Interspeech 2024
1-5 September 2024, Kos, Greece

Translating speech with just images

Dan Oneata[1], *Herman Kamper*[2]

[1]POLITEHNICA Bucharest, Romania
[2]Stellenbosch University, South Africa

dan.oneata@gmail.com kamperh@sun.ac.za

Abstract

Visually grounded speech models link speech to images. We extend this connection by linking images to text via an existing image captioning system, and as a result gain the ability to map speech audio directly to text. This approach can be used for speech translation with just images by having the audio in a different language from the generated captions. We investigate such a system on a real low-resource language, Yorùbá, and propose a Yorùbá-to-English speech translation model that leverages pretrained components in order to be able to learn in the low-resource regime. To limit overfitting, we find that it is essential to use a decoding scheme that produces diverse image captions for training. Results show that the predicted translations capture the main semantics of the spoken audio, albeit in a simpler and shorter form.

Index Terms: Visually grounded speech models, low-resource languages, speech translation.

1. Introduction

Imagine you are a linguist tasked with translating a foreign low-resource language, but that it is not possible to get parallel speech–translations. One possible approach is to ask native speakers to describe images using their own language. The idea would be to then use the images as an intermediate modality to understand new input speech [1, 2]. While there has been major advances in visually grounded speech models that learn from paired audio–image correspondences [3–7], no study has attempted to develop a model that can take speech and directly produce a written translation of the input. This is our goal.

Earlier work [8] has shown that it is possible to perform keyword detection in a foreign language using only images paired with unlabelled speech. The idea was to use a pretrained vision system to tag images with word labels in the high-resource target language. These tags were then used as targets to train an audio-to-keyword model, taking speech input in the foreign low-resource language. At test time, the audio model was then able to predict whether a keyword (in the target language) occured in the audio stream (of the low-resource source language). However, the model did not predict full translations of the speech input. Moreover, the study was done in an artificial setting where German was the high-resource target language (of the image tagger) and English audio was the low-resource source language.

An alternative approach is to do translation by retrieval: finding relevant existing captions for a given audio in a foreign language [1, 9, 10]. These methods project audio and images in a common embedding space. Then at test time they can map a novel audio to the caption of the closest image, thereby producing a full natural language translation. However, retrieval is limited to the dataset and requires manually provided captions.

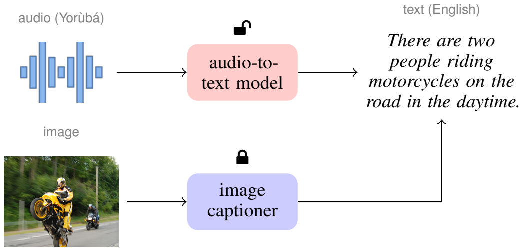

audio (Yorùbá)
audio-to-text model
text (English)
There are two people riding motorcycles on the road in the daytime.
image
image captioner

Figure 1: *Overview of our speech translation system. Given an audio in a foreign language (e.g., Yorùbá), we generate natural language translations in a high-resource language (e.g., English). We achieve this with only audio–image pairs by generating captions automatically using a pretrained image captioner and then using these as targets for an audio-to-text model.*

In this paper we propose a system that is able to directly generate natural language translations for a given foreign input audio. Our speech translation system is trained solely on audio–image pairs. The approach is illustrated in Figure 1. First, target sentences in the high-resource language (English) are generated with a pretrained image captioning system for the image associated to an audio input. Then, based on these sentences we learn an audio-to-text model, which takes as input speech in the foreign language (Yorùbá, in our case). Finally, at test time we can generate translations using the audio-to-text model, in our case translating Yorùbá speech to English text. This is done without any parallel Yorùbá–English speech–translation pairs.

In this real Yorùbá–English low-resource setting, we show that using images as an interlingua comes close to a speech translation system trained with speech–text translation pairs. In our analysis, we also show that the same system can be used in an English–English audio-to-text system that produces reasonable paraphrases of the English audio input (again using images as intermediate modality). By situating our results in terms of three toplines, we conclude that it is neither the image captioning component nor the audio-to-text architecture that limits the performance; rather, other methodological changes may be required to close the gap to human-level performance.

