LAVA: Language Audio Vision Alignment for Data-Efficient Video Pre-Training

Sumanth Gurram [1] **Andy Fang** [1] **David Chan** [1] **John Canny** [1]

Abstract

Generating representations of video data is of key importance in advancing the field of machine perception. Most current techniques rely on hand-annotated data, which can be difficult to work with, expensive to generate, and hard to scale. In this work, we propose a novel learning approach based on contrastive learning, LAVA, which is capable of learning joint language, audio, and video representations in a self-supervised manner. We pre-train LAVA on the Kinetics 700 dataset using transformer encoders to learn representations for each modality. We then demonstrate that LAVA performs competitively with the current state-of-the-art self-supervised and weakly-supervised pre-training techniques on UCF-101 and HMDB-51 video action recognition while using a fraction of the unlabeled data.

1. Introduction

Supervised learning has generally driven the progress video representation learning, however, labeling datasets is both time-consuming and expensive, making it especially hard to leverage large amounts of data using supervised learning. Moreover, while attention-based architectures such as Dosovitskiy et al. (2020) Arnab et al. (2021); Bertasius et al. (2021) have started to outperform CNNs on key benchmarks, they often require much larger amounts of training data than CNNs.

Self-supervised methods have emerged to answer the challenge; as powerful pre-training strategies for vision tasks, they can scale to larger training datasets without being constrained by labeling needs. One common approach is to use data augmentation to learn representational invariants for vision Qian et al. (2021); Jing et al. (2018). Instead of relying on hand-designed augmentations for the visual modality, another approach is to exploit the multi-modal nature of video and learn audio-visual correspondence, as seen with contrastive methods such as those in Morgado et al. (2020b); Patrick et al. (2021; 2020a); Morgado et al. (2020a); Korbar et al. (2018). Similarly, other methods use text metadata from videos to learn joint visual-text representations in a self/weakly-supervised manner Stroud et al. (2020); Patrick et al. (2020b); Li & Wang (2020); Sun et al. (2019a); Miech et al. (2020); Sun et al. (2019b).

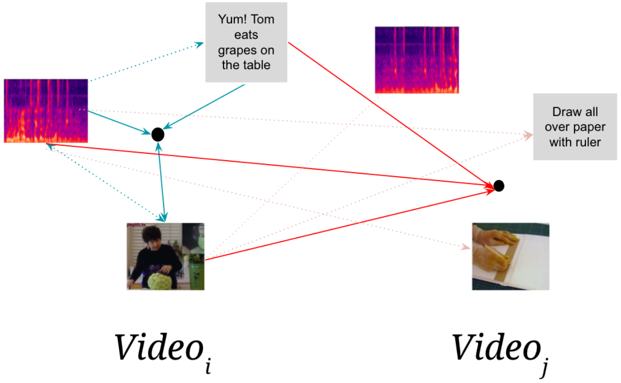
Yum! Tom eats grapes on the table
Draw all over paper with ruler
$Video_i$
$Video_j$

Figure 1. For a given sample, the video, audio, and text data are all encoded into embeddings. LAVA's pre-training objective involves contrasting video embeddings from one sample to audio and text embeddings from different samples (dotted red lines) while aligning embeddings from the same sample (dotted green lines). Additionally, LAVA will calculate a centroid from audio, video, and text embeddings for each sample and contrast (solid red lines) or align (solid green lines) embeddings to these centroids accordingly.

A less common but even more label-efficient strategy is to learn audio, visual and text representations together. While methods such as Akbari et al. (2021); Alayrac et al. (2020); Chen et al. (2021) pre-train on these three modalities, they do so leveraging HowTo100M, which has a massive 15 years of unlabeled video data. Moreover, excepting Akbari et al. (2021), most of the above techniques use highly modality-specific encoders (e.g. only CNNs for vision), instead of exploring more generic, attention-based backbones for all modalities.

