Use What You Have: Video Retrieval Using Representations From Collaborative Experts - Supplementary Material

A Supplementary Material

A.1 Paper update, result corrections and summary of differences

Following the release of the initial version of this paper (which can be viewed for reference at https://arxiv.org/abs/1907.13487v1), a bug was discovered in our open-source software implementation which resulted in: (i) an overestimate of model performance; (ii) inaccurate conclusions about the relative importance of different experts on retrieval performance.

This correction to the paper contains repeats of each of the experiments reported in the initial paper, with the following changes: (1) the removal of the bug which affected previous results; (2) a systematic approach to hyperparameter selection (discussed in more detail below); (3) the inclusion of additional "expert" pretrained features (described in Sec. A.5) to assess the influence of feature strength within a modality. In addition to results, the written analysis has also been updated to reflect the corresponding changes in results. The authors would like to express their gratitude to Valentin Gabeur who identified the bug in the software implementation and enabled this correction.

Bug details: The bug caused information about feature availability in the ground truth target video to become available to the query encoder during both training and testing when computing embedding distances. The leak occurred through incorrect weighting of the embedding distances due to: (1) a leaking broadcasting operation in an existing open-source library [20] that was imported into our codebase; (2) incorrect NaN handling (introduced in our codebase), producing the same effect. The bug has now been patched in each of the open-source codebases that were known to have used this implementation.

A.2 Detailed Description of Datasets

MSR-VTT [29]: 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 [20, 31]. In particular, when comparing with [20] (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 [26, 30][1]. Differently from the other datasets, the MSVD videos do not have audio streams.

LSMDC [23]: 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. The validation set contains 7408 clips and evaluation is performed on a test set of 1000 videos from movies disjoint from the training and val sets, as outlined by the Large Scale Movie Description Challenge (LSMDC). [2]

ActivityNet-captions [15]: ActivityNet Captions consists of 20K videos from YouTube, coupled with approximately 100K descriptive sentences. We follow the paragraph-video retrieval protocols described in [32] training up to 200 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 time-stamp annotations, we do not use time stamps in this work.

A.3 Optimisation details and hyperparameter selection

For each dataset, a grid search was first performed (using the Lookahead solver [27, 33]) over batch sizes (16, 32, 64, 128, 256), learning rates (0.1, 0.01) and weight decay (1E-3, 5E-5) for each dataset using a single expert to determine appropriate optimisation parameters. Next, an experiment on MSR-VTT compared several choices for the dimensionality of the projection operation applied to the features (described in Sec. 3.1) (choosing among 512, 768 and 1024 dimensions), which suggested that 768 was most effective. This was then fixed for all remaining experiments (this represents a difference from the original paper, in which 512 was used). Further ablations (provided below) indicate that performance is not sensitive to this hyperparameter. Next, Asynchronous Hyperband [16] was used to select all remaining hyperparameters on MSR-VTT by partially evaluating 1k configurations on the validation sets for each dataset. These hyperparameters consisted of: the number of VLAD clusters and ghost clusters [34] used for different experts, the zero-padding length applied to variable-length experts, the margin hyperparameter m in Eq. 3, the Collaborative Gating architecture (whether to use batch normalization [13], the number of layers used to form the MLP, and the choice of activation function). The architecture choices were then fixed for all datasets. Note that to ensure a fair comparison on MSR-VTT with the MoEE method of [20] in Tab. 6, MoEE was also provided with a budget of 1k sampled configurations. To determine zero-padding, margin and VLAD clusters for DiDeMo, MSVD and LSMDC further Asynchronous Hyperband searches were conducted, each with a budget

[1] Note: referred to by [22] as the JMET-JMDV split

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

of 500 sampled configurations. Since, differently from the other datasets with available validation and test sets, the validation set itself is used to assess performance on ActivityNet, hyperparameters were copied from the DiDeMo configuration. The configurations, experts, pretrained models and logs for each of the experiments reported in this paper are made available as part of the updated open-source implementation at www.robots.ox.ac.uk/~vgg/research/collaborative-experts/.

