Grounded Sequence to Sequence Transduction

Lucia Specia, Loic Barrault, Ozan Caglayan, Amanda Duarte, Desmond Elliott, Spandana Gella, Nils Holzenberger, Chiraag Lala, Sun Jae Lee, Jindrich Libovicky, Pranava Madhyastha, Florian Metze, *Senior Member, IEEE*, Karl Mulligan, Alissa Ostapenko, Shruti Palaskar, Ramon Sanabria, Josiah Wang, and Raman Arora

***Abstract*—Speech recognition and machine translation have made major progress over the past decades, providing practical systems to map one language sequence to another. Although multiple modalities such as sound and video are becoming increasingly available, the state-of-the-art systems are inherently unimodal, in the sense that they take a single modality — either speech or text — as input. Evidence from human learning suggests that additional modalities can provide disambiguating signals crucial for many language tasks. In this article, we describe the How2 dataset, a large, open-domain collection of videos with transcriptions and their translations. We then show how this single dataset can be used to develop systems for a variety of language tasks and present a number of models meant as starting points. Across tasks, we find that building multimodal architectures that perform better than their unimodal counterpart remains a challenge. This leaves plenty of room for the exploration of more advanced solutions that fully exploit the multimodal nature of the How2 dataset, and the general direction of multimodal learning with other datasets as well.**

***Index Terms*—Grounding, multimodal machine learning, speech recognition, machine translation, representation learning, summarization.**

Manuscript received September 25, 2019; revised February 21, 2020; accepted March 9, 2020. Date of publication May 28, 2020; date of current version June 24, 2020. This work was supported in part by Johns Hopkins University with unrestricted gifts from Amazon, Facebook, Google, Microsoft, and Mitsubishi Electric Research Laboratories, in part by the Extreme Science and Engineering Discovery Environment (XSEDE) by NSF Grant ACI-1548562, in part by the Bridges system by NSF Award ACI-1445606, at the Pittsburgh Supercomputing Center, in part by the French National Research Agency (ANR) through the CHIST-ERA M2CR project, under Contract ANR-15-CHR2-0006-01, in part by the MultiMT project (H2020 ERC Starting Grant 678017), in part by the MMVC project, via an Institutional Links grant, ID 352343575, under the Newton-Katip Celebi Fund partnership, in part by the UK Department business, Energy and Industrial Strategy (BEIS), and in part by Scientific and Technological Research Council of Turkey (TUBITAK) and delivered by the British Council. The work of Shruti Palaskar, Ramon Sanabria, and Florian Metze was supported in part by Facebook and in part by Amazon grants. The work of Jindrich Libovický was supported by the Czech Science Foundation, under Grant 19-26934X. The work of Amanda Duarte was supported by the "la Caixa" Foundation (ID 100010434) which the code is LCF/BQ/DI18/11660029. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong He. This paper was presented in part at the 2018 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, Baltimore, MD, USA, August 2018. *(Corresponding author: Shruti Palaskar.)*

Florian Metze, Shruti Palaskar, and Ramon Sanabria are with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: fmetze@cs.cmu.edu; spalaska@cs.cmu.edu; ramons@cs.cmu.edu).

Nils Holzenberger and Raman Arora are with the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: nholzen1@jhu.edu; arora@cs.jhu.edu).

Lucia Specia, Ozan Caglayan, Pranava Madhyastha, and Josiah Wang are with the Department of Computing, Imperial College London, SW7 2BU London, U.K. (e-mail: l.specia@imperial.ac.uk; ozancag@gmail.com; pranava@imperial.ac.uk; josiah.wang@imperial.ac.uk).

Spandana Gella is with the Institute for Language, Cognition and Computation, University of Edinburgh, EH8 9YL Edinburgh, U.K. (e-mail: spandanagella@gmail.com).

Jindrich Libovicky is with the Center for Information and Language Processing, LMU Munich 80333, Munich, Germany (e-mail: libovicky@ufal.mff.cuni.cz).

Amanda Duarte is with the Department of Signal Theory and Communications, Universitat Politcnica de Catalunya, 08034 Barcelona, Spain (e-mail: amanda.duarte@upc.edu).

Loic Barrault and Chiraag Lala are with the Department of Computer Science, University of Sheffield, S10 2TG Sheffield, U.K. (e-mail: loic.barrault@univ-lemans.fr; clala1@sheffield.ac.uk).

Desmond Elliott is with the Department of Computer Science, University of Copenhagen, 1165 Kobenhavn, Denmark (e-mail: des.elliott@googlemail.com).

Sun Jae Lee is with the University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: ssmine0104@gmail.com).

Karl Mulligan is with the Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: karl.mulligan@jhu.edu).

Alissa Ostapenko is with the Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: aostapenko@wpi.edu).

Digital Object Identifier 10.1109/JSTSP.2020.2998415

I. Introduction

MULTIMODAL machine learning covers topics at the intersection of natural language processing, speech recognition, and computer vision [1]. Research in this area is motivated by recent advances in representation learning and the reported benefits of multi-sensory inputs: e.g. visual and tactile interaction increases infant sensitivity to colour differences over purely visual inputs [2], and psycholinguistic studies show the benefits of multiple modalities in concept representation [3]. Significant progress has been made in the last decade on major problems, including image captioning [4], visual question answering [5], image–sentence retrieval [6], and video captioning [7]. A common aspect of these problems is that they typically involve *bi-modal* learning, e.g. images and sentences in image captioning, due to the nature of the freely available datasets.

In recent years, there has been a collective effort in multilingual and multimodal representation learning, and models of visually grounded speech. In multimodal machine translation, researchers have focused on methods for integrating visual information into sequence-to-sequence models [8]–[10], and in multilingual image–sentence retrieval, it has been shown that cross-lingual sentence–sentence objectives improve retrieval performance [11], and that these findings extend to working with multiple languages [12]. In multimodal speech recognition, the image modality has been used to adapt the acoustic model [13], the language model [14] and, more recently, end-to-end systems [15], [16]. In spite of these recent successes, researchers have worked with bi- or multilingual datasets [17] that are much smaller than the datasets typically used for machine translation and speech recognition research.

TABLE I
STATISTICS OF How2 DATASET

		Videos	Hours	Clips/Sentences
300h	train	13,168	298.2	184,949
	val	150	3.2	2,022
	test	175	3.7	2,305
	held	169	3.0	2,021
2000h	train	73,993	1,766.6	-
	val	2,965	71.3	-
	test	2,156	51.7	-

This paper introduces the large-scale tri-modal How2 dataset , which consists of 2,000 hours of instructional videos with audio signals and two types of English text: closed captions of the speech and a self-written summary of the video, and crowd-sourced Portuguese translations of a subset of the human annotated transcripts (Section II). The How2 dataset affords a wide variety of bi-, tri- and multi-modal experiments; here, we focus on multimodal speech recognition (Section III), multimodal machine translation (Section IV), abstractive video summarization (Section V), and multiview learning from speech, video, and multi-lingual transcripts (Section VI). The main findings from these experiments is that learning multimodal representations almost always results in better task-specific performance, and that there are numerous opportunities for future research on effective feature integration in multimodal learning.

II. THE How2 DATASET

In the How2 dataset, we collect 79,114 English instructional videos from YouTube with English subtitles. The dataset consists of a total of 2,000 hours of video. Videos have an average length of 90 seconds [18] and manual Portuguese translations. This collection of videos and translations constitutes a large-scale resource for testing a substantial part of multimodal language processing methods in a real-world scenario.[1]

An alignment process is needed to use the audio, the English subtitles, the Portuguese translations, and the video modality together. To this end, we first re-segment the English subtitles into sentences using NLTK [19]. Then, we force-align the speech signal at the word level with an HMM-GMM pre-trained on the Wall Street Journal dataset. Finally, using the timings provided by the word alignment, we create video *clips* aligned to the initial segmented sentences. This process splits a video into a sequence of clips, aligned with the speech signal and the segmented sentences. Table I presents summary statistics of the 2000 h set and 300 h subset: the *val* and *test* sets can be used for early-stopping, model selection and evaluation; the *held* set is reserved for future evaluations or challenges. The total set (*i.e.* 2000 h) contains around 22.5 M words. The tokenized training set of *300 h* subset contains around 3.8 M (43 K unique) and 3.6 M (60 K unique) words for English and Portuguese respectively. Videos are broken down into clips, as described above, with an average length of 5.8 seconds, or 20 words of spoken language.

We collected Portuguese translations using the *Figure Eight* crowdsourcing platform, where we could reliably find Portuguese speaking crowdworkers. In order to speed-up the annotation process, we framed the translation task as a post-editing task. We first selected the best online machine translation service among three state-of-the-art services based on *Figure Eight's* workers preferences. Then, we used the translations generated by this system as a proxy and paid the crowd workers to post-edit the translation. We attempted to ensure that the workers were in fact post-editing the proxies by replacing content words of the proxy with a random Portuguese word. If the substituted word remains in the post-edit, we removed the worker from the pool and re-collected the post-edit. Each of the 200 workers used in this project have a limit of post-editing 5,000 sentences. None of them reached this threshold.

We estimated the quality of this process by comparing the performance of a translation model trained on either the post-edited translations or the machine-generated proxy-translations. The model trained on the proxy-translations performed 1 BLEU point worse on predicting the post-edited translations than the model trained on the post-edited translations, which suggests that our data collection method indeed resulted in different human-edited translation data.

To estimate the topic diversity in How2 dataset, we ran a Latent Dirichlet Allocation (LDA) [20] over the English subtitles. Then, we defined 22 clusters by analyzing empirical distances between videos and centroids. Finally, we applied a topic label to each cluster by analyzing the top words.

