architectures as those in the previous multisource text-only MT experiments, with the modification of applying the image feature vectors to the models. Firstly, the image was used to initialize the encoder side for both RNN and Bi-RNN models. Secondly, the image was applied to initialize the decoder side for both models. Figure 4 illustrates the architecture of this experiment.

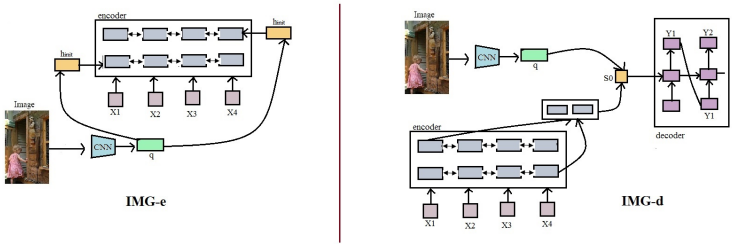


Figure 4. Multimodal multisource machine translation.

4.4. Training and Evaluation

The set of shared hyperparameters utilized throughout this research are detailed in Table 3. All models were trained for a maximum of 30 epochs. At the conclusion of each epoch, the model's performance was evaluated using the Multi30K validation dataset. After model training, we individually decoded test dataset translations from each run. We then restored all segmentation artifacts, including those from BPE, to ensure comparability of the results. We employed OpenNMT's implementation of the BLEU [8] and METEOR [55] scores for evaluating the performance of the models after generating the target sentences.

The BLEU (Bilingual Evaluation Understudy) score primarily measures the precision of n-grams (sequences of 'n' words) in the machine translation that also appear in the reference translations. It aims to see how much of the machine's output is "correct" relative to the human references.

It is computed as follows. First, we start by counting how many unigrams, bigrams, trigrams, and 4-grams in the machine translation appear in any of the reference translations. To prevent a machine translation from getting a high score just by repeating common words, the counts are "clipped" to the maximum frequency of that n-gram in any single reference. A penalty is applied if the machine translation is significantly shorter than

the reference translations. This prevents very short, but highly precise, translations from scoring artificially high. The clipped n-gram precisions (for different n-gram lengths) are combined using a geometric mean. This penalizes missing any n-gram match. The geometric mean is multiplied by the brevity penalty to get the final BLEU score, which ranges from 0 to 1 (or 0 to 100). Higher scores indicate better quality.

Table 3. The set of hyperparameters utilized in this work.

Hyper Parameter	Value
Weight initialization	ResNet50 [56]
Embedding size	500
Optimizer	Adam
Learning rate	0.002
Dropout	0.3
Number of layers	2 layers of Bi-LSTM
Epochs	30
Batches	25
Word embedding	GloVe/BERT

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) aims to address some limitations of BLEU by considering not just exact word matches but also synonyms, stemmed words, and paraphrase matches, and by incorporating recall (how much of the reference is covered by the translation) alongside precision, with a heavier weight on recall. It also considers word order.

It is computed as follows. It first creates a flexible alignment between the machine translation and a reference translation. This alignment can match words based on exact matches, stemmed matches (e.g., "running" and "ran"), and synonym matches (using resources like WordNet). Based on this alignment, it calculates unigram precision (matched words/words in machine translation) and unigram recall (matched words/words in reference translation). Precision and recall are combined into a harmonic mean (F-mean), typically with recall weighted higher than precision. A penalty is applied based on how fragmented the matched words are in the machine translation compared to the reference. This implicitly rewards correct word order. The F-mean is multiplied by $(1 - \text{Penalty})$ to get the final METEOR score, which also ranges from 0 to 1. Higher scores indicate better quality, with a stronger correlation to human judgments.

5. Results and Discussions

All experimental results are presented as percentages (0–100), with 100 indicating optimal quality. The reported BLEU and METEOR values are rounded to two decimal places for enhanced readability. These scores are calculated on the test-2016 dataset. Each experiment is discussed separately to simplify the presentation.