A.4 Ablation Studies - Full Tables

	Text ⟹ Video					Video ⟹ Text				
Experts	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Scene	$4.0_{\pm 0.1}$	$14.1_{\pm 0.1}$	$22.4_{\pm 0.3}$	$50.0_{\pm 1.0}$	$201.3_{\pm 1.6}$	$5.6_{\pm 0.6}$	$18.2_{\pm 0.6}$	$27.7_{\pm 0.3}$	$39.0_{\pm 0.0}$	$247.0_{\pm 10.1}$
Scene+Speech	$4.6_{\pm 0.1}$	$15.5_{\pm 0.2}$	$24.4_{\pm 0.2}$	$44.7_{\pm 1.2}$	$183.6_{\pm 1.7}$	$6.0_{\pm 0.2}$	$20.4_{\pm 0.5}$	$30.3_{\pm 1.0}$	$33.0_{\pm 2.0}$	$222.6_{\pm 9.9}$
Scene+Audio	$5.6_{\pm 0.0}$	$18.7_{\pm 0.1}$	$28.2_{\pm 0.1}$	$33.7_{\pm 0.6}$	$140.8_{\pm 0.3}$	$8.2_{\pm 0.4}$	$24.8_{\pm 0.4}$	$36.0_{\pm 0.1}$	$21.7_{\pm 0.6}$	$127.9_{\pm 5.9}$
Scene+Action(KN)	$5.3_{\pm 0.3}$	$17.6_{\pm 0.8}$	$27.1_{\pm 0.9}$	$36.0_{\pm 1.7}$	$158.7_{\pm 1.6}$	$7.3_{\pm 0.6}$	$22.3_{\pm 1.4}$	$33.4_{\pm 1.7}$	$25.2_{\pm 2.0}$	$151.7_{\pm 11.6}$
Scene+Obj(IN)	$5.0_{\pm 0.2}$	$16.6_{\pm 0.7}$	$25.5_{\pm 1.0}$	$40.7_{\pm 2.1}$	$173.1_{\pm 3.3}$	$6.9_{\pm 0.5}$	$21.2_{\pm 0.9}$	$31.1_{\pm 1.9}$	$28.7_{\pm 3.8}$	$188.3_{\pm 4.7}$
Scene+Obj(IG)	$7.2_{\pm 0.1}$	$22.3_{\pm 0.3}$	$33.0_{\pm 0.2}$	$25.3_{\pm 0.6}$	$125.1_{\pm 0.1}$	$10.1_{\pm 0.3}$	$29.7_{\pm 0.5}$	$41.9_{\pm 0.7}$	$15.2_{\pm 0.9}$	$91.3_{\pm 2.4}$
Scene+Action(IG)	$6.8_{\pm 0.1}$	$21.7_{\pm 0.1}$	$32.4_{\pm 0.1}$	$25.7_{\pm 0.6}$	$122.1_{\pm 0.3}$	$9.4_{\pm 0.3}$	$27.8_{\pm 0.6}$	$40.1_{\pm 1.1}$	$17.2_{\pm 1.1}$	$87.8_{\pm 4.2}$
Scene+OCR	$4.1_{\pm 0.1}$	$14.1_{\pm 0.1}$	$22.2_{\pm 0.2}$	$50.3_{\pm 1.2}$	$203.1_{\pm 4.4}$	$5.4_{\pm 0.5}$	$18.6_{\pm 1.2}$	$26.6_{\pm 1.2}$	$40.0_{\pm 1.0}$	$292.6_{\pm 9.9}$
Scene+Face	$4.1_{\pm 0.1}$	$14.2_{\pm 0.3}$	$22.4_{\pm 0.4}$	$49.7_{\pm 0.6}$	$194.2_{\pm 5.1}$	$5.6_{\pm 1.0}$	$17.9_{\pm 0.7}$	$26.7_{\pm 0.8}$	$39.1_{\pm 2.6}$	$273.5_{\pm 6.3}$