These observations are the main motivation for LAVA, which introduces a more data and label-efficient method for pre-training transformer encoders on audio, visual, and

[1] University of California, Berkeley, USA. Correspondence to: Sumanth Gurram <sumanthgurram@berkeley.edu>, David Chan <davidchan@berkeley.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

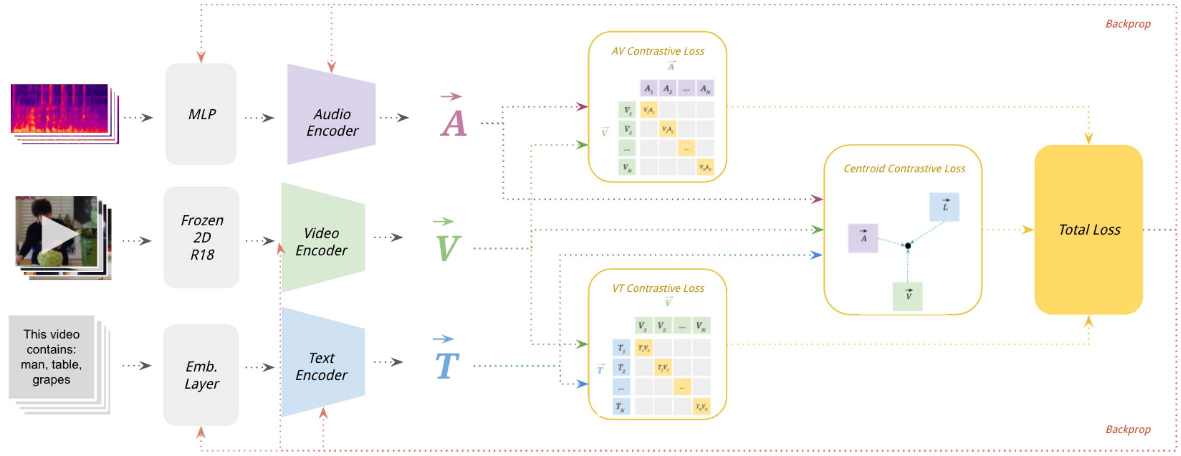

Figure 2. LAVA's pre-training architecture includes modality-specific feature extraction, attention-based encoding, cross-modal contrastive losses and centroid contrastive loss. Implementation details are in Section 3.

text modalities for video data. To achieve this, our proposed method uses novel cross-modal and centroid-based contrastive objectives seen in Figure 1. We evaluate pre-trained LAVA on UCF-101 and HMDB-51.

2. Pre-Training Approach

Overview Given a set of n unlabelled videos X, each video $x_i \in X$ is decomposed into different modalities a_i, v_i and t_i, which are audio, visual and text features, respectively. Details regarding modality-specific extraction are in Section 3. Then, LAVA's three encoders f_a, f_v, and f_t produce output embeddings given their respective modality features m_i as inputs. LAVA then uses projection functions to project these embeddings to various multi-modal latent spaces (e.g. audio-video, video-text and audio-video-text spaces) in some $\mathbb{R}^d$. Let us denote z_m as the embedding for modality m. We define projection function $g_{m,m'}$ to project z_m and $z_{m'}$, to a multi-modal latent space m, m'. In these multi-modal latent spaces, we apply a contrastive framework to jointly compare embeddings using a similarity function s, such that $\forall i, m' \neq m$ we have high $s(g_{m,m'}(z_{m,i}), g_{m,m'}(z_{m',i})$ and $\forall i \neq j, m' \neq m$ we have low $s(f_m(m_i), f_{m'}(m'_j))$.

Cross-Modal Contrastive Loss We use the function

$$s(\cdot, \cdot) = exp(g_{m,m'}(z_{m,i})^T g_{m,m'}(z_{m',j})/\tau)$$

as the similarity function, where τ denotes temperature, and the noise contrastive estimation (NCE) Gutmann & Hyvärinen (2010) as our contrastive loss, where positive pairs are embeddings from the same instance and negatives are embeddings from difference instances. This objective is intended to make LAVA a dictionary, effectively mapping input features a_i, v_i, t_i to unified representations z_i in a multi-modal latent space. The NCE loss is formulated below (for simplicity the projection functions are omitted):

$$NCE(z_m, z_{m'}) = -log(\frac{\sum_{i=0}^{N} exp(z_{m,i}^T z_{m',i}/\tau)}{\sum_{i=0}^{N} exp(z_{m,i}^T z_{m',i}/\tau)+\sum_{i=0}^{N}\sum_{j\neq i}^{N} exp(z_{m,i}^T z_{m',j}/\tau)}) \tag{1}$$