A. Features

In what follows, we detail the features that we extract for each modality.

1) Speech Features: For speech, we extract 40-dimensional filter bank features from 16 kHz raw speech signal using a time window of 25 ms and an overlap of 10 ms. 3-dimensional pitch features are then concatenated to form the final 43-dimensional feature vectors. The speech features of a given video are further normalized using the mean and variance statistics from that specific video.

2) Action Features (video-level): We extract action-level video features from a 3D ResNeXt-101 [21] pretrained on the Kinetics action recognition dataset [22] which comprises 400 different actions.

3) Object Features (frame-level): A ResNet-152 [23] trained on ImageNet [24] which consists of 1000 categories ranging from animals, flowers to devices and foods and so on.

4) Scene Features (frame-level): A ResNet-50 trained on Places365 [25] for scene recognition with 365 categories including, but not limited to: garden, valley, studio, theater and office.

5) Object-Level: A ResNet-152 [23] trained on ImageNet [24] which consists of 1000 categories, ranging from animals and flowers to devices and foods.

[1]The tools to download and construct the corpus are freely available at https://github.com/srvk/how2-dataset

Fig. 1. Example from How2 dataset where visual semantics can be helpful when transcribing *ukulele*.

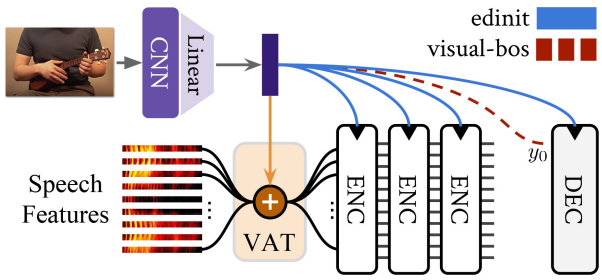


Fig. 2. Summary of proposed multimodal ASR approaches.

III. Multimodal Speech Recognition

Fig. 1 shows an illustrative example from How2 where a purely monomodal ASR is prone to transcribe the utterance *ukulele* to an homophonic equivalent *eucalylie*. Earlier work in ASR suggests that a correlated auxiliary modality can be helpful within the context of instructional videos where videos consistently provide visual cues related to the speech semantics [13]–[15], [26]. This section discusses multimodal extensions to automatic speech recognition (ASR) with vision as supporting modality. We mainly explore two different multimodal interactions: first, we apply the visual adaptive training framework [13]–[15] to S2S ASR systems; second, we propose end-to-end multimodal grounding methods inspired by previous work in image captioning [27] and multimodal neural machine translation [28], [29].

A. Training & Features

We conduct all the experiments on the *300 h* split of How2 dataset (Section II). For textual features, we first lowercase and remove punctuation from the English transcripts and then train a SentencePiece model [30] to construct a subword vocabulary of 5000 tokens. For speech, we use the 43-dimensional features (Section II) as they are. Finally for the visual modality, we explore two more pre-trained CNNs in addition to the action features described in Section II: a ResNet-152 [23] trained on ImageNet [24] for *object* recognition and a ResNet-50 trained on Places365 [25] for *scene* recognition. For all types of features, we obtain an average-pooled *(avgpool)* representation from the corresponding CNN. For object and scene-level features, we also experiment with class probabilities *(prob)* which are 1000-dimensional and 365-dimensional, respectively.

We explore two methods to obtain a single feature vector for each clip of a given video: (1) a *per-clip* representation by averaging frame level feature vectors of the clip and (2), a *per-video* representation by averaging frame level feature vectors across the whole video. We train all models with three randomly initialized instances using *nmtpy* [31]. For each instance, the best model is obtained by early-stopping on validation set word error rate (WER).

B. Baseline Model

All multimodal ASR systems in this section extend the well-known recurrent, attentive sequence-to-sequence model [32]. In the following, $\mathrm{X}=\{x_0, \ldots, x_{T-1}\}$ represents an input sequence of T speech features and f is the corresponding visual feature for that utterance. All recurrent, attention and embedding layers in the network are 320-dimensional.

The **speech encoder** is composed of 6 bidirectional LSTM layers [33], each followed by a *tanh* projection layer. The middle two LSTM layers apply a temporal subsampling [34] by skipping every other input, reducing the input sequence length T to $T/4$. The **decoder** implements the so-called *Conditional GRU* architecture [35] where an attention mechanism [32] is wrapped between two GRU [36] layers. At timestep $t{=}0$, the hidden state of the first GRU is initialized with the mean-pooled speech encoder state. The second GRU receives the output of the **attention** layer.

C. Multimodal ASR Systems

1) Visual Adaptive Training (VAT): This method fine-tunes a pre-trained ASR model using the visual modality. VAT adds a new linear layer to the model to project the visual feature vector f into the speech feature space. The projected utterance-specific *shift vector* is then added to the speech features and the network is jointly optimized until convergence:

$$x_t = x_t + (\mathbf{W_v} f + b_v) \quad t \in \{0, \ldots, T-1\} \tag{1}$$

2) Tied Initialization of Recurrent Blocks: Initializing the encoder and the decoder is an approach previously explored in multimodal machine translation [28], [29]. In order to prime the speech **encoder** with visual context, two non-linear layers are employed to learn an initial hidden state h_0^k and an initial cell state c_0^k for all the 6 LSTM layers in the encoder:

$$h_0^k = \tanh\left(\mathbf{W_h} f + b_h\right) \quad k \in \{1, \ldots, 6\} \tag{2}$$

$$c_0^k = \tanh\left(\mathbf{W_c} f + b_c\right) \tag{3}$$

The same idea can also be applied to the first GRU in the **decoder** so that its initial hidden state is visually primed:

$$h_0' = \tanh\left(\mathbf{W_d} f + b_d\right) \tag{4}$$

Finally we explore a third variant where we fuse the two approaches by *sharing* the projection parameters in equations 2 and 4. In the following, these three variants will be referred to as *einit*, *dinit* and *edinit* respectively.

3) Visual Beginning-of-Sentence: Neural decoders receive a special beginning of sentence vector as input at timestep $t{=}0$ in order to begin decoding. This vector can be either constant or learned during training, the latter being the approach taken in this

TABLE II
VISUAL ADAPTIVE TRAINING RESULTS

Granularity	CNN	Mean WER (↓) *avgpool*	*prob*
Baseline ASR		19.4	
Contrastive *restart*		19.1	
per-clip	object	18.3	18.9
	scene	18.2	19.0
per-video	object	18.2	18.7
	scene	18.1	18.8
	action	**18.0**	-

work. The disadvantage of both methods is the fact that during inference, the decoder always receives the same embedding at $t{=}0$ regardless of the input modality. Here we propose to modulate the decoder by using a visually-informed embedding for a given example i:

$$y_0^i = \mathbf{W_v} f^i + b_v \tag{5}$$

D. Experimental Results

In what follows, we report single best, mean and ensembled WER across the three training runs of each model.

1) Visual Adaptive Training: In Table II, we clearly see that *avgpool* features consistently outperform class probability features. Similarly, a *per-video* representation seems to give a slight boost compared to *per-clip* granularity. Overall, *avgpool* features reduces the WER by up to 1.4% depending on the feature type and granularity. The contrastive *restart* continues training the baseline ASR model without visual adaptation, and shows that the improvements are not a side-effect of training the model for additional epochs. But interestingly, once the learned adaptation layer is removed from the network so that the model falls back to the vanilla speech features x_t, the model still obtains around 18% WER. This seems to indicate that the effect of adaptation is indirect in the sense that it leads to a more robust ASR without necessarily relying on the visual modality.

2) End-to-End Variants: We observe that tied initialization (*edinit*) reduces the WER by 0.8% and 0.5% in terms of single best and mean scores, respectively (Table III). With ensembling, the *edinit* variant reaches the best WER (15.0%) among all the models explored. The *visual-bos* method performs on par with the *edinit*. Action features give slightly better performance for both.

Returning to example in Fig. 1, we checked how successful the systems are when transcribing the word *ukulele*. We observe that *edinit* systems with action and object features could transcribe it once (out of ten occurrences in the test set) while the baseline system could not. However, this should be taken with a grain of salt, as the *ukulele* occurs only three times in the training set.

E. Discussion

In this section, we first explored visual adaptive training for S2S ASR models and then experimented with novel multimodal extensions to S2S ASR. Our experiments showed that the method is effective for the S2S paradigm too, reaching up to 1.4% absolute WER improvement with action-level features. However, we also discovered that the adaptive system still preserves its performance even when the adaptation layer is removed during inference. We leave the analysis of this phenomenon to future work. Although end-to-end models perform better than the baseline, the difference is smaller compared to adaptive training. But when ensembling is used, the end-to-end models obtain the best WER among all models. With regard to visual representations, we show that average pooled CNN features perform better than class probabilities and the action-level features are slightly better than others.

TABLE III
END-TO-END RESULTS: ALL FEATURES ARE *AVGPOOL* AND *PER-VIDEO*. *ENS* STANDS FOR ENSEMBLE DECODING

	Feature	WER (↓) Best	Mean	Ens
Baseline	-	19.2	19.4	15.6
dinit	action	19.2	19.4	15.5
einit	action	18.8	19.2	15.6
edinit	scene	18.8	19.2	15.4
	object	18.5	18.9	15.2
	action	**18.4**	**18.9**	**15.0**
visual-bos	object	19.0	19.1	15.5
	scene	18.7	19.0	15.2
	action	**18.5**	**18.9**	**15.1**

IV. REGION-SPECIFIC MACHINE TRANSLATION

This section discusses another multimodal sequence to sequence task – Multimodal machine translation (MMT). MMT is a research field that aims to enrich textual context with additional modalities (images, videos, audio) for machine translation (MT). The assumption is that context provided by these modalities can help ground the meaning of the text and, as a consequence, generate more adequate translations. This becomes more critical when translating content that is naturally multimodal, such as picture posts on social media, audio descriptions or subtitles. MMT is especially useful when dealing with ambiguous or out-of-vocabulary words, e.g. translating *hat* into German (there is a distinction between summer hat *Hut* and winter hat *Mütze*). Even a human translator would need to see the image to decide which word to use.