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

	Text ⟹ Video					Video ⟹ Text				
Experts	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Scene	$4.0_{\pm 0.1}$	$14.1_{\pm 0.1}$	$22.4_{\pm 0.3}$	$50.0_{\pm 1.0}$	$201.3_{\pm 1.6}$	$5.6_{\pm 0.6}$	$18.2_{\pm 0.6}$	$27.7_{\pm 0.3}$	$39.0_{\pm 0.0}$	$247.0_{\pm 10.1}$
Prev.+Speech	$4.6_{\pm 0.1}$	$15.5_{\pm 0.2}$	$24.4_{\pm 0.2}$	$44.7_{\pm 1.2}$	$183.6_{\pm 1.7}$	$6.0_{\pm 0.2}$	$20.4_{\pm 0.5}$	$30.3_{\pm 1.0}$	$33.0_{\pm 2.0}$	$222.6_{\pm 9.9}$
Prev.+Audio	$5.8_{\pm 0.1}$	$19.0_{\pm 0.3}$	$28.8_{\pm 0.2}$	$32.3_{\pm 0.6}$	$136.8_{\pm 1.2}$	$8.6_{\pm 0.2}$	$26.1_{\pm 0.6}$	$37.8_{\pm 0.8}$	$19.8_{\pm 0.8}$	$117.7_{\pm 2.9}$
Prev.+Action(KN)	$6.7_{\pm 0.2}$	$21.8_{\pm 0.4}$	$32.5_{\pm 0.5}$	$25.3_{\pm 0.6}$	$115.9_{\pm 1.0}$	$9.9_{\pm 0.4}$	$28.6_{\pm 0.7}$	$41.7_{\pm 0.8}$	$15.7_{\pm 0.6}$	$77.9_{\pm 5.2}$
Prev.+Obj(IN)	$7.5_{\pm 0.1}$	$23.4_{\pm 0.0}$	$34.1_{\pm 0.2}$	$23.7_{\pm 0.6}$	$111.9_{\pm 0.6}$	$11.2_{\pm 0.3}$	$32.1_{\pm 0.8}$	$45.4_{\pm 0.6}$	$13.7_{\pm 0.6}$	$68.0_{\pm 1.4}$
Prev.+Obj(IG)	$9.5_{\pm 0.2}$	$27.7_{\pm 0.1}$	$39.4_{\pm 0.1}$	$18.0_{\pm 0.0}$	$92.6_{\pm 0.4}$	$14.7_{\pm 0.6}$	$38.9_{\pm 0.8}$	$53.1_{\pm 1.0}$	$9.3_{\pm 0.6}$	$45.6_{\pm 2.1}$
Prev.+Action(IG)	$9.9_{\pm 0.1}$	$28.6_{\pm 0.3}$	$40.7_{\pm 0.1}$	$17.0_{\pm 0.0}$	$86.4_{\pm 0.4}$	$15.5_{\pm 0.6}$	$40.1_{\pm 1.2}$	$54.4_{\pm 1.3}$	$8.7_{\pm 0.6}$	$39.4_{\pm 0.9}$
Prev.+ OCR	$10.0_{\pm 0.1}$	$28.8_{\pm 0.2}$	$40.9_{\pm 0.2}$	$16.7_{\pm 0.6}$	$87.3_{\pm 0.8}$	$15.2_{\pm 0.1}$	$41.1_{\pm 0.6}$	$54.6_{\pm 0.7}$	$8.5_{\pm 0.5}$	$38.5_{\pm 0.6}$
Prev.+ Face	$10.0_{\pm 0.1}$	$29.0_{\pm 0.3}$	$41.2_{\pm 0.2}$	$16.0_{\pm 0.0}$	$86.8_{\pm 0.3}$	$15.6_{\pm 0.3}$	$40.9_{\pm 1.4}$	$55.2_{\pm 1.0}$	$8.3_{\pm 0.6}$	$38.1_{\pm 1.8}$

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

		Text ⟹ Video					Video ⟹ Text				
Expert	Num. Caps	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Obj(IN)	1	$2.6_{\pm 0.1}$	$9.3_{\pm 0.4}$	$15.0_{\pm 0.7}$	$101.3_{\pm 15.5}$	$321.1_{\pm 35.1}$	$3.7_{\pm 0.3}$	$13.5_{\pm 0.6}$	$20.8_{\pm 0.4}$	$60.0_{\pm 2.0}$	$304.9_{\pm 15.}$
Obj(IN)	20	$4.9_{\pm 0.1}$	$16.5_{\pm 0.2}$	$25.3_{\pm 0.4}$	$40.7_{\pm 1.2}$	$169.1_{\pm 1.4}$	$6.9_{\pm 0.6}$	$21.0_{\pm 0.3}$	$31.3_{\pm 0.3}$	$30.0_{\pm 1.7}$	$201.6_{\pm 9.}$
All	1	$4.8_{\pm 0.2}$	$16.2_{\pm 0.5}$	$25.0_{\pm 0.7}$	$43.3_{\pm 4.0}$	$183.1_{\pm 19.6}$	$8.4_{\pm 0.5}$	$25.6_{\pm 0.7}$	$37.1_{\pm 0.2}$	$20.3_{\pm 0.6}$	$87.2_{\pm 6.7}$
All	20	$10.0_{\pm 0.1}$	$29.0_{\pm 0.3}$	$41.2_{\pm 0.2}$	$16.0_{\pm 0.0}$	$86.8_{\pm 0.3}$	$15.6_{\pm 0.3}$	$40.9_{\pm 1.4}$	$55.2_{\pm 1.0}$	$8.3_{\pm 0.6}$	$38.1_{\pm 1.8}$