5.1. Multisource Text-Only Results

The first experiment serves as a baseline for comparison with subsequent experiments. The results for this initial experiment were generated rapidly as images were not incorporated; it relied solely on text. Table 4 presents the results for the multisource text-only configuration. We employed 2-layer LSTM-based RNN and Bi-RNN models. For both the encoder and decoder, we applied GloVe and BERT word embeddings. We can observe that the Bi-RNN model, when combined with BERT word embeddings, achieved higher results and outperformed the other models. The results indicate that utilizing the Bi-RNN model with BERT word embeddings in both the encoder and decoder leads to superior scores, owing to its enhanced ability to recognize context in both directions.

Table 4. Results of the multisource text-only experiment.

	Model	BLEU	METEOR
Text-only + GloVe	RNN	30.38	40.41
	Bi-RNN	31.65	40.42
Text-only + BERT	RNN	31.17	40.77
	Bi-RNN	32.72	41.47

Figure 5 displays several examples generated by the RNN and Bi-RNN models. The translation produced by the Bi-RNN model is nearly identical to the reference translation.

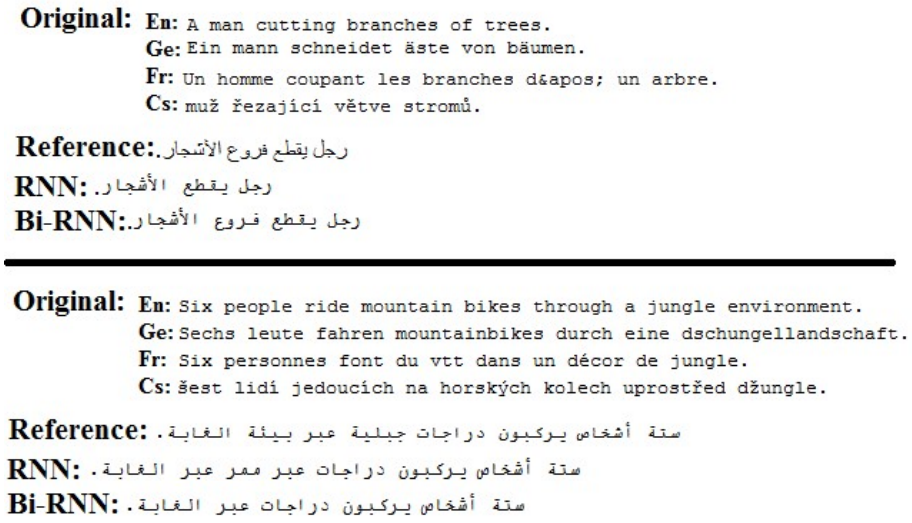


Figure 5. Examples produced by the baseline models.

Furthermore, when attempting to use the Arabic text to generate one of the other languages, the following results were achieved: Arabic to English yielded a BLEU score of 41.05; Arabic to French, a BLEU score of 36.10; Arabic to German, a BLEU score of 25.59; and Arabic to Czech, a BLEU score of 20.51. Translating into English proved to be the easiest, which is expected given that the Arabic text was originally generated from English.

An additional experiment was carried out using the FastText word representation [57]. This experiment aimed to produce an Arabic sentence (target language) by utilizing four different source languages (English, German, French, and Czech). After preprocessing the dataset, FastText was used to represent words as vectors. FastText provides two models for word embedding (Skip-Gram [58] and CBOW). We employed the Skip-Gram model for dataset representation and used a Bi-RNN model for training. The experiment yielded a BLEU score of 31.63.

5.2. Multimodal Multisource Results

For our next experiment, we incorporated images alongside the source languages. The features of these images were extracted using two pretrained models: ResNet-50 and VGG-19 networks, both of which are among the most prominent CNN architectures. We applied these extracted features at two distinct points within the RNN and Bi-RNN models. The first placement involved incorporating them into the encoder side (IMG-e) to initialize its hidden states. The second placement was in the decoder side (IMG-d). We then conducted a comparison between these two approaches.