Existing work on image-based MMT [37]–[39], especially neural network approaches, often incorporates images as context either as a single, global vector representation of the whole image, or by attending to grid-based representations of different local subregions of the image. We argue that such models do not exploit images effectively for MT. A global image representation provides only a summary of the image and is expected to apply equally to the whole text, but MT operates at the word level. For attention-based models, there is a mismatch between the visual unit (equally divided grid-like image subregions) and the textual unit (a word) because the subregions may not correspond to a word or cover multiple words. This makes it hard to learn

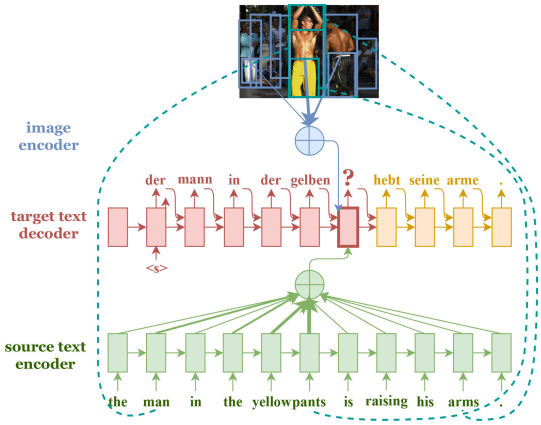


Fig. 3. Referential grounding approach uses object bounding boxes as visual units by grounding the boxes to source language words in the encoder to guide MT.

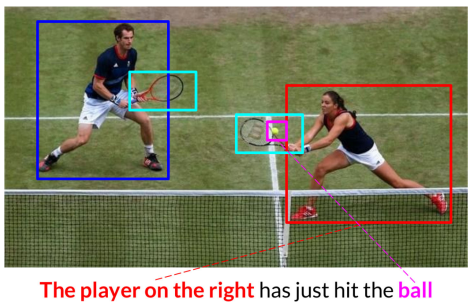


Fig. 4. Multimodal correspondences can be used to help guide translation, for example potentially resolving the gender ambiguity of the word *player* such that it can be correctly translated to its feminine form into a gender-marked language.

the correspondence between the textual and visual units during decoding due to a lack of visual consistency, especially when trained on small datasets; any assumed learned correspondences are also hard to interpret since the subregions are not well defined.

Our work in this section involves new referential grounding approaches to MT where the correspondences between the visual units (object regions) and textual units (source words) are better defined, and can then be used more effectively for translation (Fig. 3). By *object region* we mean the depiction of the entity instance from the image as single, coherent unit. The object instance can be a concrete entity, amorphous 'stuff' (*sky*, *cloud*), or a scene (*beach*, *forest*). The main motivation of using objects as a visual unit is that it may potentially result in better and more interpretable grounding. As a motivational example, Fig. 4 shows a case where the ambiguous word *player* can be translated correctly into a gender-marked language (female player) if its correspondence to the correct region in the image is identified.

Our main contributions in this section are:

1) An *implicit referential grounding* MT approach where the model jointly learns how to ground the source language in the object-level image representations, and to translate, while exploring training regimes with and without providing the correspondence as supervision;
2) An *explicit referential grounding* MT approach where object-level grounding is performed at the source side, independent of the translation model, and is subsequently used to guide MT, where we vary the ways in which the visual information is fused to the textual information.

The results of our experiments show that the proposed referential grounding models outperform existing MMT models according to automatic evaluation metrics for general quality and lexical ambiguity.

A. Dataset and Region Alignment

Unlike other sections, we build and evaluate our referential grounding MMT models on **Multi30 K** [17]. This makes the task simpler to investigate, especially as the content in the subtitles in How2 videos are often not depicted in the video. Each image in Multi30 K contains one English (EN) description taken from Flickr30K [40] and human translations into German (DE), French (FR), and Czech (CS) [37]–[39]. The dataset contains 29,000 instances in the training set, 1,014 in the development set, and 1,000 in the 2016 test set, where each instance comprises an image and its description in four languages (EN, DE, FR and CS). This setup of Multi30 K makes this dataset a "simpler" version of the real-world multi-lingual multi-modal data as compared to the `How2` dataset which is inherently a video-based human-targeted instructions corpus. Multimodal MT on the `How2` dataset [41] is explored in a follow up work.

The referential grounding models are dependent on image region annotations and their mapping to the text. We consider bounding box localisations of an object as "region," for which we have region annotations derived from **Flickr30 K Entities** [42]. In the dataset, each entity mention (noun phrase) in Flickr30 K descriptions is annotated with a bounding box of the instance(s) depicted. Any entity without a bounding box is labeled as non-visual. Each entity mention is also assigned at least one of eight high-level categories (*person*, *clothing*, *bodyparts*, *animals*, *vehicles*, *instruments*, *scene* and *others*).

B. Model

1) Implicit Grounding: We propose two new attention mechanisms for MMT, where grounding happens on the source language and where the process may be supervised by examples of aligned word-image region pairs.

a) Base Model: As a baseline, we experiment with the standard visual attention approach by Caglayan et al. [29] and its extension to hierarchical fusion by Libovický and Helcl [9]. The image features for an image I are extracted from the last convolutional layer of a 152-layer `ResNet` [23] as a $14\times14\times1024$ feature map.

b) Source Co-Attention: Our first proposed object-level attention model learns to align source words to object regions and to translate them jointly.

Let $\mathbf{V} = v_1, \ldots, v_m$ be the m oracle or detected object-level regions that have been cropped from the image. The visual representation for each object region, $\phi(v_i)$, is a 2,048-dimensional vector generated as a non-linear transform of the penultimate (`pool5`) layer of a 152-layer `ResNet` CNN.

Given these representations, we adapt the co-attention mechanism of Lu et al. [43] to ground the source words where the model jointly learns to align these words to the image regions, and to translate them. This is done by first obtaining the affinity

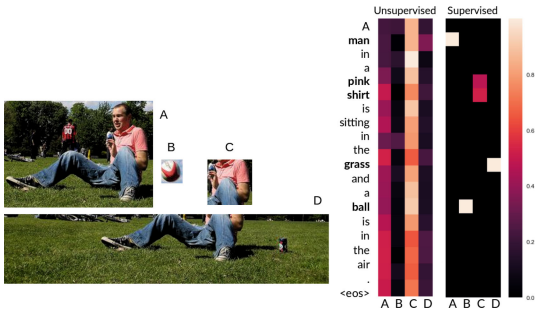


Fig. 5. Distribution of attention weights for unsupervised and supervised co-attention mechanism.

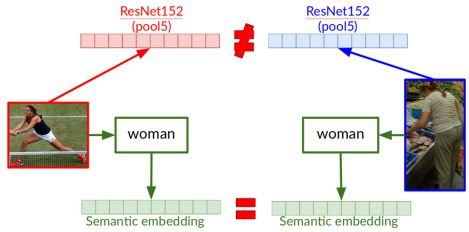


Fig. 6. Specification via category embedding or visual features.

matrix $\mathbf{A}$:

$$\mathbf{A} = \tanh\left(\mathbf{H}^{\top} W_a \mathbf{V}\right) \tag{6}$$

where $\mathbf{H} \in \mathcal{R}^{n \times d}$ are the encoder hidden states and $\mathbf{V} \in \mathcal{R}^{m \times l}$ are the object-level image representations and W_a is the bilinear parameter matrix. The image and encoder attention maps are obtained as:

$$\begin{aligned} \mathbf{C_s} &= \tanh\left(\mathbf{W_{cs}}\mathbf{H} + (\mathbf{W_{cv}}\mathbf{V})\mathbf{A}^{\top}\right) \\ \mathbf{a^s} &= \text{softmax}(w_{cs}^{\top}\mathbf{C_s}) \end{aligned} \tag{7}$$

where $\mathbf{a^s}$ computes the source affinity. Similarly, visual affinity $\mathbf{a^v}$ is computed as:

$$\begin{aligned} \mathbf{C_v} &= \tanh\left(\mathbf{W_{cv}}\mathbf{V} + (\mathbf{W_{cs}}\mathbf{H})\mathbf{A}\right) \\ \mathbf{a^v} &= \text{softmax}(w_{cv}^{\top}\mathbf{C_v}) \end{aligned} \tag{8}$$

Hierarchical attention [9] is added on top of co-attention such that, at decoding time, the model jointly attends to the source context vector computed using the standard attention and the sum of the source affinity attention and the visual affinity attention from Eq 7 and Eq 8.

c) Supervised Source Co-Attention: Our second proposed model learns to ground source words to bounding box regions with explicit correspondence annotations as supervision. We expand the co-attention approach by adding an auxiliary loss to the standard cross-entropy loss. The auxiliary loss penalises cases where the co-attention weights are highest for regions other than the correct one. Inspired by phrase localisation work by Rohrbach et al. [44], given a correct region j we define the grounding loss as:

$$\mathcal{L}_{grounding} = -\frac{1}{B}\sum_{b=1}^{B} \log(\Pr(j|\mathbf{a^v})) \tag{9}$$

where B is the number of phrases per batch and $\mathbf{a^v}$ is from Eq 8. The loss is only active if the ground truth has an alignment; otherwise, it is set to zero.

