Beyond Instructional Videos:
Probing for More Diverse Visual-Textual Grounding on YouTube

Jack Hessel
Cornell University
jhessel@cs.cornell.edu

Zhenhai Zhu   Bo Pang   Radu Soricut
Google Research
{zhenhai,bopang,rsoricut}@google.com

Abstract

Pretraining from unlabelled web videos has quickly become the de-facto means of achieving high performance on many video understanding tasks. Features are learned via prediction of *grounded* relationships between visual content and automatic speech recognition (ASR) tokens. However, prior pretraining work has been limited to only instructional videos, a domain that, *a priori,* we expect to be relatively "easy:" speakers in instructional videos will often reference the literal objects/actions being depicted. Because instructional videos make up only a fraction of the web's diverse video content, we ask: can similar models be trained on broader corpora? And, if so, what types of videos are "grounded" and what types are not? We examine the diverse YouTube8M corpus, first verifying that it contains many non-instructional videos via crowd labeling. We pretrain a representative model on YouTube8M and study its success and failure cases. We find that visual-textual grounding is indeed possible across previously unexplored video categories, and that pretraining on a more diverse set still results in representations that generalize to both non-instructional and instructional domains.

1   Introduction

Self-supervised pretraining approaches have recently been adapted to web videos (Sun et al., 2019a,b; Miech et al., 2019, 2020; Zhu and Yang, 2020; Amrani et al., 2020); the resulting models have achieved state-of-the-art performance on a wide range of video understanding tasks, e.g., dense caption generation, action localization, etc.

In general, the pretraining step requires a large, unlabelled corpus of web videos. The training objective aligns visual content (i.e., video segments) with automatic speech recognition (ASR) tokens, and the resulting representations are fine-tuned for downstream tasks. The assumption underlying this family of approaches is that the words spoken in a video scene have *some* consistent relationship with the temporally corresponding visual content.

However, in contrast to the highly diverse corpora utilized for text-based pretraining (Wikipedia, Common Crawl, etc.), pretraining for web videos so far has been limited to instructional videos. This domain restriction is motivated by the commonly-accepted notion that "procedural knowledge tends to be inherently multimodal" (Malmaud et al., 2015): we *expect* that the semantic information in video frames and ASR tokens is readily correlated in instructional videos. But corpus diversity brings significant benefits: in the text-only case, models can effectively represent diverse real-world entities (Roberts et al., 2020) precisely because pretraining is not restricted to, e.g., only fictional stories (Zhu et al., 2015).

In search of more general representations, our main question is: **does video-ASR pretraining "work" for more diverse pretraining corpora?** Are certain categories of non-instructional videos "groundable," thus enabling diverse representation learning? Or are some types too difficult, only acting as training noise? We conclude that: 1) grounding is indeed possible in a wide range of yet-to-be-computationally-exploited YouTube video categories, e.g., walk-throughs, vehicles, tech reviews, etc., with some harder than others; 2) for the model we consider, not much representational power is gained or lost by switching from a pure instructional training set to a diverse one, which may additionally provide more versatility.

2   Related Work

ASR is known to be a useful signal source in various instructional video understanding tasks (Gupta et al., 2017; Huang et al., 2017; Huang* et al.,

2018; Moriya et al., 2019), e.g., action detection/classification (Yu et al., 2014; Alayrac et al., 2017; Chang et al., 2019; Kuehne et al., 2019), segmentation/captioning (Sener et al., 2015), and instruction alignment (Malmaud et al., 2015; Alayrac et al., 2016). A number of multimodal instructional video datasets have been proposed (Wang et al., 2019; Tang et al., 2019; Sanabria et al., 2018). A notable recent example of a non-instructional video corpus is Ignat et al. (2019), who analyze grounded-ness in a "lifestyle vlogs" corpus. Fouhey et al. (2018) highlight the difference between keyword search, and implicitly mining action data of interest from a broader corpus (e.g., Bregler (1997); Gu et al. (2018)).

Operational grounding. Our work builds upon prior operational notions of grounding: if an algorithm is able to consistently predict specific visual-textual relationships, then that relationship is said to be "grounded" (Lu et al., 2008; Berg et al., 2010; Parikh and Grauman, 2011; Hill and Korhonen, 2014; Hessel et al., 2018). For example, Yanai and Barnard (2005) rank "substrings of text by how well their occurrence can be predicted from visual features."