First, we present the results obtained from the ResNet-50 network and compare them with the baseline models. Subsequently, we present the results from the VGG-19 network and compare them with the baseline models as well.

5.2.1. ResNet-50

Table 5 presents the results of the M^{3}S-NMT model using image features extracted from the ResNet-50 network, comparing them with the text-only models. From the table, it is clear that incorporating images into the MT process yields a slight but observable improvement in translation quality. The Bi-RNN model consistently outperforms the RNN model. When comparing the placement of image features, initializing the decoder's hidden states with images (IMG-d) results in slightly better translation performance compared to the encoder side (IMG-e), showing improvements of 0.8% in BLEU and 0.34% in METEOR, respectively. Furthermore, when using pretrained BERT word embeddings, the results improve to 33.19% for BLEU and 41.97% for METEOR.

Table 5. Results of the M^{3}S-NMT model with ResNet-50.

	Model	**BLEU**	**METEOR**
Text-only + GloVe	RNN	30.38	40.41
	Bi-RNN	31.65	40.42
Text-only + BERT	RNN	31.17	40.77
	Bi-RNN	32.72	41.47
With images + GloVe	RNN + IMG-e	31.72	41.01
	Bi-RNN + IMG-e	31.94	41.07
	RNN + IMG-d	32.52	41.35
	Bi-RNN + IMG-d	32.74	41.41
With images + BERT	RNN + IMG-e	32.07	41.42
	Bi-RNN + IMG-e	32.23	41.96
	RNN + IMG-d	33.07	41.90
	Bi-RNN + IMG-d	33.19	41.97

Figure 6 displays two interesting illustrative examples from our top-performing multimodal multisource MT configurations, namely IMG-e and IMG-d. In the first example, both IMG-e and IMG-d models produce a perfect translation that matches the reference. In contrast, the text-only baseline model yields an incorrect translation. Observing the image reveals the subject is 'starring' (يحدق), not 'writing' (يكتب), highlighting the baseline's error. However, in the second example, the IMG-d model nearly produces the same translation as the reference. The IMG-e model, on the other hand, translates 'at the same time' (في نفس الوقت), which is an acceptable alternative. The text-only baseline model produces a translation: 'path in the woods' (درب في الغابة), that is significantly divergent from both the image and the reference.



Original: **En:** A man in an orange hat starring at something.
Ge: Ein mann mit einem orangefarbenen hut , der etwas anstarrt .
Fr: Un homme avec un chapeau orange regardant quelque chose .
Cs: Muž v oranžovém klobouku na něco zírá .

Reference: رجل يرتدي قبعة برتقالية يحدق في شيء ما.
Bi-RNN (text-only): رجل يرتدي قبعة برتقالية يكتب على شيء ما .
Bi-RNN + IMG-e: رجل يرتدي قبعة برتقالية يحدق في شيء ما .
Bi-RNN + IMG-d: رجل يرتدي قبعة برتقالية يحدق في شيء ما .



Original: **En:** A group of people are climbing in cold weather.
Ge: Eine gruppe klettert bei kaltem wetter.
Fr: Un groupe de personnes gravissent une pente dans le froid.
Cs: Skupina lidí leze v chladném počasí.

Reference: مجموعة من الناس يتسلقون في الجو البارد.
Bi-RNN (text-only): مجموعة من الناس يتسلقون درب في الغابة .
Bi-RNN + IMG-e: مجموعة من الناس يتسلقون في نفس الوقت.
Bi-RNN + IMG-d: مجموعة من الناس يتسلقون في يوم بارد .

Figure 6. Examples of translations produced via the multimodal multisource systems using ResNet-50 and a text-only model.