In Fig. 5 we show an example of attention weights learned for image regions (indicated by letters A-D on the grids) for a source sentence with both the unsupervised and supervised versions of the source co-attention mechanism. The supervised version clearly learns to assign the attention weights to the correct regions for each given content source word.

2) Explicit Grounding: While attention is a well-established approach, for a dataset as small as ours (30 K training instances), the models do not observe enough instances of similar visual representations with the same textual context for attention to be effective. The exception is supervised attention, as shown in the previous section, but it requires region annotations and their alignments to source words for training.

Here we introduce a different approach: regions and their correspondences (alignments) to words in the source sentence are identified beforehand, and then fed to the model as a way of further specifying the source words.

Previous work has explored word-level information in neural MT as morphological features [45] and as topics [46]. In both cases, every word was specified with a vector containing the additional information (e.g. POS tags). We follow a similar approach; however, our setting is more complex in that we do not have an image region associated to every word in the sentence. We experiment with different strategies for words that do not align to a region in the image, including function words, as we discuss below. As for the content of the external vector, we experiment with two types of additional information: (i) object categories, and (ii) CCA projections.

a) Object Categories: The idea is to specify a word with the category of the object in the image it aligns to. We focus on nouns, which are more commonly depicted in images. Instead of using `pool5` features, for visual representations we rely on the category of the objects in the image for which an alignment exists. Fig. 6 shows a motivational example, where the `pool5` visual representation for the two *woman* regions would be very different despite belonging to the same semantic category. To make the representation more flexible, instead of the category label itself, we use pre-trained word embeddings for the word representing the category. By doing so, visual representations for *woman* and *girl* would be closer than those for *woman* and *dog*, for example. We refer to this representation as $\mathbf{E_{obj}}$.

b) CCA Projections: Since the specification involves relating words to image representations, we evaluate the utility of projecting the image representation such that it is highly correlated with the word representations by using canonical correlation analysis (CCA) [47]. Formally, given paired matrices $\mathbf{V}$ and $\mathbf{E}$, where each row of $\mathbf{V}$ is a visual region and its corresponding word represented by its embedding $\mathbf{E}$, we generate a linear projection using CCA. We then use these projections to obtain transformed representations of $\mathbf{V}$ as $\mathbf{V_{cca}}$ and use them as visual features. $\mathbf{V}$ can contain either category embeddings or `pool5` representations.

TABLE IV
COMPARISON OF MODELS USING ORACLE OBJECT ANNOTATIONS AND ALIGNMENTS, ACCORDING TO METEOR. RESULTS ARE AVERAGE OF THREE RUNS WITH DIFFERENT SEEDS. THE FIRST ROW INDICATES THE BEST SYSTEM FOR EN-DE, THE ONLY PAIR TESTED ON THIS TEST SET AT WMT16 [37]

Systems	EN-CS	EN-DE	EN-FR
Best WMT16	–	53.20	–
Text-only	28.90	**57.35**	74.09
SubrAttention	28.84	55.45	73.31
CoAttention	30.37	57.15	**75.85**
SupCoAttention	30.34	56.48	75.10
ExplicitProj	**30.63**	57.05	75.02
ExplicitConc	**30.61**	57.26	75.17
ExplicitCCA	30.52	57.12	75.34

For both object categories and CCA projections, for unaligned words we specify them with an empty vector or with the vector containing pre-trained word embeddings of the word itself. We experiment with specifying every single word in the phrase for multi-word alignments, or specifying the head nouns only. We explore two methods to specify visual information for words: *concatenation* and *projection*.

3) Concatenation: The source word embedding is specified with region-grounded information via concatenation:

$$\tilde{\phi}(s_i) = [\phi(s_i); \phi(r)] \tag{10}$$

where, $\phi(s_i)$ is the source word embedding and $\phi(r)$ is the object-level region information (category label embedding or CCA projection). These are the initial representations of the words for the encoder bidirectional recurrent units.

4) Projection: Alternatively, we learn a linear projection W over the region-grounded information:

$$\tilde{\phi}(s_i) = \phi(s_i) + W\phi(r) \tag{11}$$

C. Experimental Results

We build attention-based sequence-to-sequence models [32] with bidirectional recurrent neural networks with gated recurrent units [48] as the encoder and decoder. We use the `nmtpytorch` tool [31] with the following settings: early stop by Meteor (max 100 epochs), selection of best model according to Meteor, beam size = 6, batch size = 64, Adam as optimizer, word embedding dimensionality = 256, and no sub-word units (they do not improve performance in our case).

For category embeddings and CCA representations we use `fasttext` 300-dimensional pre-trained word embeddings [49]. In the results reported for explicit alignments we specify only head nouns for which an alignment exists to a region in the image, and use the pre-trained embeddings of the words themselves for the remaining words.

1) MMT Results: Table IV summarises the results for the following models, using BLEU [50] and Meteor [51], where the latter is the official metric used for this task (following the MMT shared tasks):

1) **Text-only**: NMT baseline without visual information.
2) **SubrAttention**: Visual attention over image subregions at decoding time (Section IV-B1a) with hierarchical fusion.
3) **CoAttention**: Co-attention over image regions (`pool5` features for objects) and source words (Section IV-B1b).
4) **SupCoAttention**: Supervised co-attention over (`pool5` image region features for objects) and source words (Section IV-B1c).
5) **ExplicitProj**: Projection of category embedding information $\mathbf{E_{obj}}$ (Section IV-B2b).
6) **ExplicitConc**: Concatenation of category embedding $\mathbf{E_{obj}}$ and learned word embeddings (Section IV-B3).
7) **ExplicitCCA**: Concatenation of $\mathbf{V_{cca}}$ (`pool5` object features) and learned word embeddings (Section IV-B3).

TABLE V
COMPARISON OF MODELS USING ORACLE OBJECT ANNOTATIONS AND ALIGNMENTS, ACCORDING TO LTA

Model	EN-CS	EN-DE	EN-FR
Text-only	10.44	37.00	53.62
SubrAttention	10.84	37.82	53.62
CoAttention	12.45	38.06	**55.16**
SupCoAttention	13.25	37.47	**55.16**
ExplicitProj	13.65	**38.41**	54.08
ExplicitConc	12.85	38.06	53.78
ExplicitCCA	**14.06**	38.17	54.08

The results in Table IV show that the proposed multimodal models outperform text-only counterparts as well as the standard multimodal approach SubrAttention for EN-CS and EN-FR. As it has been shown in the WMT shared tasks on MMT [38], [39], automatic metrics often fail to capture nuances in translation quality such as the ones we expect the visual modality to help with, which – according to human perception – lead to better translations. This may be particularly the case for EN-DE, where rich morphology and compounding may result in better translations, even though these do not match the reference sentences.

2) Lexical Ambiguity Evaluation: To deal with the weaknesses of the automatic metrics above, we also evaluate systems using Lexical Translation Accuracy (LTA) [52] following the methodology used at the WMT18 shared task on MMT [39]. LTA measures how accurately a system translates a subset of ambiguous words found in the Multi30 K corpus. A word is said to be ambiguous in the source language if it has multiple translations (as given in the Multi30 K training corpus) with different meanings. A lexical translation is considered correct if it matches exactly the (lemmatised) word aligned to it on the reference test set. The test set of 1,000 sentences contains 1,708 such words for EN-DE, 1,298 for EN-FR, and 249 for EN-CS. Table V shows that all multimodal models are better than their text-only counterpart.

3) Oracle Versus Predicted Regions: Thus far we showed results where the oracle bounding boxes and object-word alignments are used. In the implicit grounding models this is not a major issue given that the alignments are only needed at training time. For the explicit grounding models, however, this information is also needed at test-time. Therefore, we also investigate using predicted objects and object-word alignments [53].[2]

[2]We use the **w2v-max** and **union** model described in their paper.

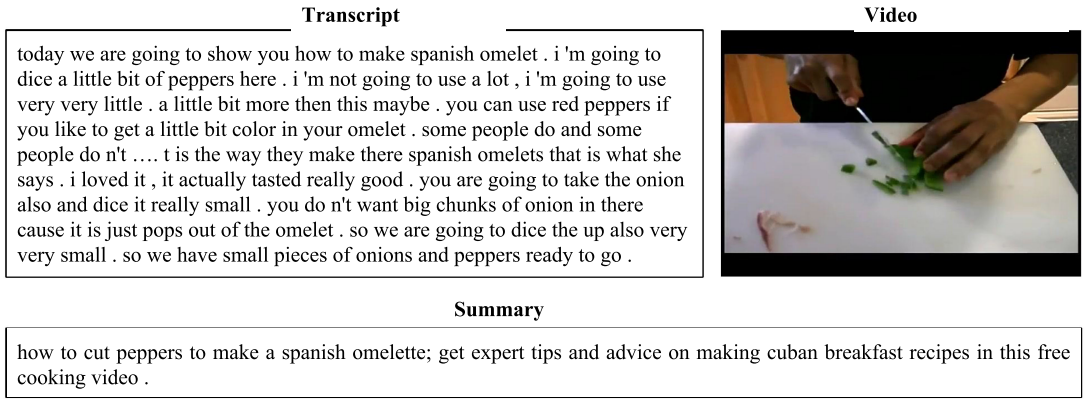


Fig. 7. How2 dataset example with different modalities. "Cuban breakfast" and "free cooking video" are not mentioned in the transcript and must be derived from other sources.