2. Related work

Our approach is an example of cross-modal learning (also referred to as cross-modal knowledge distillation). This type of learning is applied to transfer knowledge across different modalities, for example, from vision to depth data [11] or from vision to radio signals [12]. In terms of vision and audio—the modalities

10.21437/Interspeech.2024-903

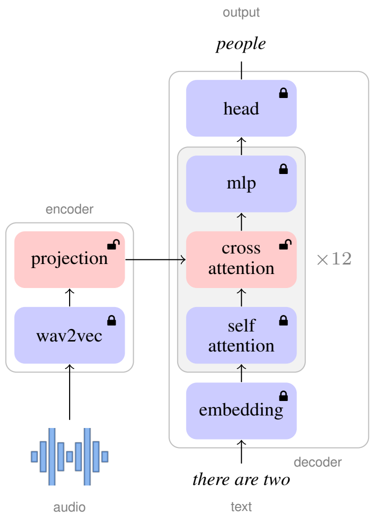


Figure 2: *Our audio-to-text model is a transformer that generates text autoregressively conditioned on audio. The network consists of learnable (🔓) cross-attention layers interspersed in a frozen (🔒) GPT-2 decoder to integrate wav2vec audio features.*

of interest here—Ayatar *et al.* [13] used visual information to perform scene detection on audio, while Owens *et al.* [14] used ambient sound to learn visual scene information. The work of Kamper *et al.* [8, 15] is more similar to our approach since it works on *speech* audio. But, as previously mentioned, they only transfer unstructured information (image tags for a fixed number of classes) rather than trying to capture the richness of natural language. In this direction, Kim and Rush [16] transfer natural text by distilling the output of large translation models to smaller ones; we differ by working across modalities.

Recently, the community has also explored aligning audio features to CLIP's [17] visual features using audio–image pairs [2, 18]. This approach allows to implicitly align the audio to a text embedding (via the visual channel), since CLIP provides by default a visual–text alignment. However, these methods are unable to directly generate novel text: they can only provide a compatibility score for a given text–audio input pair. These models are therefore used for retrieval or keyword detection.

3. Method

Our task is speech translation: given an audio in a foreign low-resource language (Yorùbá) we want to generate a natural language translation in a high-resource language (English). To this end, we learn an audio-to-text network that generates text autoregressively conditioned on the input audio signal. We assume that training data consists only of images paired with audio files that describe the contents of the corresponding image. However, in order to be able to train the audio-to-text network we need audio–text pairs. We propose to use existing state-of-the-art image captioning systems (such as BLIP [19] and GIT [20]) to generate captions for the images in the training set. These text captions paired with the associated audio files then serve as data to train a speech translation model.

While translation is our main task, our method does not make any assumptions on the input and output languages. If the two languages are the same, for example both the audio files and the image captions are in English, then our system will perform a type of paraphrasing: both the input audio and output target text would describe the same semantic information present in the image, but not necessarily using the same words. This speech paraphrasing task is related but different from the more standard task of automatic speech recognition, where the output text should contain exactly the same words as the spoken input.

3.1. Audio-to-text model

As illustrated in Figure 2, our audio-to-text model is a transformer model that is composed of two pretrained unimodal models. The encoder is the wav2vec2 XLS-R 2B model [21], which maps the input audio to a sequence of 1920-dimensional embeddings. The decoder is the GPT-2 model [22], which generates text in an autoregressive manner. We couple the encoder and decoder through cross-attention layers, which are inserted after the self-attention layers in each of the twelve GPT-2 blocks. All parameters of our model are kept fixed, with the exception of the cross-attention layers and a projection layer that maps the audio embeddings (1920D) down to the text space (768D). Leveraging existing strong pretrained models directly allows for efficient learning in our low-resource setting. Concretely, our combined transformer has over 2.3B parameters, but only 1.3% of those (29M) are learnable, making our model lean and more efficient to train. Our architecture is reminiscent of Flamingo [23] or SmallCap [24], but these operate on different modalities (images and text) and have not been employed in our tasks.

We also experimented with an alternative audio-to-text variant: mapping the audio to a soft prompt to guide the decoding [25, 26]. But we found the proposed architecture to work better for our problem. Another variant that we tried was mapping the audio to image features (instead of text) and use those as input to a frozen image captioner. But we were not able to make this alternative work as we found it difficult to model the continuous and high-dimensional image embedding space.

