Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval Supplementary Material

Max Bain[1] Arsha Nagrani[1†] Gül Varol[1,2] Andrew Zisserman[1]
[1] Visual Geometry Group, University of Oxford
[2] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS
{maxbain, arsha, gul, az}@robots.ox.ac.uk

Contents

1. Additional Benchmark Results

We evaluate our model on two other video retrieval benchmark datasets: LSMDC (Table 1) and ActivtyNet Captions (Table 2). We additionally evaluate on a standard `image` retrieval benchmark: Flickr30K (Table 3), demonstrating the versatility of our model to perform competitively for both images and video. These datasets are described in detail below.

1.1. LSMDC

LSMDC [15] consists of 118,081 video clips sourced from 202 movies. The validation set contains 7,408 clips and evaluation is done on a test set of 1,000 videos from movies disjoint from the train and val sets. This follows the protocol outlined in [16]. We outperform all previous methods, except for MMT in Median Rank, which pretrains on HowTo100M, a dataset consisting of over 100M clip-text pairs and contains multiple experts as well as audio modalities. Our model uses visual information alone.

Table 1: Text-to-video retrieval results on the LSMDC test set.

Method	**R@1**	**R@5**	**R@10**	**MedR**
JSFusion [19]	9.1	21.2	34.1	36.0
MEE [13]	9.3	25.1	33.4	27.0
CE [12]	11.2	26.9	34.8	25.3
MMT (HowTo100M) [6]	12.9	29.9	40.1	**19.3**
Ours	**15.0**	**30.8**	**40.3**	20.0

Table 2: Text-to-video retrieval results on the ActivityNet val1k set. **R@k:** Recall@K. **MedR:** Median Rank.

Method	E2E	VT PT	**R@1**	**R@5**	**MedR**
FSE			18.2	44.8	8.3
CE [12]			18.2	47.7	13.0
CLIPBERT	✓		21.3	49.0	6.0
MMT			22.7	54.2	5.0
SupportSet [14]			26.8	58.1	**3.0**
MMT [6]		HowTo	28.7	61.4	**3.0**
SupportSet [14]		HowTo	**29.2**	**61.6**	**3.0**
Ours	✓	CC,WebVid-2M	28.8	60.9	**3.0**

1.2. ActivityNet Captions

ActivityNet Captions [8] contains 20K YouTube videos focused on actions, annotated with 100K sentences. The training set consists of 10K videos, and we use the 'val1' set of 4.9K videos to report results. At test time we use paragraph-to-video retrieval as is standard protocol set by other works, where the segment descriptions are concatenated to give a video-level description. We compare to prior work in Table 2 and achieve comparable results to the state of the art by using much less training data.

Table 3: Text-to-**image** retrieval results on the Flickr30K test set. ++ indicates additional datasets: COCO Captions, SBU Captions. VisGenObjects denotes Visual Genome object bounding box annotations used to pretrain an FRCNN object feature extractor.

Method	Vis PT. size	R@1	R@5	R@10
SCANM [9]	VisGenObj (3.8M)	48.6	77.7	85.2
IMRAM [2]	VisGenObj (3.8M)	53.9	79.4	87.2
SGRAF [5]	VisGenObj (3.8M)	58.5	83.0	88.8
Ours	CC (3.0M)	54.2	83.2	89.8
Ours	CC,WV-2M (5.5M)	61.0	87.5	92.7

1.3. Flickr30K

We also evaluate on a text-to-image retrieval benchmark to demonstrate the versatility of our model in that it can be used to achieve competitive performance in image settings as well as state-of-the art in video retrieval. The Flickr30K [18] dataset contains 31,783 images with 5 captions per image. We follow the standard protocol of 1,000 images for validation, 1,000 images for testing and the remaining for training. We report the results in Table 3. Unlike other works [2, 5, 9] which utilise high resolution regions extracted using a Faster-RCNN detector, our model is single stage and does not require any object detections. We compare to works with a similar number of training image-text pairs, and find that our model is comparable. We also note that training on WebVid2M provides a sizeable boost (5% improvement in R@1). Note that there are other recent text-image works such as UNITER [3] and OSCAR [11], however these are trained on almost twice the number of samples. Recent works scale this up even further to billions of samples (ALIGN [7]).

2. Architectural Details

2.1. Video Encoder

The video encoder is composed of: (i) the patch embedding layer; (ii) learnable positional space, time and [CLS] embeddings; and (iii) a stack of $|\ell| = 12$ space-time attention blocks

1. The patch embedding layer is implemented as a 2D convolutional layer with a kernel and stride size equivalent to the target patch size $P = 16$, and $d = 768$ output channels (the chosen embedding dimensionality of the video encoder).

2. The positional space and time embeddings are instantiated with shape $M \times d$ and $N \times d$ respectively, where M is the maximum number of input video frames and N is the maximum number of non-overlapping patches of size P within a frame (196 for a video resolution of 224×224). The [CLS] embedding is instantiated with shape $1 \times d$.