TABLE VI
ROUGE-L AND CONTENT F1 FOR DIFFERENT SUMMARIZATION MODELS: RANDOM BASELINE (1), RULE-BASED EXTRACTED SUMMARY (2A), NEAREST NEIGHBOR SUMMARY (2B), DIFFERENT TEXT-ONLY (3,4,5A), POINTER-GENERATOR (5B), ASR OUTPUT TRANSCRIPT (5C), VIDEO-ONLY (6-7) AND TEXT-AND-VIDEO MODELS (8-9)

Method		ROUGE-L	Content F1
Naive baselines			
1	Language Model sampling	27.5	8.3
2a	Rule-based Extractive summary	16.4	18.8
2b	Next-neighbor Summary	31.8	17.9
Text-only models			
3	S2S on 2a only	46.4	36.0
4	S2S on 200 tokens of Transcript	40.3	27.5
5a	S2S on Transcript	**53.9**	**47.4**
5b	PG on Transcript	50.2	42.0
5c	S2S on ASR	46.1	34.7
Video-only models			
6	AF only	38.5	24.8
7	RNN over AF	**46.3**	**34.9**
Multimodal models			
8	Transcript + AF w/ Hier. Attn	**54.9**	**48.9**
9	ASR + AF w/ Hier. Attn.	46.3	34.7

The results indicate that there are no significant differences in performance.

D. Discussion

We proposed referential grounding approaches for MMT that use clearly defined correspondences between a source word and an object in the image to guide translation. We showed that MMT models using such groundings at object-level can better exploit image information, leading to better performance, especially when translating challenging cases such as ambiguous words.

V. SUMMARIZATION

All videos in the How2 dataset are accompanied by a manually written summary that should attract the attention of viewers and increase the chance of the video being found in a keyword search. The goal of the summarization task on this dataset is to generate this type of video summary. An example video summary is shown in Fig. 7.

A. Characteristics of the Summaries

In order to get a reliable estimate of the summarization quality, we use a different split than for ASR and MT. The standard splits contain enough text for sentence-level evaluation; however, there is only one summary per video. We use 73,993 videos for training, 2,965 for validation and 2,156 for testing. The average length of transcripts is 291 words and and the average length of summaries is 33 words.

B. Baseline Methods

a) Language Input: For text-based input, we use the transcripts of the videos. We leverage the speech modality by using the outputs from a pre-trained speech recognizer trained with other data, as inputs to a text summarization model. We use state-of-the-art models for distant-microphone conversational speech recognition, ASpIRE [54] and EESEN [55], [56]. The word error rate of these models on the How2 test data is 35.4%. This high error mostly stems from normalization issues in the data. For example, recognizing and labeling "20" as "twenty" etc. We accept these as-is for this task. Also, note that this is the WER on the larger 2000-hour corpus rather than 300-hour subcorpus.

b) Visual Input: We represent videos by a sequence of 2048-dimensional action feature vectors (see Section II).

C. Models

We study various summarization models. First, we use a Sequence-to-Sequence (S2S) model [57] consisting of an encoder RNN to encode (text or video features) with the attention mechanism [32] and a decoder RNN to generate summaries. Our second model is a Pointer-Generator (PG) model [58], [59] that has shown strong performance for abstractive summarization [60], [61]. As our third model, we use hierarchical attention approach [9] originally proposed for multimodal machine translation to combine textual and visual modalities to generate text. This model first computes the context vector independently for each of the input modalities (text and video). In the next step, the context vectors are treated as states of another encoder, and a new

TABLE VII
EXAMPLE OUTPUTS OF GROUND-TRUTH TEXT-AND-VIDEO WITH HIERARCHICAL ATTENTION (8), TEXT-ONLY WITH GROUND-TRUTH (5A), ACTION FEATURES WITH RNN (7) AND THE TOPIC-BASED NEXT NEIGHBOR (2B)

No.	Model	R-L	C-F1	Output
-	Reference	-	-	watch and learn how to tie thread to a hook to help with fly tying as explained by out expert in this free how - to video on fly tying tips and techniques .
8	Ground-truth text + Action Feat.	54.9	48.9	learn from our expert how to attach thread to fly fishing for fly fishing in this free how - to video on fly tying tips and techniques .
5a	Text-only (Ground-truth)	53.9	47.4	learn from our expert how to tie a thread for fly fishing in this free how - to video on fly tying tips and techniques .
7	Action Features + RNN	46.3	34.9	learn about the equipment needed for fly tying , as well as other fly fishing tips from our expert in this free how - to video on fly tying tips and techniques .
2b	Next Neighbor	31.8	17.9	use a sheep shank knot to shorten a long piece of rope . learn how to tie sheep shank knots for shortening rope in this free knot tying video from an eagle scout .

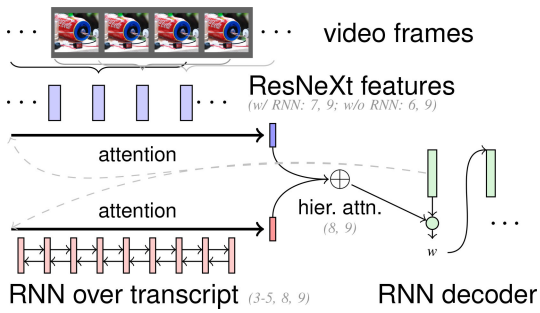


Fig. 8. Building blocks of the sequence-to-sequence models, gray numbers in brackets indicate which components are utilized in which experiments.

vector is computed. When using a sequence of action features instead of a single averaged vector, the RNN layer helps capture context. In Fig. 8, we present the building block of our models.

D. Evaluation

To evaluate the generated summaries we use ROUGE-L [62], a standard metric for abstractive summarization that measures the longest common sequence between the reference and the generated summary. Additionally, we introduce a new metric *Content F1* that fits the template-like structure of the summaries observed in our dataset.

a) Content F1: We compute the F1 score of the content words in the summaries based over a monolingual alignment obtained using METEOR toolkit [63]. Then, we remove function words and task-specific stop words that appear in most of the summaries from the reference and the hypothesis. These stop words (how, learn, tips, free, etc.) are very frequent in the reference summaries making it easy for the decoder to predict these and thus increase the ROUGE score. We treat the remaining words from the reference and the hypothesis as two bags of words and compute the F1 score over the alignment. Note that the score ignores the fluency of output in line with recently proposed metrics such as HighRes [64].

b) Human Evaluation: In addition to automatic evaluation, we also evaluated system summaries by eliciting human judgments. Following the abstractive summarization human annotation work of Grusky *et al.* [65], we ask our annotators to label the generated output on a scale of 1–5 on metrics of *informativeness, relevance, coherence*, and *fluency*. We perform this on randomly sampled 500 videos from the test set. We evaluate three models: two unimodal (text-only, 5a; video-only 7) and one multimodal (text-and-video, 8). Three workers annotated each video on Amazon Mechanical Turk.

TABLE VIII
HUMAN EVALUATION SCORES ON 4 DIFFERENT MEASURES OF INFORMATIVENESS (INF), RELEVANCE (REL), COHERENCE (COH), FLUENCY (FLU)

Model (No.)	INF	REL	COH	FLU
Text-only (5a)	3.86	**3.78**	3.78	3.92
Video-only (7)	3.58	3.30	3.71	3.80
Text-and-Video (8)	**3.89**	3.74	**3.85**	**3.94**

E. Output Examples from Different Models

Table VII shows the example outputs from our different text-only and text-and-video models. The text-only model produces a fluent output which is close to the reference. The action features with the RNN model, which sees no text in the input, produces an in-domain ("fly tying"' and "fishing") abstractive summary that involves more details like "equipment" which is missing from the text-based models but is relevant. The next neighbor model is related to "knot tying" but not related to "fishing". The scores for each of these models reflect their respective properties. Observing other outputs of the model, we noticed that although predictions were usually fluent and thus getting high ROUGE scores, there is a large room for improvement by predicting all details from the ground truth summary, like the subtle selling point phrases, or by using the visual features in a different adaptation model.

In Table VIII, we report human evaluation scores of the best text-only, video-only, and multimodal models. We observe that text-only summaries dominate on relevance but multimodal models are the most informative, coherent and fluent, indicating that these models can fuse complementary information from

multiple modalities to generate relevant summaries. The example presented in Table VII shows how the generated summaries vary with different models and features.

Our parallel work [66], [67] demonstrates the use of our summarization models trained in this work for a transfer learning-based summarization task on the Charades dataset [68], which has audio, video, and text (summary, caption, and question-answer pairs) modalities just like the How2 dataset. Pre-training and transfer learning with the How2 dataset led to significant improvements in unimodal and multimodal adaptation tasks on the Charades dataset.

VI. Correlation-Based Unsupervised Learning

All machine learning involves learning representations on top of the input features [69]. In deep learning, representation is learned implicitly, as a result of finding a local minimum of a loss function. In contrast to this implicit representation learning stand several explicit representation learning paradigms [70]–[72]. How to best exploit multiple views is an open problem, especially when there is a latent alignment between views, such as between an image and its spoken caption [73]. We treat the How2 dataset as a 4-way parallel corpus, and explore an advanced, correlation-based representation learning objective.

A. Deep Generalized Canonical Correlation Analysis

It has been shown that the availability of a second view in addition to a primary input can help with any task. For instance, the video stream of a speaker's face, in addition to the audio recording, helps perform speech recognition [74]. This qualitative result is still true when the second view is reconstructed from the primary input by a trained predictor. However, in some cases, it may be difficult to learn such a predictor, as in the speech recognition example above. Instead of reconstructing the secondary view, it is simpler to learn a representation for each view that is maximally reconstructive of the representations learned for the other views [72], [75]. This intuition was first formalized as Canonical Correlation Analysis (CCA) [47], extended to pairs of views [76] and arbitrary feature extractors [77]. We use the formulation of [78], [79] which we describe next.