4. Experimental setup

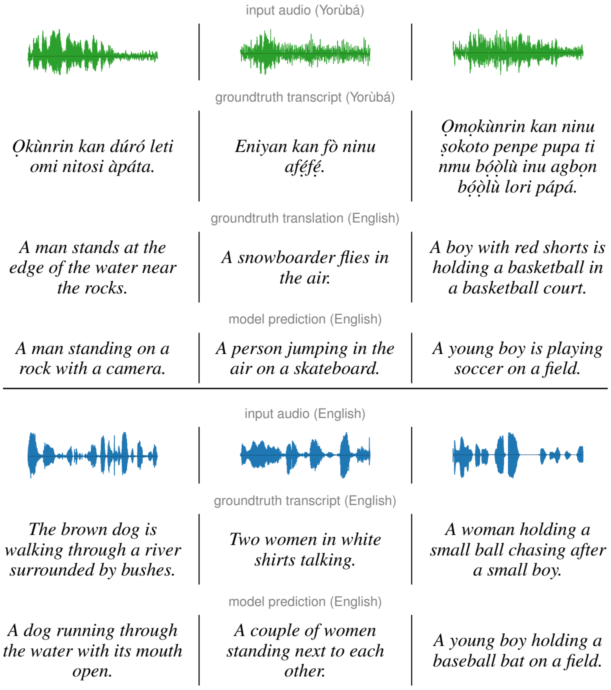
Datasets. We use two datasets in our experiments: the Flickr8k Audio Caption Corpus (FACC) [27, 28] for speech paraphrasing and its Yorùbá counterpart (YFACC) [29] for speech translation. FACC is derived from Flickr8k [27], which contains 8k images, each annotated with five text captions. Audio recordings of these captions were subsequently collected by Harwath and Glass [28], resulting in 65 hours of data. YFACC [29] consists of a subset of the FACC data (one caption per image) that was translated and recorded by a single speaker in Yorùbá; YFACC totals 13.3 hours. Although Yorùbá is spoken by roughly 44M people as a first language in West Africa, it is still considered a low-resource language.

Metric. To evaluate our model, we employ the BLEU metric, a common measure of the similarity of natural texts. Intuitively, BLEU measures the precision of a hypothesis against a set of reference sentences: what fraction of the n-grams in a prediction occurs in any of the reference sentences. We include up to four n-grams, referred to as BLEU-4. We use the `sacrebleu` library [30]. Both the speech translation and speech paraphrasing tasks are evaluated using BLEU.

Implementation. We experiment with three families of image captioning systems (BLIP [19], BLIP2 [31], GIT [20]) and three types of text decoding techniques (beam search, multinomial sampling, diverse beam search decoding [32]). Some examples are displayed in Figure 3. For each image we generate five captions using the image captioner. When training the

beam	sample	diverse
A young boy standing on a dirt road next to a field of sunflowers.	*Many small kids on a path running through a field.*	*A young boy standing on a dirt road in a field.*
A young boy standing on a dirt road next to a field.	*A boy is standing and taking photos of his plants on a dirt road.*	*A small child standing in the middle of a dirt road.*
A young boy standing on a dirt road in a field.	*A little boy using a camera to look for watermelons.*	*A child standing on a dirt road in the middle of a field.*

Figure 3: *Sample captions for the image on top using three types of decoding on the GIT image captioning model.*



input audio (Yorùbá)		
groundtruth transcript (Yorùbá)		
Ọkùnrin kan dúró leti omi nitosi àpáta.	*Eniyan kan fò ninu afẹ́fẹ́.*	*Ọmọkùnrin kan ninu ṣokoto penpe pupa ti nmu bọ́ọ̀lù inu agbọn bọ́ọ̀lù lori pápá.*
groundtruth translation (English)		
A man stands at the edge of the water near the rocks.	*A snowboarder flies in the air.*	*A boy with red shorts is holding a basketball in a basketball court.*
model prediction (English)		
A man standing on a rock with a camera.	*A person jumping in the air on a skateboard.*	*A young boy is playing soccer on a field.*
input audio (English)		
groundtruth transcript (English)		
The brown dog is walking through a river surrounded by bushes.	*Two women in white shirts talking.*	*A woman holding a small ball chasing after a small boy.*
model prediction (English)		
A dog running through the water with its mouth open.	*A couple of women standing next to each other.*	*A young boy holding a baseball bat on a field.*

Figure 4: *Examples of Yorùbá-to-English translations (top) and English-to-English paraphrases (bottom) for the visually grounded speech models trained on captions generated by GIT with diverse beam search.*

audio-to-text model we randomly pair each of the five spoken captions with each of the five generated captions. We use the AdamW optimizer [33] with a learning rate of $1 \cdot 10^{-4}$, warmed up linearly for 200 steps and then decayed linearly until the end of training. Training is run for 50 epochs and it takes around six hours on four Tesla T4 GPUs on the YFACC dataset. We keep the best model as monitored on the development set. For the translation experiments, we initialize the audio-to-text model from the best model trained on English (FACC), since this was shown to work better than random initialization [29, 34]. Our implementation is based on the HuggingFace library [35] and is available at `https://github.com/danoneata/strim`.