3. Each space-time attention block consists of norm layers, temporal and spatial self-attention layers, and an MLP. The order and connections of these layers is shown in Figure 1.

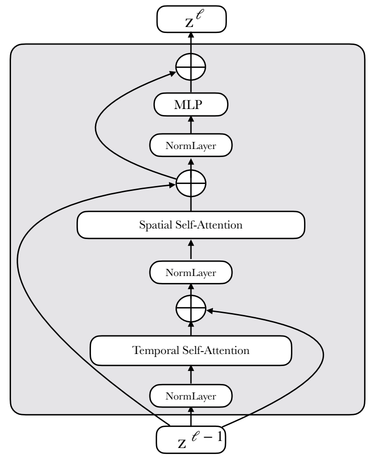

Figure 1: Detailed diagram of the space-time self attention block.

2.2. Text Encoder

Our text encoder is instantiated as `distilbert-base-uncased` [17]. Distilbert follows the same general architecture as BERT [4], but with the number of layers reduced by a factor of 2 and the token-type embeddings and the pooler removed. We use the HuggingFace[1] transformers library implementation.

3. Architectural Ablations

3.1. Video Backbone

We investigate the effects of using different video backbone architectures (Table 4) and find that the space-time transformer encoder leads to large improvements in performance on MSR-VTT when compared to ResNets and 3D variants thereof.

During testing, all frame-variants see an equal number of frames, since the video embeddings are averaged over multiple strides.

For the video backbone ablation, we fix the text backbone to `distilbert-base-uncased`. For the text backbone ablation, we fix the video backone to the base space-time transformer with an input resolution of 224 and a patch size $P = 16$.

[1] https://huggingface.co/

Table 4: **Video backbone.** Text-to-video retrieval results on MSR-VTT test set with different video backbones. All models were pretrained on WebVid-2M and finetuned on MSR-VTT train set. 4 frames were given as input, except for the ResNet-101 which only supports image (1-frame) inputs. The text backbone is fixed to distilbert-base-uncased.

Video Backbone	#params	R@1	R@10	MedR
ResNet-101	45M	11.5	44.1	14.5
S3D-G	76M	3.6	20.4	59.5
R(3D)-101	85M	9.3	38.3	20.0
S-Tformer 224_{16} B	114M	**26.8**	**68.2**	**4.0**

Table 5: **Text backbone.** Text-to-video retrieval results on MSR-VTT test set with different text backbones. All models were pretrained on WebVid-2M and finetuned on MSR-VTT train set. The video backbone is fixed to the base space-time transformer with an input resolution of 224 and a patch size $P = 16$.

Text Backbone	#params	R@1	R@10	MedR
t5-small	60.5M	15.1	51.4	10.0
t5-base	222.9M	24.0	62.8	6.0
distilbert-base-uncased	66.4M	26.8	**68.2**	**4.0**
bert-base-uncased	109.5M	**27.5**	67.3	**4.0**

Table 6: **Space-time attention method:** Zero-shot results are presented on 1K-A MSR-VTT test set for text-video retrieval. The models were trained on WebVid-2M.

Attention Method	R@1	R@10	MedR
Divided Space-Time [10]	13.0	40.2	18.0
Ours	14.6	42.7	16.0

3.2. Text Backbone

The choice of text backbone has a significant impact on downstream performance (Table 5), with the t5 models performing significantly worse with more or similar numbers of parameters. DistilBERT and normal BERT achieve similar performance, with DistilBERT having far fewer parameters, therefore we chose to use DistilBERT in our work for efficiency.

3.3. Space-Time Attention

Space-time attention. Our modified space-time attention block, shown in Fig. 2, improves retrieval performance, as show in Table 6. We compare both variants during pretraining on WebVid-2M by reporting zero-shot results on MSR-VTT. We find once again that our modification leads to modest performance gains.

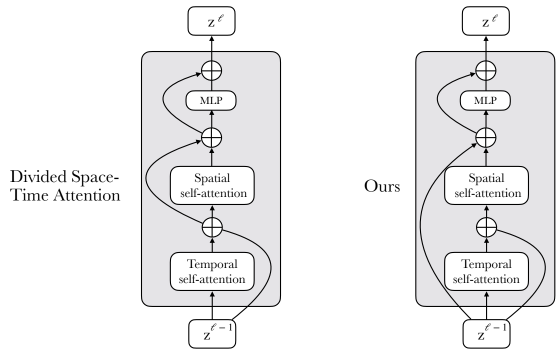

Figure 2: **Attention block:** The original divided block used in the Timesformer [1] architecture (left) compared to ours (right). We find that this minor modification of the input residual connection trains more quickly and is more stable than the original.