For each view $j \in \{1..J\}$, all N points of the dataset are stored in a matrix $X_j \in \mathbb{R}^{d_j \times N}$, where d_j is the dimensionality of the feature vector. We denote $f_j : \mathbb{R}^{d_j} \rightarrow \mathbb{R}^{h_j}$ the j-th learned feature extractor — in our case a neural net — and $U_j \in \mathbb{R}^{h_j \times k}$ a linear transformation matrix. The $\{f_j\}_j$ and $\{U_j\}_j$ are trained jointly to reconstruct an unknown shared representation, under constraints, resulting in the following problem:

minimize $\sum_{j=1}^{J} ||G - U_j^T f_j(X_j)||_2^2$ subject to $GG^T = I_k$, with respect to parameters $\{G, \{f_j, U_j\}_j\}$. Here, $G \in \mathbb{R}^{k \times N}$ can be viewed as the learned representation for the dataset, and k is the dimensionality of said representation. The constraint on G prevents trivial solutions. Note that each learned feature extractor (f_j, U_j) tries to reconstruct G from X_j. We refer to this method as deep generalized CCA (DGCCA). Deep CCA [77] is equivalent to the case $J = 2$.

TABLE IX
Recall@10 for Retrieving Reference Modality Given Source Modality ("Source - Reference"). Swapping Source and Reference Change Retrieval Scores by Less Than 1% Absolute

	Linear CCA		Deep CCA		
	dev	test	dev	test	k
text (en) - text (pt)	82.5	81.4	95.1	94.6	400
speech - text (en)	98.3	96.9	92.1	90.1	160
video - text (en)	0.9	0.8	2.3	1.6	400
video - speech	0.8	0.6	1.9	1.8	160

TABLE X
Recall@10 for Retrieving Column Modality Given Source Row Modality, for a DGCCA Model Trained on 3 Views. Results From the Bottom Left Triangle Can be Compared to Those in Table IX.

		text (en)	speech	video
text (en)	dev	-	92.1	1.7
	test	-	89.8	1.8
speech	dev	92.1	-	1.9
	test	89.1	-	1.2
video	dev	1.4	1.9	-
	test	1.7	1.2	-

B. Experiments and Results

Within the framework of DGCCA, we use the How2 dataset as a 4-way parallel corpus: video, speech, transcription in English, translation in Portuguese. Each data point in that corpus corresponds to one utterance. For the text and speech modalities, we use encoder-decoder sequence-to-sequence models trained on the How2 dataset to extract the $\{X_j\}_j$ features. We average either the encoder-side embeddings or the sequence of context vectors to obtain a single vector for each sequence, following [80]. For the video modality, we first break up the videos into keyframes, then average the outputs of a ResNet [23] over the time window corresponding to a given utterance. We thus obtain a single vector representing the video modality for each utterance.

1) Retrieval Experiments: We start with an intrinsic evaluation of our learned representations. We use a retrieval task to probe the reliability of the learned embedding space. Given a source point v, we return the 10 closest points within a reference set $\{u_i\}_i$. The source and reference points come from different views of the dev and test sets of the How2 dataset . This allows us to score the retrieval based on whether the correct point is within the 10 closest points, and we report this as Recall@10. Picking the 10 closest points at random results in a Recall@10 of 0.5% for the dev set and 0.4% for the test set. Using our DGCCA model, retrieving the 10 closest points involves projecting the source point and the reference set into the shared space, computing pairwise distances (we use mean-centered cosine distance) and taking the 10 closest points.

To validate the approach, we compare linear and deep CCA on pairs of modalities. Linear CCA corresponds to f_j being set to the identity mapping for all j. We report retrieval results in Table IX. With the exception of speech-to-text retrieval, deep CCA performs systematically better than linear CCA.

We train models on 3 and 4 modalities, and report retrieval scores in Tables X and XI. In both cases, $k = 160$. When adding modalities, we note that retrieval scores decrease, since the

TABLE XI
RECALL@10 FOR RETRIEVING COLUMN MODALITY GIVEN SOURCE ROW MODALITY, FOR A DGCCA MODEL TRAINED ON 4 VIEWS. RESULTS FROM THE BOTTOM LEFT TRIANGLE CAN BE COMPARED TO THOSE IN TABLE IX.

		Text (pt)	Text (en)	Speech	Video
Text (pt)	dev	-	98.8	73.5	2.1
	test	-	98.3	71.0	1.1
Text (en)	dev	98.8	-	88.2	1.4
	test	98.4	-	85.4	0.9
Speech	dev	73.0	88.1	-	1.1
	test	70.7	85.4	-	1.0
Video	dev	2.1	1.1	1.0	-
	test	1.1	1.1	0.9	-

TABLE XII
SCORING TOP-1 RETRIEVAL RESULT FROM DGCCA MODELS WITH ASR, MT AND ST METRICS. MODELS USED (FROM LEFT TO RIGHT) WERE TRAINED USING SPEECH AND TEXT (EN); TEXT (EN) AND TEXT (PT); SPEECH, TEXT (EN), TEXT (PT) AND VIDEO. SOURCE SENTENCES FOR THE RETRIEVAL ARE FROM THE TEST SET

Reference Set	WER	BLEU (MT)	BLEU (ST)
train	134%	5.2	0.2
train + test	27.4%	80.7	19.8
Baseline S2S	24.3%	57.3	27.9

model needs to accommodate additional views. Some retrieval scores are higher than others; most likely, the model trades off higher scores for easier pairs of views (e.g. Portuguese text and English text) against lower scores for harder pairs of views (e.g. video and speech). This could be compensated by adding weights $\{w_j\}_j$ for each reconstruction loss, or by tuning the architectures of the $\{f_j\}_j$ separately.

Overall, retrieval scores between language modalities are high, ranging from 71.0% to 98.4%. There are several reasons which could explain the lower scores involving the video modality. First, it is not quite clear how much temporal coherence the video modality has in the How2 dataset . For instance, objects mentioned by the speaker might appear much later in the video, very briefly, or not at all. Further, ResNet features might not be able to adequately represent the domain of the How2 dataset . We experimented with representations from action networks [21] trained on an action dataset [22], and obtained similar results. Most likely, given the noisy input features, our models lack either the expressive power or a sufficient amount of training data to capture the correspondence between the language modalities and the video [73], [81].

2) Scoring top-1 Retrieval Results: Given our high retrieval scores between language modalities, we attempt to measure their performance with conventional ASR, MT and ST metrics — WER and BLEU scores. For each data point in the test set, we retrieve the closest point from a reference set, and use it as the output hypothesis of either an MT, ASR or ST model, which can then be scored with the relevant metric. If using the test set as a reference set, given the high retrieval scores, the WER or BLEU scores would be almost perfect. We thus report two more challenging settings in Table XII: the reference set can be either the train set, or the union of the train set and the test set. As compared to the baseline sequence-to-sequence neural model, our models perform reasonably well, and are consistent with our retrieval scores: MT works best, then ASR, then ST. When the reference set is the train set, the scores drop considerably, also because the train set does not necessarily contain adequate sentences. To quantify this, we pick, for each target sentence from the test set, the closest sentence from the train set in terms of edit distance, which yields a BLEU of 10.6 and a WER of 63.0%.

C. Discussion

We framed the How2 dataset as a multiview representation learning problem, and probed the quality of the learned representations using intrinsic evaluations. While our results show it is possible to learn high-quality representations on the language modalities, the video modality remains a major challenge, possibly calling for specialized architectures or transfer learning. Further integrating the learned representations into supervised tasks is left for future work.

VII. CONCLUSION

This paper describes (1) the How2 dataset , a collection of large-scale open-domain user-generated instructional videos, and (2) a detailed study of different multi-modal learning experiments on this dataset or other proxy datasets like Multi30 K for MT. This corpus brings together English audio, English transcripts, Portuguese transcripts, videos, and summaries, along with meta-data such as topic of the video. This makes the How2 dataset a good resource for research at the intersection of vision, language and speech. By releasing this dataset, we hope to enable research on multi-lingual, multi-modal, highly correlated and well-aligned parallel modalities. We presented numerous uni-, multi- and cross-modal tasks such as speech recognition, machine translation, summarization, and multi-view representation learning. With this study, we hope to shed light on the current state of vision, language and speech grounding and to help researchers with designing new tasks in this space.

APPENDIX

A. The How2 Dataset

Fig. 9 show the LDA topic distribution and segment length analysis of the *300 h* subset of the How2 dataset .

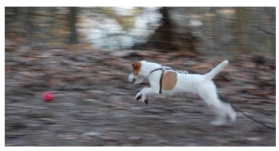
B. Region-Specific Multimodal Machine Translation

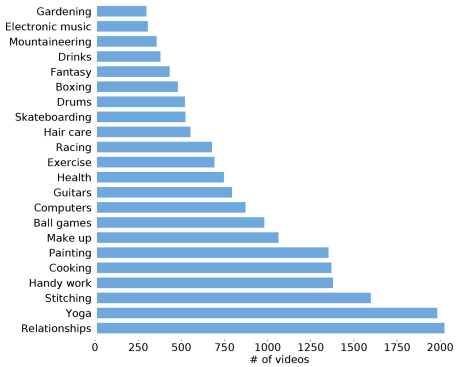
Table XIII shows qualitative examples for results presented in Section IV.