3 Video-ASR pretraining + our model

Recent work in designing pretraining objectives assumes that: 1) ASR tokens have, on average, some correspondence to temporally co-occurring video frames within the same video; and 2) video clips lacking ASR tokens can be ignored, i.e., models are not expected to align or reason about clips that lack ASR.

We build and analyze a model that encapsulates both of these assumptions.[1] The model is a slight simplification of Miech et al. (2019), where a joint embedding for the visual content and ASR tokens is learned. In the supplementary material, we detail replications of key experiments of their work: our model achieves comparable results on a difficult action localization task, CrossTask (Zhukov et al., 2019), when pretrained on the same corpus of 1M instructional videos.

Model details. The similarity between clip i and ASR caption j, $s_{i,j}$, is estimated by computing the cosine similarity between their corresponding embeddings in the joint space.[2] During training, tem-

---

[1]While more complicated models are possible, our goal is to conduct an error analysis of a simple, representative model, not to necessarily achieve state-of-the-art results.

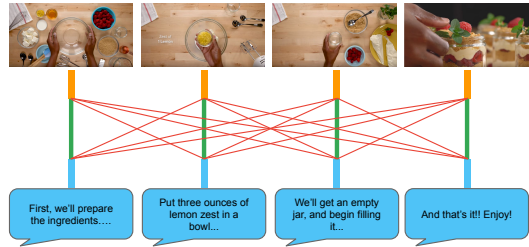[2]Joint embedding models are parameterized using gated

Figure 1: Intra-video AUC metric: the model scores all possible links between clips and ASR captions within a single video; the model is rewarded for assigning higher similarity to **temporally-aligned segments** versus **mismatched ones**.

porally corresponding (clip, ASR caption) pairs are sampled ("$\mathcal{P}$ositive" cases). For each positive case, a set of mismatched "$\mathcal{N}$egative" cases is also sampled both from other videos and from the same video in equal proportion. In contrast to Miech et al. (2019), we control for clip length, and sample temporally *fixed-length* segments: this simplifying choice makes our error analysis significantly more straightforward, and results in minimal performance change.[3] The following hinge loss is minimized for margin δ:

$$\sum_{i,j \in \mathcal{P}, \mathcal{N}} \max(0, \delta + s_{i,j} - s_{i,i}) + \max(0, \delta + s_{j,i} - s_{i,i}) \tag{1}$$

Measuring visual-textual alignment. Viewed through the lens of link prediction between truly co-occuring (clip, ASR) pairs, Eq. 1 can be seen as a differentiable approximation of AUC (Rendle et al., 2009). Thus, we propose to operationally measure the groundedness using *intra-video* AUC: a single score is assigned to each video, rewarding the model if it is able to successfully align temporal pairs within the same video (and penalizing it if not). Fig. 1 presents a visualization of this method.

4 A More Diverse Corpus

YouTube-600K. YouTube8M (Abu-El-Haija et al., 2016) is a dataset of 6.1M YouTube videos,[4] where each video is labeled across 3K

---

multi-layer feedforward networks. Visual features are: frame-wise 2D Inception-v3 pretrained for object detection (Szegedy et al., 2016; Sun et al., 2017) and 3D CNN S3D-G features pretrained for action recognition (Xie et al., 2018; Kay et al., 2017).

[3]We use 5 second segments, and initially randomly sample 256 segments per video before discarding ones that have no temporally-accompanying ASR. Segments may overlap.

[4]v3 of the dataset is smaller than v1/v2, due to videos becoming unavailable over time and other refinements.
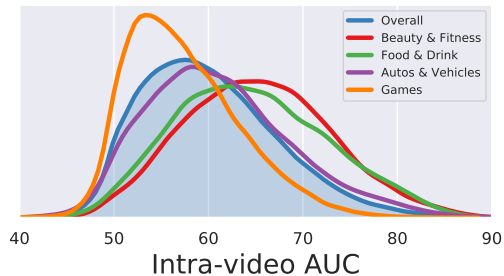
Figure 2: Distribution of intra-video AUC scores, grouped by meta category, compared to the overall distribution.

