Use What You Have: Video retrieval using representations from collaborative experts - Supplementary Material

A Supplementary Material

A.1 Detailed Description of Datasets

MSR-VTT [20]: This large-scale dataset comprises approximately 200K unique video-caption pairs (10K YouTube video clips, each accompanied by 20 different captions). The dataset is particularly useful because it contains a good degree of video diversity, but we noted a reasonably high degree of label noise (there are a number of duplicate annotations in the provided captions). The dataset allocates 6513, 497 and 2990 videos for training, validation and testing, respectively. To enable a comparison with as many methods as possible, we also report results across other train/test splits used in prior work [14, 22]. In particular, when comparing with [14] (on splits which do not provide a validation set), we follow their evaluation protocol, measuring performance after training has occurred for a fixed number of epochs (100 in total).

MSVD [5]: The MSVD dataset contains 80K English descriptions for 1,970 videos sourced from YouTube with a large number of captions per video (around 40 sentences each). We use the standard split of 1,200, 100, and 670 videos for training, validation, and testing [19, 21][1]. Differently from the other datasets, the MSVD videos do not have audio streams.

LSMDC [17]: This dataset contains 118,081 short video clips extracted from 202 movies. Each video has a caption, either extracted from the movie script or from transcribed DVS (descriptive video services) for the visually impaired. Evaluation is performed on a test set consisting of 1000 videos from movies disjoint from the training set, as outlined by the Large Scale Movie Description Challenge (LSMDC). [2]

ActivityNet-captions [12]: ActivityNet Captions consists of 20K videos from YouTube, coupled with approximately 100K descriptive sentences. We follow the paragraph-video retrieval protocols described in [23], training up to 15 epochs and reporting performance on `val1` (this train/test split allocates 10,009 videos for training and 4,917 videos for testing).

DiDeMo [1]: DiDeMo contains 10,464 unedited, personal videos in diverse visual settings with roughly 3-5 pairs of descriptions and distinct moments per video. The videos are collected in an open-world setting and include diverse content such as pets, concerts, and sports games. The total number of sentences is 40,543. While the moments are localised with

 *Equal Contribution

[1] Note: referred to by [16] as the JMET-JMDV split

[2] https://sites.google.com/site/describingmovies/lsmdc-2017

time-stamp annotations, we do not use time stamps in this work.

A.2 Ablation Studies - Full Tables

	Text ⟹ Video					Video ⟹ Text				
Method	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
RGB	4.6	15.8	24.6	42	171.9	8.7	23.2	33.4	27	185.8
RGB+Scene	4.9	16.4	25.1	41	166.8	8.4	23.2	33.2	27	174.5
RGB+Motion	5.7	18.3	28.1	33	144.8	10.0	26.8	38.8	19	122.7
RGB+Audio	7.0	22.0	32.4	26	107.2	12.0	30.2	43.2	14	83.1
RGB+OCR	7.0	22.4	33.2	25	102.2	11.3	29.8	41.0	16	102.9
RGB+ASR	7.3	22.6	33.4	25	101.1	12.0	31.2	43.6	15	102.3
RGB+Face	7.6	23.0	33.9	24	102.0	11.6	30.8	43.2	15	102.3

Table 1: Ablation study of importance of each expert when combined with RGB features.

	Text ⟹ Video					Video ⟹ Text				
Method	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
RGB	4.6	15.8	24.6	42	171.9	8.7	23.2	33.4	27	185.8
RGB+S	4.9	16.4	25.1	41	166.8	8.4	23.2	33.2	27	174.5
RGB+S+M	5.8	18.3	28.0	33	143.1	10.9	28.1	39.6	18	118.6
RGB+S+M+A	7.9	23.8	34.7	22	96.3	13.9	35.1	48.3	11	60.9
RGB+S+M+A+OCR	12.1	33.4	46.3	13	56.8	18.8	44.1	59.3	7	41.4
RGB+S+M+A+OCR+ASR	15.8	41.3	55.1	8	37.0	24.8	52.9	67.5	5	23.3
RGB+S+M+A+OCR+ASR+Face	22.2	51.9	65.2	5	22.6	33.3	63.8	77.0	3	14.5

Table 2: Ablation study of the importance experts on the MSR-VTT dataset.