C. Correlation-based Multiview Learning

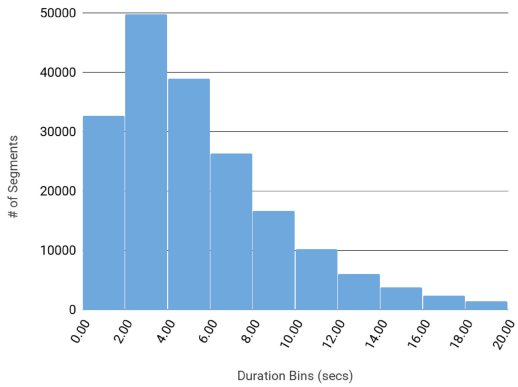
1) Feature Extraction: We use baseline ASR and MT models from Sections III and IV. For each input sequence, the encoder produces a corresponding sequence of feature vectors $h_1, \ldots, h_T$. We use $\frac{1}{T}\sum_{i=1}^{T} h_i$ to represent that input sequence. The decoder with attention produces a sequence of context vectors $c_1, \ldots, c_S$, and we use $\frac{1}{S}\sum_{i=1}^{S} c_i$ to represent the target sequence. Since we use word-based ASR and MT systems, each c_j and h_i roughly represents a word in context. For the video

TABLE XIII
QUALITATIVE EXAMPLES COMPARING TEXT-ONLY NMT AND MULTIMODAL MODELS. WE SHOW THE SOURCE (SRC), TEXT-ONLY MT (NMT) AND A MULTIMODAL MODEL (MMT). IN BOTH CASES WE ALSO SHOW THE BACK TRANSLATION INTO ENGLISH FOR CLARITY. UNDERLINED WORDS REPRESENT TRANSLATION ERRORS, WHILE BOLD FACE WORDS, THE CORRECT (OR BETTER) VERSION

EN-FR	
	SRC: A man on a tag line going into the water. NMT: Un homme sur une ligne de métro en train de marcher dans l'eau. *(A man on the metro line walking to the water.)* MMT: Un homme sur une **ligne de sable allant** dans l'eau. *(A man on the **sand line going** into the water.)*
	SRC: A large group of people of various ages and genders sit outside together. NMT: Un grand nombre de personnes de différents âges et des accessoires sont assis ensemble. *(A large number of people of different ages and accessories sit together.)* MMT: Un grand nombre de personnes de différentes áges et **d'autres** sont assis ensemble . *(A large number of people of different ages and **others** sit together.)*
EN-DE	
	SRC: A man in a gray shirt jumps over the top of a sand dune in the desert . NMT: Ein mann in einem grauen hemd springt über das dach einer sanddüne . *(A man in a grey shirt is jumping over the roof of a sand dune.)* MMT: Ein mann in einem grauen hemd springt über **die spitze** einer sanddüne in der wüste. *(A man in a grey shirt is jumping over **the peak** of a sand dune in the desert.)*
	SRC: A fox terrier leaps after a ball. NMT: Ein metzger springt nach einem ball. *(A butcher jumps for a ball.)* MMT: Ein **terrier** springt nach einem ball. *(A **terrier** jumps for a ball.)*



(a) Topic distribution.



(b) Segment duration.

Fig. 9. LDA topic distributions and segment durations for the 300 h subset. The overall 2000 h corpus exhibits very similar characteristics.

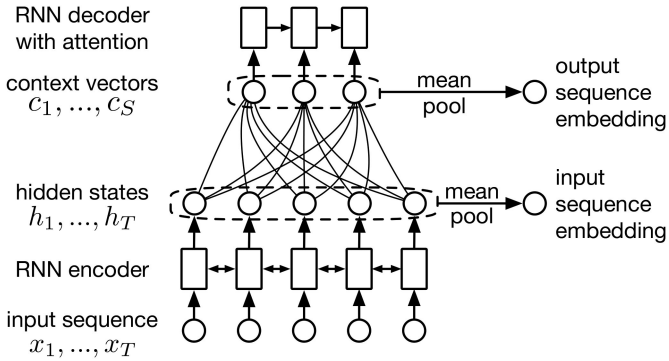


Fig. 10. Extracting sequence embeddings from trained sequence to sequence models.

modality, we first break up the videos into keyframes, then use a ResNet [23] to map each keyframe to a multi-class posterior, based on the 1000 ImageNet classes. For each speech utterance, we then compute the average of the posteriors corresponding to the time window of the speech utterance. The averaging process is meant to capture the most persistent predictions and reduce the variability due to noise. We thus obtain a single vector representing the video modality for each utterance. As a result, for text, speech and video, the X_j features are 320-, 800- and 1000-dimensional, respectively.

2) Models and Training: The features described above are kept fixed, while we use feed-forward neural networks with 2 hidden layers and tanh non-linearities for the $\{f_j\}_j$. The first

layer has the same dimensionality as the input, and the second layer the same size as k. To avoid under-defining the objective, k should be no larger than the smallest of the $\{h_j\}_j$. We set k to half the smallest of the $\{h_j\}_j$ involved, as a heuristic to retain most of the informative components and discard uninformative ones. For numerical stability, we add the identity matrix scaled by 10^{-16} to all the view-specific covariance matrices. We use stochastic gradient descent with batch size 5500 and Adam optimizer with default parameters. The analytical expression of the gradient was taken from [78]. In the experiments involving video, we use a weight decay of 10^{-5}. After each full pass over the training set, we measure retrieval scores between all possible pairs of different views on the dev set, using the highest of these scores to measure the performance of our model. We use this score to do early stopping.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.

[2] T. Wilcox, R. Woods, C. Chapa, and S. McCurry, "Multisensory exploration and object individuation in infancy." *Developmental Psychol.*, vol. 43, no. 2, p. 479, 2007.

[3] L. W. Barsalou, W. K. Simmons, A. K. Barbey, and C. D. Wilson, "Grounding conceptual knowledge in modality-specific systems," *Trends Cogn. Sci.*, vol. 7, no. 2, pp. 84–91, Mar. 2003.

[4] R. Bernardi *et al.*, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, 2016.

[5] S. Antol *et al.*, "VQA: Visual question answering," *in Proc. Int. Conf. Comput. Vision*. IEEE, 2015, pp. 2425-2433.

[6] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, 2013, pp. 853-899.

[7] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *IEEE Int. Conf. Comput. Vision 2015*, pp. 4534-4542.

[8] O. Caglayan, L. Barrault, and F. Bougares, "Multimodal attention for neural machine translation," 2016, *arXiv:1609.03976*.

[9] J. Libovický and J. Helcl, "Attention strategies for multi-source sequence-to-sequence learning," in *Proc. 55th Annual Meeting of the Assoc. Comput. Linguistics (Vol. 2: Short Papers)*, 2017, pp. 196–202.

[10] D. Elliott and Á. Kádár, "Imagination improves multimodal translation," in *Proc. Eighth Int. Joint Conf. Natural Lang. Process.*, Taipei, Taiwan, 2017, pp. 130–141.

[11] S. Gella, R. Sennrich, F. Keller, and M. Lapata, "Image pivoting for learning multilingual multimodal representations," in *Conf. Empirical Methods Natural Lang. Process*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2839–2845.

[12] Á. Kádár, D. Elliott, M.-A. Côté, G. Chrupała, and A. Alishahi, "Lessons learned in multilingual grounded language learning," in *Proc. 22nd Conf. Comput. Natural Lang. Learn*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 402–412.

[13] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach," in *Interspeech 2016*, 2016, pp. 3414–3418.

[14] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2017, pp. 5020–5024.

[15] S. Palaskar, R. Sanabria, and F. Metze, "End-to-end multimodal speech recognition," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech Signal Process., 2018*, 2018, pp. 5774–5778.

[16] O. Caglayan, R. Sanabria, S. Palaskar, L. Barraul, and F. Metze, "Multimodal grounding for sequence-to-sequence speech recognition," *in Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech Signal Process*. IEEE, 2019, pp. 8648–8652.

[17] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual English-German image descriptions," in *Proc. Workshop Vision Lang*. ACL, 2016.

[18] "How2," [Online]. Available: https://github.com/srvk/how2, 2018.

[19] E. Loper and S. Bird, "NLTK: The natural language toolkit," *Natural Lang. Process. Python. OReilly Media Inc*, 2002.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[21] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and ImageNet?" in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6546–6555.

[22] W. Kay *et al.*, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[25] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 99 pp. 1-1, 2017.

[26] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *Proc. Conf. Comput Vision Pattern Recognit.*, 2016, pp. 4575–4583.

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.

[28] I. Calixto and Q. Liu, "Incorporating global visual features into attention-based neural machine translation." in *Proc. Conf. Empirical Methods Natural Lang. Process*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 992–1003.

[29] O. Caglayan *et al.*, "LIUM-CVC submissions for WMT17 multimodal translation task," in *Proc. Mach. Transl*. ACL, 2017, pp. 432–439.

[30] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2018, pp. 66–71.

[31] O. Caglayan, M. García-Martínez, A. Bardet, W. Aransa, F. Bougares, and L. Barrault, "NMTPY: A flexible toolkit for advanced neural machine translation systems," *Prague Bull. Math. Linguistics*, vol. 109, pp. 15–28, 2017.

[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend, and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. Int. Conf. Acoust., Speech Signal Process*. IEEE, 2016, pp. 4960-4964.

[35] R. Sennrich *et al.*, "Nematus: A toolkit for neural machine translation," in *Proc. Software Demonstrations 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*. Assoc. Comput. Linguistics, 2017, pp. 65–68.