5.2.2. VGG-19

In Table 6, we display the results of the M^{3}S-NMT model utilizing image feature vectors extracted using the VGG-19 network. Using these feature vectors improves translation performance compared to both the baseline text-only models and models using ResNet-50 for feature extraction. The results indicate that the Bi-RNN model, when utilizing IMG-e with BERT word embedding, outperforms other models, achieving BLEU and METEOR scores of 34.57% and 42.52%, respectively. Comparing the results of GloVe and BERT word embeddings, the pretrained BERT word embedding consistently improves the results.

Table 6. Results of the M^{3}S-NMT model with VGG-19.

	Model	BLEU	METEOR
Text-only + GloVe	RNN	30.38	40.41
	Bi-RNN	31.65	40.42
Text-only + BERT	RNN	31.17	40.77
	Bi-RNN	32.72	41.47
With images + GloVe	RNN + IMG-e	32.08	41.58
	Bi-RNN + IMG-e	33.52	42.20
	RNN + IMG-d	32.68	41.67
	Bi-RNN + IMG-d	33.06	41.78
With images + BERT	RNN + IMG-e	33.00	41.87
	Bi-RNN + IMG-e	34.57	42.52
	RNN + IMG-d	33.30	41.85
	Bi-RNN + IMG-d	34.22	42.35

Figure 7 displays an example generated by the top-performing configurations of this experiment, namely IMG-e and IMG-d. The example illustrates how the Bi-RNN with IMG-e produced a translation identical to the reference, while the text-only model and the Bi-RNN with IMG-d yielded different translations. Specifically, the text-only model translates that they are sitting in the 'museum' (متحف), which is incorrect according to both the reference and the image. The Bi-RNN with IMG-d, however, translates that they are sitting in an 'elevator' (مصعد), which is also incorrect.



Figure 7. Examples of translations produced via the multimodal multisource systems using VGG-19 and a text-only model.

5.3. Discussion of the Results

We evaluated M^{3}S-NMT models that incorporated image features into both the decoder and encoder. One of our primary objectives was to demonstrate how multisource input from different European languages could be translated into Arabic, a language that presents significant differences from its European counterparts. Additionally, we showcased methods that lead to improvements in translation performance when employing multimodal approaches compared to baseline text-only models. Furthermore, we observed that utilizing the VGG-19 network for extracting image features yielded better results than the ResNet-50 network. The use of pretrained BERT word embeddings also consistently improved the performance of the models.

To sum up, the contributions of this work are manifold. We began by creating and extending the Multi30K dataset to include an Arabic target language version, translated from the original English version. Following this, we investigated various techniques to translate the meaning of sentences from multiple source languages (English, German, French, and Czech) into Arabic, incorporating the visual modality. We employed RNN and Bi-RNN with word embeddings from GloVe and BERT, applying them to both the encoder and decoder sides to enhance results. Notably, the Bi-RNN model with BERT word embeddings consistently outperformed other models. Regarding the visual modality, we observed that the VGG model yielded better results for image feature extraction than ResNet.

6. Conclusions

This research explores multimodal multisource neural machine translation (M^{3}S-NMT), a task that leverages the Multi30K dataset. This dataset features more than 30 K images, each accompanied by descriptions in four distinct languages: English, German, French, and Czech. The images themselves serve as an additional input, providing a valuable visual context.

To push the boundaries of this task, we created an Arabic version of the Multi30K dataset, positioning Arabic as our target language. This is particularly challenging as Arabic possesses a unique linguistic style, differing significantly from the European languages in the dataset. Furthermore, research involving Arabic in this domain is relatively sparse, adding to the complexity.

6.1. Key Findings and Contributions

Our experiments demonstrate that incorporating visual context significantly supports machine translation (MT) systems, helping to identify words or concepts that might otherwise be missed in text-only translations. The findings also underscore the importance of multimodal machine translation (MMT) for a deeper understanding of multimodal systems across diverse languages.