5. Experimental results

We present our main results and then do a sensitivity analysis to measure the impact of different image captioning methods.

5.1. Main results

Our main results are given in Table 1. These are given in terms of the BLEU score (higher is better) against a variable number n of references (captions) for each image, where n ranges from one to five. With more references, the model gets credit if a predicted n-gram occurs in any of the references; this is reasonable since different people could translate the same sentence differently. The subset of references is randomly selected from the five captions available for each image. We repeat each experiment five times and report the mean and two times standard deviation. The results for three visually grounded speech models (bottom section) are contextualized with three topline systems (top section).

Speech translation with images. The results for our visually grounded speech translation system are given in rows 4 and 5. We consider two variants, both using captions generated with the GIT image captioning model, but differing in the type of decoding used: beam search (row 4) or diverse beam search (row 5), as described in Section 4. We see that using more diverse captions rather than beam search give slightly better performance. Performance in absolute terms are modest, but BLEU can be difficult to interpret; so to give a qualitative indication of performance, the top part of Figure 4 shows sample predictions. We see that while the audio files are not transcribed verbatim, the predictions do capture the gist of the message being conveyed. The predictions are valid English sentences, but they tend to be shorter and more direct then the ground truth transcripts. There are some semantic failures, as the model hallucinates the existence of “camera” in one example and mistakes “snowboard” for “skateboard” and “basketball” for “soccer” in the other two cases.

Comparison to humans. To situate the speech translation results quantitatively, we can compare them to the three topline approaches at the top of Table 1. The first (row 1) can be seen as human performance on this dataset [36]: for a given image, we measure how well the caption given by one annotator (hypothesis) matches the captions of others annotators (reference set). Since each image has five captions, the reference set is limited to a maximum n of four. The results remain moderate in the absolute: humans reach a BLEU score of only 21.59% for $n = 4$. This suggests that even among humans there is a noticeable variance on how they describe the images. Our best visually grounded speech model, achieving 14.22% with $n = 4$, is only 7.36% behind this topline in absolute BLEU.

Comparison to supervised speech translation. Next we consider a direct audio-to-text speech translation model trained on ground truth text annotations (row 2, Table 1). This model corresponds to the typical speech translation model and we include it to both validate our architecture and put a limit on what is achievable for the visually grounded speech models. The results are even better than the annotator topline for low values of n. For this experiment (as for all those using audio at the input, rows 4–6) we always include the caption of the input audio caption in the reference set (hence the zero variance when $n = 1$). So although it might seem surprising at first that this model outperforms the annotators, the model has the advantage of having access to the Yorùbá audio. As such, this speech translation model can infer the exact words used, while the humans are likely to use different words to cover the semantics. Comparing this topline to our best visually grounded speech translation approach, we see at $n = 5$ that we are 6.19% behind in absolute BLEU.

Table 1: *BLEU scores against the English annotations from the Flickr8k test set (rows 1, 3, 6) or its corresponding Yorùbá subset (rows 2, 4, 5). All experiments involving generated captions (rows 3–6) use the GIT image captioning model.*

		input		targets		num. references				
	method	language		language	decoding	1	2	3	4	5
Toplines										
1	annotator	N/A		N/A	N/A	8.32±0.5	13.95±1.0	17.84±0.9	21.59±0.7	N/A
2	translation	Yorùbá	→	English	annotations	15.23±0.0	18.25±0.3	19.87±0.4	21.07±0.3	22.01±0.0
3	generated captions	N/A		English	beam search	9.62±0.9	17.07±1.0	22.16±0.8	25.88±0.6	29.37±0.6
Visually grounded speech models										
4	translation	Yorùbá	→	English	beam search	6.65±0.0	9.37±0.5	11.32±0.5	12.72±0.2	13.71±0.0
5	translation	Yorùbá	→	English	diverse	6.10±0.0	9.54±0.6	12.28±0.9	14.22±0.4	15.82±0.0
6	paraphrasing	English	→	English	diverse	6.56±0.5	10.45±0.8	13.10±0.7	15.45±0.4	17.46±0.9