		Text ⟹ Video					Video ⟹ Text				
Expert	No.Captions	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
RGB	20	4.6	15.8	24.6	42	171.9	8.7	23.2	33.4	27	185.8
Generic	1	5.8	18.3	28.0	33	143.1	10.9	28.1	39.6	18	118.6
All	1	12.4	34.8	47.8	12	47.0	20.1	46.1	60.7	7	27.2
All	20	22.2	51.9	65.2	5	22.6	33.3	63.8	77.0	3	14.5

Table 3: Ablation study of the number of captions in training on MSRVTT

A.3 Implementation Details

Appearance frame-level embeddings of the visual data are generated with a SENet-154 model [9] (pretrained on ImageNet for the task of image classification) from frames extracted at 5fps, where each frame is resized to 224 × 224 pixels. Features are collected from the final global average pooling layer, and have a dimensionality of 2048.

Motion embeddings are generated using the I3D inception model following the procedure described by [4]. Frames extracted at 25fps and processed with a window length of 64 frames and a stride of 25 frames. Each frame is first resized to a height of 256 pixels (preserving aspect ratio), before a 224 × 224 centre crop is passed to the model. Each temporal window produces a (1024x7)-matrix of features.

Face embeddings are extracted in two stages: (1) Each frame (also extracted at 25 fps) is resized to 300 × 300 pixels and passed through an SSD face detector [2, 13] to extract

bounding boxes; (2) The image region of each box is resized such that the minimum dimension is 224 pixels and a centre crop is passed through a ResNet50 [7] that has been trained for task of face classification on the VGGFace2 dataset [3], producing a 512-dimensional embedding for each detected face.

Audio embeddings are obtained with a VGGish model, trained for audio classification on the YouTube-8m dataset [8]. To produce the input for this model, the audio stream of each video is re-sampled to a 16kHz mono signal, converted to an STFT with a window size of 25ms and a hop of 10ms with a Hann window, then mapped to a 64 bin log mel-spectrogram. Finally, the features are parsed into non-overlapping 0.96s collections of frames (each collection comprises 96 frames, each of 10ms duration), which is mapped to a 128-dimensional feature vector.

Scene embeddings of 2208 dimensions are extracted from 224×224 pixel centre crops of frames extracted at 1fps using a DenseNet-161 [10] model pretrained on Places365 [24].

Speech to Text The audio stream of each video is re-sampled to a 16kHz mono signal. We then obtained transcripts of the spoken speech for MSRVTT, MSVD and ActivityNet using the the Google Cloud Speech to Text API [3] from the resampled signal. The language for the API is specified as English. For reference, of the 10,000 videos contained in MSRVTT, 8,811 are accompanied by audio streams. Of these, we detected speech in 5,626 videos.

Optical Character Recognition are extracted in two stages: (1) Each frame is resized to 800×400 pixels) and passed through Pixel Link [6] text detection model to extract bounding boxes for texts; (2) The image region of each box is resized to 32×256 and then pass through a CRNN model [18] that has been trained for text of scene text recognition on the Synth90K dataset[11], producing a character sequence for each detect box. They are then encoded via a pretrained word2vec embedding model [15].

Text We encode each word using the Google News[4] trained word2vec word embeddings [15]. All the word embeddings are then pass through a pretrained OpenAI-GPT model to extract the context-specific word embeddings (i.e., not only learned based on word concurrency but also the sequential context). Finally, all the word embeddings in each sentence are aggregated using NetVLAD.

[3] https://cloud.google.com/speech-to-text/

[4] GoogleNews-vectors-negative300.bin.gz found at: https://code.google.com/archive/p/word2vec/

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812, 2017.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

[6] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[8] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. URL https://arxiv.org/abs/1609.09430.

[9] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[11] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017.

[13] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[14] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[16] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27. ACM, 2018.

[17] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[18] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2017.

[19] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[20] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016.

[21] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[22] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018.

[23] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018.

[24] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.