[36] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vision*. Springer, 2016.

[37] L. Specia, S. Frank, K. Sima'an, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *Proc. First Conf. Mach. Transl.: Vol. 2, Shared Task Papers. Assoc. Comput. Linguistics*, 2016, pp. 543–553.

[38] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, "Findings of the second shared task on multimodal machine translation and multilingual image description," in *Proc. 2nd Conf. Mach. Transl*. ACL, 2018, pp. 215–233.

[39] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, and S. Frank, "Findings of the third shared task on multimodal machine translation," in *Proc. Mach. Transl.: Shared Task Papers*. Assoc. Comput. Linguistics, 2018, pp. 304–323.

[40] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014.

[41] V. Raunak, S. K. Choe, Q. Lu, Y. Xu, and F. Metze, "On leveraging the visual modality for neural machine translation," 2019, *arXiv:1910.02754*.

[42] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vision*. IEEE, Dec. 2015, pp. 2641–2649.

[43] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 289–297.
[44] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *Proc. Eur. Conf. Comput. Vision* Springer Int. Publishing, 2016, pp. 817–834.
[45] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *Proc. Mach. Transl.: Vol. 1, Res. Papers. Assoc. Comput. Linguistics*, 2016, pp. 83–91.
[46] S. Deena, R. W. Ng, P. Madhyashtha, L. Specia, and T. Hain, "Exploring the use of acoustic embeddings in neural machine translation," in *Proc. IEEE Automatic Speech Recognit. Understanding Workshop*. Okinawa, Japan: IEEE, 2017, pp. 450–457.
[47] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.
[48] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Transl*. Doha, Qatar: Assoc. Comput. Linguistics, Oct. 2014, pp. 103–111.
[49] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.
[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, 2002.
[51] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proc. Workshop on Statist. Mach. Transl*. ACL, 2007.
[52] C. Lala and L. Specia, " Multimodal lexical translation," in *Proc. Lang. Resour. Eval. Conf.*, N. C. C. chair, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Eur. Lang. Resour. Assoc., May 2018, pp. 3810–3817.
[53] J. Wang and L. Specia, "Phrase localization without paired training examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*. Seoul, South Korea: IEEE, Oct. 2019, pp. 4662–4671.
[54] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with TDNNs, ivector adaptation and RNN-lms," in *Proc. Autom. Speech Recognit. Understanding, 2015 IEEE Workshop on*. IEEE, 2015, pp. 539–546.
[55] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. Autom. Speech Recognit. Understanding, 2015 IEEE Workshop on*. IEEE, 2015, pp. 167–174.
[56] A. Le Franc, E. Riebling, J. Karadayi, W. Yun, C. Scaff, F. Metze, and A. Cristia, "The ACLEW DiViMe: An easy-to-use diarization tool," *in Proc. Interspeech*, Interspeech. ISCA, 2018, pp. 1383–1387.
[57] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst. 27*. Curran Assoc., Inc., 2014, pp. 3104–3112.
[58] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Adv. Neural Inf. Process. Syst*. Curran Assoc., Inc., 2015, pp. 2692–2700.
[59] Ç. Gülçehre, S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio, "Pointing the unknown words," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, ACL 2016, Volume 1: Long Papers*, 2016.
[60] R. Nallapati, B. Zhou, C. dos Santos, Ç. glar Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," *CoNLL 2016*, p. 280, 2016.
[61] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1073–1083.
[62] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proc. 42nd Meeting Assoc. Comput. Linguistics*. ACL, 2004, pp. 605–612.
[63] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th workshop on Statistical Mach. Trans.*, 2014, pp. 376–380.
[64] Hardy, S. Narayan, and A. Vlachos, "Highres: Highlight-based reference-less evaluation of summarization," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 3381–3392.
[65] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," in *Proc. 2018 NAACL:HLT*, 2018, pp. 708–719.
[66] R. Sanabria, S. Palaskar, and F. Metze, "CMU Sinbads submission for the DSTC7 AVSD challenge," in *Proc. 7th Dialog Syst. Technol. Challenges Workshop AAAI*, Honolulu, Hawaii, USA, Jan. 2019.
[67] S. Palaskar, J. Libovickỳ, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," 2019, *arXiv:1906.07901*.
[68] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2016, pp. 510–526.
[69] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
[71] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Sci.*, vol. 313, no. 5786, pp. 504–507, 2006.
[72] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.
[73] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 506–517.
[74] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th ICML*, 2011, pp. 689–696.
[75] R. Arora and K. Livescu, "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7135–7139.
[76] P. Horst, "Generalized canonical correlations and their applications to experimental data," *J. Clin. Psychol.*, vol. 17, no. 4, pp. 331–347, 1961.
[77] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, 2013, pp. 1247–1255.
[78] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," in *Proc. 4th Workshop Representation Learn. NLP, RepL4NLP@ACL 2019.*, 2019, pp. 1–6.
[79] P. Rastogi, B. Van Durme, and R. Arora, "Multiview LSA: Representation learning via generalized CCA," in *Proc. 2015 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2015, pp. 556–566.
[80] N. Holzenberger, S. Palaskar, P. Madhyastha, F. Metze, and R. Arora, "Learning from multiview correlations in open-domain videos," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech Signal Process*. IEEE, 2019, pp. 8628–8632.
[81] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," 2019, *arXiv:1906.03327*.

Lucia Specia is currently a Professor of Natural Language Processing at Imperial College London and University of Sheffield. Her current research focuses on various aspects of data-driven approaches to language processing, with a particular interest in multimodal and multilingual context models and work at the intersection of language and vision. Her work can be applied to various tasks such as machine translation, image captioning, quality estimation and text adaptation.

Loic Barrault is a Senior Lecturer in the Natural Language Processing (NLP) group at the University of Sheffield. His current research work focuses on various aspect of statistical and neural machine translation (e.g. domain adaptation, multimodality, multilinguality).

Ozan Caglayan received the Ph.D. from Le Mans University in 2019 where he had worked with Loïc Barrault on Multimodal Machine Translation. His current research interests include multimodal machine learning and language understanding. He is also the developer of the sequence-to-sequence toolkit.

Amanda Duarte is a graduate student at the Barcelona Supercomputing Center, advised by Prof. Xavier Giro and Prof. Jordi Torres. Her current research interests lie at the intersection of speech, vision and language modalities applied to accessibility.

Desmond Elliott is an assistant professor in the Laboratory for Multimodal Processing at the University of Copenhagen. His research interests include multimodal and multilingual machine learning.

Spandana Gella is a PhD candidate at the School of Informatics, University of Edinburgh advised by Prof. Mirella Lapata and Prof. Frank Keller. Her research interests include weakly supervised action recognition, multilingual multi-modal representation learning, image and video description generation.

Nils Holzenberger is a graduate student at Johns Hopkins University, affiliated with the Center for Language and Speech Processing, and advised by Prof. Ben Van Durme and Prof. Raman Arora. His research interests include representation learning and unsupervised learning for speech and text.

Chiraag Lala biography is not available at the time of publication.

Sun Jae Lee is an MS candidate in Computer Science at the University of Pennsylvania. Her research interests include machine translation and multimodal machine learning.

Jindřich Libovický is a research fellow at Center for Information and Language Processing, Ludwig-Maximilian University of Munich. He recently finished his Ph.D. at Charles University in Prague advised by Prof. Pavel Pecina. His main research focus is neural machine translation including multimodal translation.

Pranava Madhyastha biography is not available at the time of publication.

Florian Metze (Senior Member, IEEE) is an Associate Research Professor at the Carnegie Mellon University Language Technologies Institute and a senior member of the IEEE. His research interests revolve around automatic speech recognition with end-to-end methods, low resource speech recognition, and multi-media analysis.

Karl Mulligan is a graduate student in Cognitive Science at Johns Hopkins University, advised by Prof. Tal Linzen. His current research uses methods from formal linguistics and machine learning to study how semantic representations emerge from language data. Karl is also interested in the representation of information across modalities, with a particular interest in spatial language.

Alissa Ostapenko is an undergraduate student studying Computer Science and Mathematics at Worcester Polytechnic Institute (WPI) in Worcester, MA, USA. Her research experience lies primarily in the field of text mining, including text classification, machine translation, and sentiment analysis. Advised by Dr. Rodica Neamtu (WPI), Alissa's most recent research has explored machine learning based website content classification for the financial domain.

Shruti Palaskar is a graduate student at Carnegie Mellon University's Language Technologies Institute advised by Prof. Florian Metze. Her research interests include multimodal machine learning, representation learning, speech recognition and summarization.

Ramon Sanabria is a graduate student advised by Prof. Florian Metze at the Language Technologies Institute (LTI), School of Computer Science, Carnegie Mellon University (CMU). His main research interests are in the areas of machine learning and sequence analysis. More concretely, his current research focuses on adding physical and abstract context to speech recognition systems. He is also the maintainer of the EESEN toolkit.

Josiah Wang is a Senior Teaching Fellow at Imperial College London. His main research interest is in generalisable Machine Learning approaches that use few or no task-specific examples to tackle tasks at the intersection of Computer Vision and Natural Language Processing, with an emphasis on integrating both modalities for better general text and image understanding. He received his Ph.D. from the University of Leeds in 2013, working with the late Dr. Mark Everingham (from the Vision group) and Prof. Katja Markert (from the Natural Language Processing group) to develop algorithms for visual object recognition by learning from textual descriptions.

Raman Arora is an Assistant Professor in the Department of computer science at the Johns Hopkins University. His research interests include machine learning, representation learning, stochastic optimization, and privacy and robustness in machine learning.