Table 3: Ablation study of the number of captions in training on MSR-VTT

A.5 Implementation Details

Object frame-level embeddings of the visual data are generated with two models, *Obj(IN)* and *Obj(IG)*. *Obj(IN)* is an SENet-154 model [11] (pretrained on ImageNet for the task of image classification) from frames extracted at 25 fps, where each frame is resized to 224 × 224 pixels. *Obj(IG)* is a ResNext-101 [28] pretrained on Instagram hashtags [19], using

	Text ⟹ Video					Video ⟹ Text					
Dimension	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR	Params.
384	$9.4_{\pm 0.2}$	$27.8_{\pm 0.4}$	$39.8_{\pm 0.4}$	$17.7_{\pm 0.6}$	$88.8_{\pm 0.5}$	$14.0_{\pm 0.5}$	$38.7_{\pm 0.5}$	$52.7_{\pm 1.4}$	$9.3_{\pm 0.6}$	$41.8_{\pm 1.0}$	88.62M
512	$9.8_{\pm 0.3}$	$28.6_{\pm 0.4}$	$40.6_{\pm 0.4}$	$17.0_{\pm 0.0}$	$88.0_{\pm 0.7}$	$14.8_{\pm 0.4}$	$40.4_{\pm 0.6}$	$53.9_{\pm 0.4}$	$8.8_{\pm 0.3}$	$38.8_{\pm 1.5}$	119.51M
640	$10.1_{\pm 0.1}$	$28.8_{\pm 0.1}$	$40.9_{\pm 0.2}$	$16.7_{\pm 0.6}$	$87.6_{\pm 0.2}$	$15.6_{\pm 0.6}$	$41.3_{\pm 0.7}$	$55.0_{\pm 0.5}$	$8.3_{\pm 0.6}$	$37.3_{\pm 1.8}$	151.12M
768	$10.0_{\pm 0.1}$	$29.0_{\pm 0.3}$	$41.2_{\pm 0.2}$	$16.0_{\pm 0.0}$	$86.8_{\pm 0.3}$	$15.6_{\pm 0.3}$	$40.9_{\pm 1.4}$	$55.2_{\pm 1.0}$	$8.3_{\pm 0.6}$	$38.1_{\pm 1.8}$	183.45M
1024	$9.9_{\pm 0.1}$	$28.6_{\pm 0.3}$	$40.7_{\pm 0.4}$	$17.0_{\pm 0.0}$	$87.6_{\pm 1.1}$	$14.7_{\pm 0.4}$	$40.7_{\pm 0.8}$	$54.4_{\pm 0.3}$	$8.5_{\pm 0.5}$	$39.1_{\pm 1.7}$	250.27M

Table 4: Ablation study of the importance of model capacity by varying the shared embedding dimension used by CE on MSR-VTT.

the same frame preparation as *Obj(IN)*. Features are collected from the final global average pooling layer of both models, and have a dimensionality of 2048.

Action embeddings are similarly generated from two models, *Action(KN)* and *Action(IG)*. *Action(KN)* is an I3D inception model that computes features 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. *Action(IG)* is a 34-layer R(2+1)D model [25] trained on IG-65m [8] which processes clips of 8 consecutive 112×112 pixel frames, extracted at 30 fps (we use the implementation provided by [6]).

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, 17] 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 [9] 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 [10]. 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 [12] model pretrained on Places365 [35].

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 MSR-VTT, MSVD and ActivityNet using 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 MSR-VTT, 8,811 are accompanied by audio streams. Of these, we detected speech in 5,626 videos.

Optical Character Recognition is extracted in two stages: (1) Each frame is resized to 800×400 pixels) and passed through Pixel Link [7] 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 model [18, 24] that has been trained for text of scene text recognition on the Synth90K dataset[14], producing a character sequence for each detect box. They are then encoded via a pretrained word2vec embedding model [21].

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

Text We encode each word using the Google News[4] trained word2vec word embeddings [21]. 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.

[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] J. H. Daniel. ig65m-pytorch. https://github.com/moabitcoin/ig65m-pytorch, 2019.

[7] 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.

[8] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019.

[9] 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.

[10] 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.

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

[12] 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.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] 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.

[15] 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.

[16] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. Massively parallel hyperparameter tuning. *arXiv preprint arXiv:1810.05934*, 2018.

[17] 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.

[18] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Synthetically supervised feature learning for scene text recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[19] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[26] 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.

[27] Less Wright. Project title. https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer, 2019.

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pages 9593–9604, 2019.

[34] Yujie Zhong, Relja Arandjelović, and Andrew Zisserman. Ghostvlad for set-based face recognition. In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.

[35] 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.