How well can we translate with generated captions? To answer this question, we consider the performance of the generated image captions (row 3, Table 1), which are used as targets by our speech translation system. For each image we pick a random image-generated caption as the hypothesis and n annotations as the reference set. The captions are generated using the GIT model and beam search decoding. We see that the image captions yield strong results relative to the human annotations, even surpassing the inter-annotator agreement: a BLEU of 25.88% for $n = 4$. This might be caused by the fact that the BLEU metric, being a precision metric, prefers simpler descriptions, which are typically output by image captioning systems. The performance here are therefore the real upper bound for our visually grounded approach; by comparing our best BLEU of 15.82% to the 29.37% at $n = 5$, we can conclude that the generated captions are not the bottleneck if we want to improve performance. Rather, other methodological improvements are needed to take advantage of the rich supervision signal present in images.

Paraphrasing with images. As mentioned in Section 3, by using English speech input, we can easily use exactly the same approach as above to do visually grounded speech paraphrasing. Results for this model is given in row 6 of Table 1. We see that this speech paraphrasing model comes closer to the annotator and generated caption toplines (rows 1 and 2) than the speech translation models (rows 4 and 5). But note here that this English–English model is trained and evaluated on the full FACC data, which contains five times more utterances for each image than the YFACC data used for the Yorùbá-to-English speech translation experiments. The improvement over the Yorùbá–English variants (rows 4 and 5) is therefore presumably due to a combination of the larger training dataset and the language match between input and output. For a qualitative view, sample paraphrases are given in the bottom of Figure 4.

5.2. Impact of image captioning

The image captioning system directly influences the speech translation results since it provides the targets for the audio-to-text module. We therefore conduct a sensitivity analysis on three aspects: the image captioning model, the text decoding strategy, and the number of generated captions.

Concretely, we consider three image captioning models (BLIP, BLIP2, GIT) and three decoding techniques (beam search, diverse beam search, multinomial sampling) and evaluate all nine combinations. Figure 5 shows the performance of the image captions and of the resulting visually grounded models (both translation and paraphrasing). The translation models are initialized from the best paraphrasing system. In terms of the captions performance (Figure 5-left), we observe that multinomial sampling performs consistently worse than the other two variants, with beam provides attaining the best results. The performance across image models is more comparable, but the best results are achieved by the BLIP2 system.

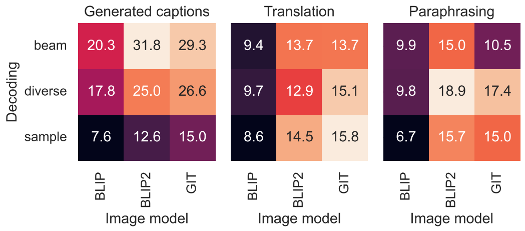


Figure 5: *Performance in terms of the BLEU score of the generated captions, speech translation and speech paraphrasing, for all nine combinations of image models and decoding strategies.*

However, these conclusions do not translate for the tasks of interest: the best variant for translation is the GIT image model with multinomial sampling (Figure 5-middle), while for paraphrasing it is BLIP2 with diverse beam search (Figure 5-right). Notably, multinomial sampling (bottom row) now yields the best performance for translation when coupled with the GIT or BLIP2 models. This suggests that more diverse targets, as illustrated in Figure 3, are important to prevent overfitting.

Since diversity is an important factor, we generated for the translation task a varying number of captions (from one to ten) using GIT captioning with multinomial sampling. Indeed, when the number of captions is very low (one or two) the performance suffers (9 to 12% BLEU), but after three captions, the performance stabilizes at around 15% BLEU score, with the maximum of 17.21% being reached when the number of captions is nine.

6. Conclusions

We have shown that it is possible to translate Yorùbá audio to English text using only visual information present in images. We are able to achieve this by training an audio-to-text model supervised by the text output of an image captioning system. To build an efficient model, we leverage state-of-the-art components such as wav2vec and GPT-2, and train only a small subset of parameters. The output predictions convey the semantics of the spoken message in natural language, but they tend to be simpler and shorter than human translations. A limitation of our model is that it tends to hallucinate words, especially when training data is limited. Future work will explore confidence estimation techniques [37] to flag these unreliable predictions.