Domain	Example Categories (AUC)
Vehicles	Crossover SUV (70); Sedan (69); Minivan (69); Station wagon (68)
Walkthroughs	Hotel Suite (71); Apartment (69); Dining room (68); Living room (68)
Advertising	Advertising (68); Television advertisement (66); Infomercial (63)
Tech Reviews	CNET (66), Netbook (65), Asus (63) IPhone 5S (64), MacBook (64)
Toys	Funko (66); Monster High (64); Figurine (64) Action Figure (64)
Appliances	Home appliance (65); Washing machine (64); Kitchen stove (64)
Places	Greenhouse (62) University (61); Amusement park (59)

Table 1: Domains with high operational groundability.

categories, ranging from "cooking" to "games" to "nature." It is among the largest and most diverse publicly available dataset of YouTube videos. Due to user deletions and videos without detected spoken words, we are able to collect ASR via the YouTube API for 1.4M (29%) videos; we further filtered to 817K videos tagged with English ASR. Maintaining the train / validation split of the original data release yields 639K training videos (henceforth referred to as YouTube-600K) and 167K validation-set videos.

Human annotation of "Is-it-instructional" While a qualitative examination of YouTube8M reveals clear topical and stylistic diversity compared to domain-restricted corpora, we quantitatively verify that YouTube8M does not consist of mostly instructional videos.

To that end, we sample 6.8K videos with English ASR from the validation set for human labeling. Each video is shown to three paid annotators, who must each provide a Yes/No answer to the question: "Does this video focus on real-world human actions accompanied by procedural language that explains what is happening on screen in reasonable detail?" [5] After a few iterations over the guidelines and examples, the annotators reach high agreement: in 96% of cases, all three judges are unanimous. Based on these annotations, we estimate that around 74% of the videos in the YouTube-600K corpus are not instructional.

Which categories are easiest/hardest? We train our model on YouTube-600K, and compute intra-

video AUC for each of the 178K validation videos. First, we average all videos labeled with a particular category to produce a per-category AUC score. The performance in a vast majority of categories is above the 50 AUC (random baseline), and ranges from 51 ("Mixtape") to 76 ("Muffin").[6] This suggests that at least *some* aspect of most categories of videos can be visual-textually grounded.

We next coarsely aggregate the YouTube8M categories into meta-categories, e.g., "Food and Drink."[7] The AUC distribution of 4 popular meta categories relative to the overall AUC distribution is given in Fig. 2. In general, the grounding succeeds most readily on makeup/hair videos (e.g., "Eye liner" AUC = 74, "Updo" AUC = 68, etc.) and cooking videos (e.g., "Vegetarian cuisine" AUC = 71), domains that have been previously used in video grounding work. Besides these already-studied domains, other high-scoring category types emerge (Table 1). Conversely, some categories are more difficult for the model. Video games tend to be difficult, e.g., "RuneScape" AUC = 54, "First-person shooter AUC = 55; speakers in these videos often reference diverse topics unrelated to the game itself. Non–video-game categories can also be difficult, e.g., "Unidentified flying object" AUC = 56, "Dashcam" AUC = 54.

We next ask: *are instructional videos indeed easier to ground?* While human judgements of instructional-ness and intra-video AUC are positively correlated $\rho = .20$ ($p \ll 0$), the low magnitude of this correlation provides additional empirical confirmation that other types of videos are

---

[5] Note that our definition for "is-instructional" intended to include the usual "how-to" videos, but also attempted to capture a more general notion of "instructional-ness". For instance, an un-boxing video where parts of a product are *taken out and assembled* along with corresponding narration should receive "Yes", whereas a video showing only a product from different angles should receive "No", due to a lack of narrated human actions.

[6] To make sure that the model is not succeeding simply because a category happened to be frequent in the dataset, we note the correlation between category AUC and category frequency is essentially zero ($\rho = .02, p > .58$).

[7] These meta-categories are called "verticals," and are released with YouTube8M.

(a) Action Figure (64 category AUC)
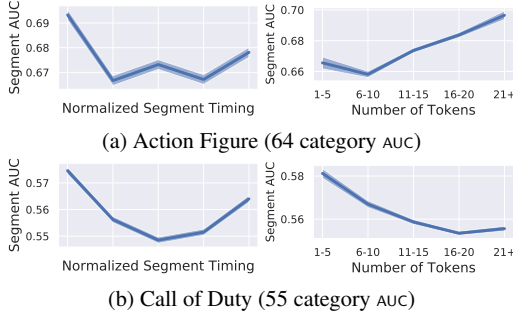
(b) Call of Duty (55 category AUC)

Figure 3: Correlation between per-segment AUC scores and segment timing within video (left column) and the number of tokens in a segment (right column)

also promising.