Here are some of our key observations:

- Visual Modality's Impact: This work clearly shows that using visual modality enhances MMT performance.
- Image Feature Extraction: We primarily used image features extracted from pretrained convolutional neural networks (CNNs), specifically Residual Networks (ResNets) and VGG. Interestingly, we observed that using the VGG model yielded better results compared to ResNet for image feature extraction.
- Word Embeddings: We also incorporated pretrained word embeddings from GloVe and BERT. Our findings indicate that using the BERT model for word embedding consistently improved translation results.
- Performance Comparison: Our overall conclusion is that MMT outperforms multi-source text-only machine translation.

6.2. Limitations and Future Work

One of the limitations of this work is the modest enhancement observed for the multimodal setting on the translation quality compared to text-only MT. While the results are promising, this direction warrants further exploration into which situations can bring out the powers/benefits of multimodality and multilinguality. For example, it would be valuable to investigate how multimodal translation specifically reduces ambiguity by leveraging information from the image channel. This could offer deeper insights into the interplay between visual and linguistic data in translation. Along the lines of this reasoning, it would be useful to fully explore the effect of having multiple source languages on the Arabic translation. While the results of our work and previous ones showed that having multiple source languages can produce better results, it would be useful to consider different settings/situations and how they affect the quality of translation to Arabic.

Another important future direction is to consider benefiting from newly released Arabic-specific resources that aid in multimodal MT such as Arabic ImageNet [39] and AraTraditions10k [40]. A final limitation of this work is that we restricted ourselves to one of the most basic yet well-known architectures for the baseline models. So, as a final direction, we plan on exploring newer multimodal architectures such as Vision Transformers (ViTs) [59], Swin Transformer [60], Mamba [61], Extended LSTM (xLSTM) [62], Vision-LSTM (ViL) [63], Florence-2 [64], etc. While exploring these different architectures, it would be beneficial to probe them to try to understand what makes one model work in certain situations while others fail.

Author Contributions: Conceptualization, I.A. and M.A.-A.; methodology, R.M., I.A. and M.A.-A.; software, R.M. and A.F.; validation, R.M., I.A., M.A.-A. and A.F.; investigation, R.M., I.A. and M.A.-A.; resources, R.M. and A.F.; data curation, R.M. and A.F.; writing—original draft preparation, R.M., I.A. and M.A.-A.; writing—review and editing, M.A.-A.; visualization, R.M.; supervision, I.A. and M.A.-A.; project administration, I.A. and M.A.-A.; funding acquisition, M.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Ajman University.

Data Availability Statement: The data created in this work is available at https://zenodo.org/records/4394718#.YECQ32gzZPZ, accessed on 21 June 2025.