7. Acknowledgements

We thank the anonymous reviewers for their comments and suggestions. This work was funded by European Union's Horizon research and innovation program under grant agreement No. 101070190 (AI4Trust).

8. References

[1] D. Harwath, G. Chuang, and J. R. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *Proc. ICASSP*, 2018.

[2] Y. Zhao, J. Hessel, Y. Yu, X. Lu, R. Zellers, and Y. Choi, "Connecting the dots between audio and text without parallel data through visual knowledge transfer," in *Proc. NAACL*, 2022.

[3] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *Proc. ICLR*, 2020.

[4] O. Scharenborg, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, E. Dupoux, L. Besacier, A. W. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, and M. Müller, "Speech technology for unwritten languages," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, 2020.

[5] R. Sanabria, A. Waters, and J. Baldridge, "Talk, don't write: A study of direct speech-based image retrieval," in *Proc. Interspeech*, 2021.

[6] A. Rouditchenko, A. Boggust, D. Harwath, D. Joshi, S. Thomas, K. Audhkhasi, R. Feris, B. Kingsbury, M. Picheny, A. Torralba *et al.*, "AVLnet: Learning audio-visual language representations from instructional videos," in *Proc. Interspeech*, 2021.

[7] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *J. Artif. Intell. Res.*, vol. 73, 2022.

[8] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," in *Proc. SLTU*, 2018.

[9] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, "Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms," in *Proc. ICASSP*, 2020.

[10] L. Berry, Y.-J. Shih, H.-F. Wang, H.-J. Chang, H.-Y. Lee, and D. Harwath, "M-SpeechCLIP: Leveraging large-scale, pre-trained models for multilingual speech to image retrieval," in *Proc. ICASSP*, 2023.

[11] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. CVPR*, 2016.

[12] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proc. CVPR*, 2018.

[13] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. NeurIPS*, 2016.

[14] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. ECCV*, 2018.

[15] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, 2019.

[16] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. EMNLP*, 2016.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.

[18] Y.-J. Shih, H.-F. Wang, H.-J. Chang, L. Berry, H. yi Lee, and D. Harwath, "SpeechCLIP: Integrating speech with pre-trained vision and language model," in *Proc. SLT*, 2022.

[19] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. ICML*, 2022.

[20] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "GIT: A generative image-to-text transformer for vision and language," *Trans. Mach. Learn. Res.*, 2022.

[21] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," in *OpenAI Blog*, 2019.

[23] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: A visual language model for few-shot learning," in *Proc. NeurIPS*, 2022.

[24] R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva, "SmallCap: Lightweight image captioning prompted with retrieval augmentation," in *Proc. CVPR*, 2023.

[25] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP prefix for image captioning," *CoRR*, vol. abs/2111.09734, 2021.

[26] O. Mañas, P. R. Lopez, S. Ahmadi, A. Nematzadeh, Y. Goyal, and A. Agrawal, "MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting," in *Proc. EACL*, 2023.

[27] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, 2013.

[28] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proc. ASRU*, 2015.

[29] K. Olaleye, D. Oneata, and H. Kamper, "YFACC: A Yoruba speech-image dataset for cross-lingual keyword localisation through visual grounding," in *Proc. SLT*, 2022.

[30] M. Post, "A call for clarity in reporting BLEU scores," in *Proc. WMT*, 2018.

[31] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. ICML*, 2023.

[32] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *CoRR*, vol. abs/1610.02424, 2016.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[34] L. Nortje, D. Oneata, and H. Kamper, "Visually grounded few-shot word learning in low-resource settings," *CoRR*, vol. abs/2306.11371, 2023.

[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *CoRR*, vol. abs/1910.03771, 2019.

[36] J. Wang and R. J. Gaizauskas, "Cross-validating image description datasets and evaluation metrics," in *Proc. LREC*, 2016.

[37] D. Tran, J. Z. Liu, M. W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu, N. Band, T. G. J. Rudner, K. Singhal, Z. Nado, J. van Amersfoort, A. Kirsch, R. Jenatton, N. Thain, H. Yuan, K. Buchanan, K. Murphy, D. Sculley, Y. Gal, Z. Ghahramani, J. Snoek, and B. Lakshminarayanan, "PLEX: Towards reliability using pretrained large model extensions," *CoRR*, vol. abs/2207.07411, 2022.