Within-category observations. To this point, we have identified broad categories of YouTube videos that are more groundable than others. However — it is not yet clear *why*, e.g., the algorithm gets 64 AUC on "Action Figure," or 55 AUC on "Call of Duty" (a first-person shooter game). We now define a *segment-level* AUC metric, analogous to the intra-video AUC metric previously defined: it quantifies how readily *individual ASR captions* are temporally localized by the model within the same video (see Menon and Elkan (2011) for a description of different AUC variants).

While visual/textual content plays a role in groundability, contextual factors must be considered too. Fig. 3 illustrates clear relationships 1) between ASR caption placement within a video and segment AUC (segments at the very beginning and very end of videos tend to be easier); and 2) between the number of tokens in an ASR caption and segment AUC. For "Action Figure" — ASR segments with more words are easier (this is the case with most categories), but for "Call of Duty", the opposite is true.

Additionally, we train OLS regression models to predict segment AUC from lexical unigram features, while controlling for timing/length features. Lexical features add predictive capacity ($p \ll .01$, F-test). While we find some patterns, e.g., intro/outro-language (e.g., "hey", "welcome", "peace") predictive of segment AUC for both categories, we also observe topical patterns, e.g., several unigrams associated with specific action figure body parts ("knee", "shoulder", "joint", etc.) are positively associated with segment AUC.

5 Implications for Pretraining

While self-grounding is possible for a diverse set of domains, do we gain anything by training on a more diverse corpus? Or do difficult-to-ground videos introduce noise, rendering its representations useless for downstream tasks? We compare two versions of our model: one with parameters learned from training on the diverse YouTube-600K corpus (M_{Diverse}), and one with parameters learned from a domain-specific corpus of 1M instructional videos ($M_{\text{Instructional}}$).

First, we evaluate each model's capacity to localize instructional steps on the CrossTask (Zhukov et al., 2019) dataset. Both models have similar performance for this instructional video task: macro-average task recall drops by only 15% when swapping from $M_{\text{Instructional}}$ to M_{Diverse}, even though YouTube-600K only contains 26% instructional videos (full results in supplementary).

We next evaluate each model's performance on the same-video clip alignment task over a diverse set of videos: the sample of 6.8K human-annotated videos from the YouTube8M validation set. In terms of intra-video AUC, M_{Diverse} outperforms $M_{\text{Instructional}}$ on 61% of videos. If we split the data across the "Is-it-instructional" human judgements and compare the two models in each subset, $M_{\text{Instructional}}$ "wins" in 57% of the instructional videos, whereas M_{Diverse} "wins" in 65% of non-instructional cases.

In short, both models achieve reasonable performance under instructional vs. non-instructional train/test domain mismatch. This is a promising result for future pretraining work with more diverse corpora: at least for these evaluations, good performance on an instructional video grounding task is still possible under domain shift. While the comparison of intra-video AUC is not necessarily definitive, it suggests that diverse corpora may provide more versatility, and we look forward to exploring this further in future work.

6 Conclusion

Peeking through the lens of a joint embedding model, we probe into learning visual-textual grounding over a more diverse corpus of YouTube videos vs. prior work. We find that learning visual-textual grounding is possible across many yet-to-be-explored categories of YouTube videos, and that it's possible to learn generalizable representations from a more diverse video set.

Acknowledgements

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *CVPR*.

Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. 2017. Joint discovery of object states and manipulation actions. In *ICCV*.

Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. 2020. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*.

Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *ECCV*.

Christoph Bregler. 1997. Learning and recognizing human dynamics in video sequences. In *Computer Society Conference on Computer Vision and Pattern Recognition*.

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. 2019. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *CVPR*.

David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. 2018. From lifestyle vlogs to everyday interactions. In *CVPR*.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*.

Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. 2017. Visual features for context-aware speech recognition. In *ICASSP*.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *NAACL*.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably cant see what i mean. In *EMNLP*.

De-An Huang*, Shyamal Buch*, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding "it": Weakly-supervised, reference-aware visual grounding in instructional videos. In *CVPR*.

De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Niebles. 2017. Unsupervised visual-linguistic reference resolution in instructional videos. In *CVPR*.