Acknowledgments: The authors would like to thank Ajman University for partially supporting this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Brown, P.F.; Cocke, J.; Della Pietra, S.A.; Della Pietra, V.J.; Jelinek, F.; Lafferty, J.; Mercer, R.L.; Roossin, P.S. A statistical approach to machine translation. *Comput. Linguist.* **1990**, *16*, 79–85.
2. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
3. Choi, G.H.; Shin, J.H.; Kim, Y.K. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. *arXiv* **2017**, arXiv:1709.08898.
4. Barrault, L.; Bougares, F.; Specia, L.; Lala, C.; Elliott, D.; Frank, S. Findings of the third shared task on multimodal machine translation. In Proceedings of the Third Conference on Machine Translation (WMT18), Brussels, Belgium, 31 October–1 November 2018; Volume 2, pp. 308–327.
5. Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; Specia, L. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv* **2017**, arXiv:1710.07177. [CrossRef]
6. Specia, L.; Frank, S.; Sima'An, K.; Elliott, D. A shared task on multimodal machine translation and crosslingual image description. In Proceedings of the First Conference on Machine Translation, Berlin, Germany, 11–12 August 2016; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2016; pp. 543–553.
7. Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; et al. Findings of the WMT24 general machine translation shared task: The LLM era is here but mt is not solved yet. In Proceedings of the Ninth Conference on Machine Translation, Miami, FL, USA, 15–16 November 2024; pp. 1–46.
8. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
9. Imankulova, A.; Kaneko, M.; Hirasawa, T.; Komachi, M. Towards multimodal simultaneous neural machine translation. *arXiv* **2020**, arXiv:2004.03180. [CrossRef]
10. Werbos, P.J. Backpropagation through time: What it does and how to do it. *Proc. IEEE* **2002**, *78*, 1550–1560. [CrossRef]
11. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
12. Mikolov, T.; Kombrink, S.; Burget, L.; Černockỳ, J.; Khudanpur, S. Extensions of recurrent neural network language model. In Proceedings of the 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 5528–5531.
13. Sutskever, I.; Martens, J.; Hinton, G.E. Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 1017–1024.
14. Liu, S.; Yang, N.; Li, M.; Zhou, M. A recursive recurrent neural network for statistical machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 1491–1500.
15. Auli, M.; Galley, M.; Quirk, C.; Zweig, G. Joint language and translation modeling with recurrent neural networks. In Proceedings of the EMNLP, Seattle, WA, USA, 18–21 October 2013.
16. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
17. Peris, A.; Casacuberta, F. A bidirectional recurrent neural language model for machine translation. *Procesamiento del Lenguaje Natural 55* **2015**, 109–116.
18. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259. [CrossRef]
19. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
21. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
22. Makarenkov, V.; Shapira, B.; Rokach, L. Language models with pre-trained (GloVe) word embeddings. *arXiv* **2016**, arXiv:1610.03759.
23. Hirasawa, T.; Komachi, M. Debiasing word embeddings improves multimodal machine translation. *arXiv* **2019**, arXiv:1905.10464. [CrossRef]
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 3–5 June 2019; pp. 4171–4186.
25. Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; Liu, T.Y. Incorporating bert into neural machine translation. *arXiv* **2020**, arXiv:2002.06823. [CrossRef]