3.4. Temporal Expansion

Table 7: **Temporal expansion method**. The effect of different expansion methods increasing the input number of frames from 4⟹8. Results are presented on 1K-A MSR-VTT test set for text-video retrieval. The models were pretrained on CC3M & WebVid-2M and finetuned on MSR-VTT train set.

Method	R@1	R@10	MedR
Zero-pad	**30.7**	68.3	4.0
Nearest Neighbour	29.4	69.5	4.0
Bilinear	28.3	**69.9**	4.0

We explore 3 different methods for expanding temporal positional embeddings (zero-padding and two interpolation methods), and observe robustness to all 3 (see Table 7).

4. Pretraining on other datasets

In Table 8, we restrict the pretraining of our model to COCO Captions, a dataset with only 600k image-text pairs. We demonstrate that we are able to achieve generally competitive performance on MSR-VTT. We outperform ClipBERT – which trains on both COCO Captions and Visual Genome (totalling 5.6M image-text pairs) – by several percentage points, demonstrating the strength of our proposed architecture.

5. WebVid-2M Dataset Details

In this section, we show further details of the new WebVid-2M dataset. More qualitative examples of video-text pairs can be found in Figure 3 and histograms of caption

1990s: man driving excavator, rotates seat, opens windows in cab. hand presses lever.

Frying pancakes in the kitchen at home. a woman is cooking traditional russian pancakes. modern kitchen, skillet and batter.

Twilight zhuhai famous mountain park top cityscape aerial panorama 4k timelapse china

A child with a suitcase. a happy little girl sits on a suitcase with a passport and money.

Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing'

Cockatoos on the fence

Runners feet in a sneakers close up. realistic three dimensional animation.

Ontario, canada january 2014 heavy pretty snow on tree branches

Figure 3: **WebVid-2M dataset examples:** We provide additional examples from our dataset by showing video-text pairs, using video thumbnails.

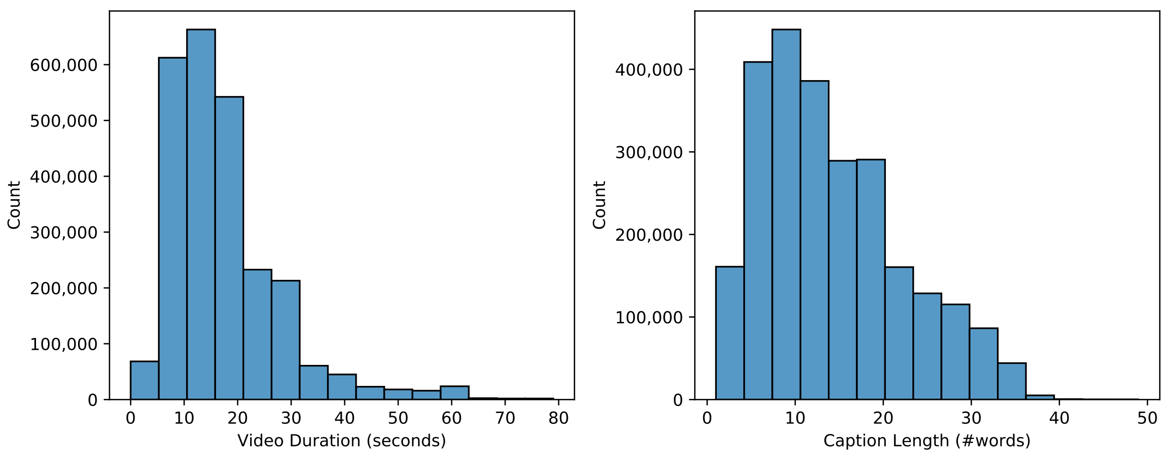

Figure 4: **WebVid-2M dataset statistics:** We report the histogram of video duration in seconds **(top)** and the histogram of caption length in words (**bottom**).

Table 8: **Pretraining sources extended:** The effect of different other pretraining sources. We use 4 frames per video when finetuning. Results are presented on the 1K-A MSR-VTT test set for text-video retrieval.

Method	Pre-training	#pairs	R@1	R@10	MedR
ClipBERT	COCO, VisGen	5.6M	22.0	59.9	6.0
Ours	COCO	0.6M	25.5	64.6	5.0

lengths and video durations can be found in Figure 4. Note that 275,000 videos are longer than 30 seconds, providing many examples of videos which can be used for training long-range video models.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv:2102.05095*, 2021. 3

[2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval, 2020. 2

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning, 2020. 2

[4] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2

[5] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching, 2021. 2

[6] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 1

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 2

[8] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1

[9] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching, 2018. 2

[10] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *arXiv preprint arXiv:2102.06183*, 2021. 3

[11] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[12] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *Proc. BMVC*, 2019. 1

[13] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv*, 2018. 1

[14] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 1

[15] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 1

[16] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 1

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. 2

[18] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[19] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 1