Oana Ignat, Laura Burdick, Jia Deng, and Rada Mihalcea. 2019. Identifying visible actions in lifestyle vlogs. In *ACL*.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. 2019. Mining youtube-a dataset for learning fine-grained action concepts from webly supervised video data. *arXiv preprint arXiv:1906.01012*.

Yijuan Lu, Lei Zhang, Qi Tian, and Wei-Ying Ma. 2008. What are the high-level concepts with small semantic gaps? In *CVPR*.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? interpreting cooking videos using text, speech and vision. In *NAACL*.

Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In *ECML-PKDD*.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Yasufumi Moriya, Ramon Sanabria, Florian Metze, and Gareth JF Jones. 2019. Grounding object detections with transcriptions. *arXiv preprint arXiv:1906.06147*.

Devi Parikh and Kristen Grauman. 2011. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *NeurIPS Workshops*.

Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Unsupervised semantic parsing of video collections. In *ICCV*.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. VideoBERT: A joint model for video and language representation learning. In *ICCV*.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*.

Weiying Wang, Yongcheng Wang, Shizhe Chen, and Qin Jin. 2019. Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. In *EMNLP*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*.

Keiji Yanai and Kobus Barnard. 2005. Image region entropy: a measure of visualness of web images associated with one concept. In *ACM MM*.

Shoou-I Yu, Lu Jiang, and Alexander Hauptmann. 2014. Instructional videos for unsupervised harvesting and learning of action examples. In *ACM MM*.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *CVPR*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *CVPR*.

Supplementary Material for "Beyond Instructional Videos: Probing for More Diverse Visual-Textual Grounding on YouTube"

Jack Hessel
Cornell University
jhessel@cs.cornell.edu

Zhenhai Zhu **Bo Pang** **Radu Soricut**
Google Research
{zhenhai,bopang,rsoricut}@google.com

We adapt Miech et al. (2019)'s joint embedding model that pre-trains by aligning ASR tokens with corresponding video frames. The main difference between our implementation and theirs is how we generated (ASR, caption) pairs. While we considered generating clips according to their methodology, we ran into two problems. First, in early experiments, we found that the interpretability our error analysis was significantly impacted by varying clip length. For example: we were worried that it might not be consistent to compare the model's ability to temporally ground a 1s clip vs. a 15 second clip. There was also high correlation between caption length and temporal clip duration, which further complicated interpretation. Sampling clips of uniform duration solved these problems. Second, Miech et al. (2019)'s temporal segmentation was generated by relying on the *scrolling timing* of the ASR tokens on the YouTube, i.e., the time that YouTube decides to generate a linebreak, removing a line of caption from the screen. Via manual inspection, we found that scrolling time was temporally unreliable, e.g., the time in which ASR captions scroll on YouTube often differs significantly from when particular words were said. Instead, we sample 256 candidate 5 second segments uniformly at random from the video, and then discard segments that have no corresponding ASR.

Visual features. Following (Xu et al., 2016; Miech et al., 2019), we extract clip features from both 2D and 3D convolutional networks. For 2D features, we sample frames at 1FPS from all of the videos in our corpus, resize frames to be 256 by 256, and pass them through Inception-v3 (Szegedy et al., 2016) pretrained on JFT (Sun et al., 2017). For 3D convolutional networks, we follow a similar procedure to (Sun et al., 2019), sample frames at 30FPS, aggregate frames into one second non-overlapping clips of 1 second each, and run an S3D-G (Xie et al., 2018) network that is pretrained on the Kinetics action recognition dataset (Kay et al., 2017). Both 2D and 3D features are L2 normalized. The result of this process is a 2524-D feature vector for each second of video in our corpus. To extract a single feature vector for each clip, following Miech et al. (2019), we max pool token embeddings.[1]

ASR features. To compute ASR caption features for a given segment, after decapitalizing and tokenizing, each vocabulary item is assigned a 300 dimensional embedding parameter; these are fine-tuned during the training process. Once again following Miech et al. (2019), we max pool token embeddings to achieve a single embedding for an ASR caption.[1] When training on YouTube-600K the vocabulary size is 61K.

Joint embedding. The similarity scores between (clip, caption) pairs is estimated by the cosine similarity between their corresponding embeddings in a joint space. To project into this joint space, clip/ASR caption embedding models are parameterized using gated multi-layer perceptions, as described in Miech et al. (2019); these parameters are trained to maximize the training objective.