26. Clinchant, S.; Jung, K.W.; Nikoulina, V. On the use of BERT for neural machine translation. *arXiv* **2019**, arXiv:1909.12744. [CrossRef]
27. Caglayan, O.; Bardet, A.; Bougares, F.; Barrault, L.; Wang, K.; Masana, M.; Herranz, L.; van de Weijer, J. LIUM-CVC submissions for WMT18 multimodal translation task. *arXiv* **2018**, arXiv:1809.00151. [CrossRef]
28. Grönroos, S.A.; Huet, B.; Kurimo, M.; Laaksonen, J.; Merialdo, B.; Pham, P.; Sjöberg, M.; Sulubacak, U.; Tiedemann, J.; Troncy, R.; et al. The MeMAD submission to the WMT18 multimodal translation task. *arXiv* **2018**, arXiv:1808.10802. [CrossRef]
29. Gwinnup, J.; Sandvick, J.; Hutt, M.; Erdmann, G.; Duselis, J.; Davis, J. The AFRL-Ohio State WMT18 multimodal system: Combining visual with traditional. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 612–615.
30. Helcl, J.; Libovickỳ, J.; Variš, D. CUNI system for the WMT18 multimodal translation task. *arXiv* **2018**, arXiv:1811.04697. [CrossRef]
31. Lala, C.; Madhyastha, P.S.; Scarton, C.; Specia, L. Sheffield submissions for WMT18 multimodal translation shared task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 624–631.
32. Zheng, R.; Yang, Y.; Ma, M.; Huang, L. Ensemble sequence level training for multimodal mt: OSU-Baidu WMT18 multimodal machine translation system report. *arXiv* **2018**, arXiv:1808.10592. [CrossRef]
33. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. *arXiv* **2020**, arXiv:2003.00104.
34. Gashaw, I.; Shashirekha, H. Amharic-Arabic Neural Machine Translation. *arXiv* **2019**, arXiv:1912.13161. [CrossRef]
35. Aqlan, F.; Fan, X.; Alqwbani, A.; Al-Mansoub, A. Improved Arabic–Chinese machine translation with linguistic input features. *Future Internet* **2019**, *11*, 22. [CrossRef]
36. Wahdan, K.A.; Hantoobi, S.; Salloum, S.A.; Shaalan, K. A systematic review of text classification research based on deep learning models in arabic language. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 6629–6643. [CrossRef]
37. Dashtipour, K.; Gogate, M.; Cambria, E.; Hussain, A. A novel context-aware multimodal framework for persian sentiment analysis. *arXiv* **2021**, arXiv:2103.02636. [CrossRef]
38. Bensalah, N.; Ayad, H.; Adib, A.; El Farouk, A.I. Arabic Machine Translation Based on the Combination of Word Embedding Techniques. In *Intelligent Systems in Big Data, Semantic Web and Machine Learning*; Springer: Cham, Switzerland, 2021; pp. 247–260.
39. Alsudais, A. Extending ImageNet to Arabic using Arabic WordNet. In Proceedings of the First Workshop on Advances in Language and Vision Research, Online, 9 July 2020.
40. Al-Buraihy, E.; Wang, D.; Hussain, T.; Attar, R.W.; AlZubi, A.A.; Zaman, K.; Gan, Z. AraTraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging. *Sci. Rep.* **2025**, *15*, 1–29. [CrossRef]
41. Elliott, D.; Frank, S.; Sima'an, K.; Specia, L. Multi30k: Multilingual english-german image descriptions. *arXiv* **2016**, arXiv:1605.00459.
42. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
43. Frank, S.; Elliott, D.; Specia, L. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Nat. Lang. Eng.* **2018**, *24*, 393–413. [CrossRef]
44. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A.M. OpenNMT: Open-source toolkit for neural machine translation. *arXiv* **2017**, arXiv:1701.02810.
45. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the NIPS, Long Beach, CA, USA, 4–9 December 2017.
46. Meng, Y.; Ren, X.; Sun, Z.; Li, X.; Yuan, A.; Wu, F.; Li, J. Large-scale pretraining for neural machine translation with tens of billions of sentence pairs. *arXiv* **2019**, arXiv:1909.11861. [CrossRef]
47. Senellart, J.; Zhang, D.; Wang, B.; Klein, G.; Ramatchandirin, J.P.; Crego, J.M.; Rush, A.M. OpenNMT system description for WNMT 2018: 800 words/sec on a single-core CPU. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, Melbourne, Australia, 20 July 2018; pp. 122–128.
48. Gwinnup, J.; Anderson, T.; Erdmann, G.; Young, K. The AFRL WMT18 systems: Ensembling, continuation and combination. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 394–398.
49. Zenkel, T.; Sperber, M.; Niehues, J.; Müller, M.; Pham, N.Q.; Stüker, S.; Waibel, A. Open source toolkit for speech to text translation. *Prague Bull. Math. Linguist.* **2018**, *111*, 125–135. [CrossRef]
50. Parida, S.; Motlicek, P. Abstract text summarization: A low resource challenge. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5996–6000.

51. Dixit, S. Summarization on SParC. 2019. Available online: https://yale-lily.github.io/public/shreya_s2019.pdf (accessed on 21 June 2025).
52. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
54. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
55. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
57. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
58. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781. [CrossRef]
59. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
60. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
61. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. [CrossRef]
62. Beck, M.; Pöppel, K.; Spanring, M.; Auer, A.; Prudnikova, O.; Kopp, M.; Klambauer, G.; Brandstetter, J.; Hochreiter, S. xlstm: Extended long short-term memory. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107547–107603.
63. Alkin, B.; Beck, M.; Pöppel, K.; Hochreiter, S.; Brandstetter, J. Vision-LSTM: xLSTM as generic vision backbone. *arXiv* **2024**, arXiv:2406.04303.
64. Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; Yuan, L. Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 4818–4829.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.