Comparison to HowTo100M. To verify that our model simplifications didn't significantly hinder performance, we attempted to replicate key experiments from Miech et al. (2019). In particular, we sought to gather the pretraining corpus they used, HowTo100M, which consists of 1.22M videos. Because of, e.g., users deleting videos, we were able to gather features for 87% of the original set, 1.06M videos. We trained with the Adam optimizer (Kingma and Ba, 2014) starting with a learning rate of .001, with the margin parameter set to .1, but didn't undertake significant hyperparameter optimization. We terminate training after

---

[1]We experimented with more expressive models, e.g., recurrent networks, but none outperformed this simple approach.

	Make Kimchi Rice	Pickle Cucumber	Make Banana Ice Cream	Grill Steak	Jack Up Car	Make Jello Shots	Change Tire	Make Lemonade	Add Oil to Car	Make Latte	Build Shelves	Make Taco Salad	Make French Toast	Make Irish Coffee	Make Strawberry Cake	Make Pancakes	Make Meringue	Make Fish Curry	Average
Zhukov et al. (2019)	15.6	10.6	7.5	14.2	9.3	11.8	17.3	13.1	6.4	12.9	27.2	9.2	15.7	8.6	16.3	13.0	23.2	7.4	13.3
Supervised upper-bound (Zhukov et al., 2019)	19.1	25.3	38.0	37.5	25.7	28.2	54.3	25.8	18.3	31.2	47.7	12.0	39.5	23.4	30.9	41.1	53.4	17.3	31.6
HowTo100M (1.2M videos) → Crosstask (Miech et al., 2019)	33.5	27.1	36.6	37.9	24.1	35.6	32.7	35.1	30.7	28.5	43.2	19.8	34.7	33.6	40.4	41.6	41.9	27.4	33.6
↳ only 600K instructional videos																			32.6
↳ only 200K instructional videos																			31.1
Our HowTo100M (1.06M videos) → Crosstask	34.0	31.3	37.8	32.0	19.6	29.8	15.9	35.4	23.1	24.0	38.6	25.4	35.8	27.6	43.3	42.0	45.8	29.0	31.7
Our YouTube-600K (639K videos; 166K instr) → Crosstask	21.5	24.7	35.2	26.2	19.6	29.5	25.8	30.1	20.9	22.9	32.7	18.4	26.7	27.2	31.0	43.3	37.7	20.5	27.4

Table 1: Comparison between our simplified model and Miech et al. (2019)'s model on CrossTask, and the effect of pretraining the model on YouTube-600K instead of HowTo100M. Note that HowTo100M → Crosstask results are pre-trained on less data when compared to the original works due to video deletion.

300K steps.

We verify the performance of our model is comparable using the CrossTask localization task (Zhukov et al., 2019). While we defer details to the original paper, the goal is to temporally localize a set of procedural steps for a task in an unlabelled/unsegmented video depicting that task. An algorithm's performance is evaluated with a recall metric (higher is better). We follow the evaluation procedure given in Miech et al. (2019), except instead of embedding each frame individually, we embed a sliding 5-second window of video clips. We also experimented with Zhukov et al. (2019)'s dynamic programming postprocessing method, and found that it usually resulted in a small performance increase.

Our model simplified model performs only slightly worse (5%) than Miech et al. (2019)'s. While we argue that our model is certainly still representative, there are several reasons why this gap might exist. For example, there may be a regularizing effect when the model is allowed to view clips of varying length. Furthermore, our feature set was different — we used different base neural networks for feature extraction. Also, our model is trained on less data due to authors deleting their videos. Finally — we didn't tune the training hyperparameters for our model/implementation, e.g., hinge size, learning rate, batch size, etc.

Stability of results. To ensure the results in our paper related to intra-video AUC were insensitive to the particular choice of model checkpoint, we re-did the experiments in §4 using a version of our model checkpointed at 140K iterations vs. the 300K presented in the main paper; these experiments were conducted over 21K dev videos instead of the full 167K dev videos presented in the main paper. Figures+tables in that section were consistent with the presented results, and the qualitative observations about the "Action Figure" category held.

References

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijaya-narasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A joint model for video and language representation learning. In *ICCV*.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gok-berk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *CVPR*.