Following Miech et al. (2020), we use NCE for audio-video and video-text pairs:

$$L_{AV}(z_a, z_v) = NCE(g_{av}(z_a), g_{av}(z_v)) \tag{2}$$

$$L_{VT}(z_v, z_t) = NCE(g_{vt}(z_v), g_{vt}(z_t)) \tag{3}$$

Centroid Contrastive Loss LAVA also enforces audio-video-text correspondence via a novel centroid contrastive loss. Following Chen et al. (2021), for a given instance x_i, we calculate a centroid c_i by averaging projected LAVA embeddings: $c_i = (g_{avt}(z_{a,i}) + g_{avt}(z_{v,i}) + g_{avt}(z_{t,i}))/3$. g_{avt} projects modal embeddings to a joint, tri-modal latent space. However, unlike Chen et al. (2021) we do not k-means cluster the centroids via k-means clustering and align embeddings to their centroid's cluster assignment. Thus, we avoid potential detriments of grouping representations into a fixed K clusters across all batches and the additional training time needed for clustering. Instead, we instead directly optimize for alignment between embeddings and their centroid via the following loss:

$$L_{AVT}(c, z_a, z_v, z_t) = \sum_{m \in a,v,t} NCE(g_{avt}(z_m), c) \tag{4}$$

Combining all losses, we define the total loss for LAVA's pre-training:

$$L_{LAVA} = L_{AV} + L_{VT} + L_{AVT} \tag{5}$$

3. Experiments

Pre-training We use the training set of Kinetics-700: 480k videos of which 300k videos have usable audio and

Method	Dataset (years)	GPU Hours	Mod.	UCF	HMDB
RotNet3D Jing et al. (2018)	K600 (0.1)	-	V	47.7	24.8
CBT Sun et al. (2019a)	K600 (0.1)	1536	VT	54.0	29.5
MemDPC Han et al. (2020a)	K600 (0.1)	-	VF	54.1	30.5
AVSF Xiao et al. (2020)	K400 (0.1)	-	AV	54.1	30.5
CoCLR Han et al. (2020b)	K400 (0.1)	-	VF	77.4	44.1
STiCA* Patrick et al. (2021)	K400 (0.1)	2930	AV	77.0	48.2
CPD Li & Wang (2020)	IG300 (0.1)	-	VT	83.7	54.7
CVRL Qian et al. (2021)	K600 (0.1)	-	V	90.8	59.7
LAVA* (video only)	K700 (0.1)	408	AVT	81.3	50.1
LAVA* (audio+video)	K700 (0.1)	408	AVT	84.3	-
MIL-NCE Miech et al. (2020)	HT (15)	4608	VT	83.4	54.8
ELo Piergiovanni et al. (2020)	Y8M	4608	V	-	64.5
VATT* Akbari et al. (2021)	HT (15)	18432	AVT	89.6	65.2
BraVe Recasens et al. (2021)	AS (1)	-	AV	93.6	70.8
MMV† Alayrac et al. (2020)	AS+HT (16)	2304	AVT	95.2	75.0
WTS Stroud et al. (2020)	WTS70M (22)	9216	VT	95.8	77.7

Table 1. * denotes that the vision backbone has a transformer encoder. † denotes that the vision backbone is fine-tuned downstream, rather than being frozen for linear evaluation. The duration of each dataset is also measured in years.

web-crawled titles. To the best of our knowledge, LAVA is the first method to extract this text data for Kinetics-700 and we intend to release this data for others to use. Video clips are [16 x 224 x 224 x 3] at 10 fps, with random h-flip and space-crop augmentations. Audio features are [80, 256] log mel-spectrograms augmented with Gaussian noise. Text sequences have a maximum length of 128 from a 48k vocab size using BPE tokenization. Most video transformers use ImageNet-pre-trained ViT Arnab et al. (2021); Bertasius et al. (2021), pre-train on massive video datasets like HowTo100M as in Akbari et al. (2021) or use convolutions followed transformers as in Patrick et al. (2021). We follow the latter, except we freeze a ImageNet-pre-trained ResNet-18 backbone and use it to extract frame-wise patch features for each clip. Audio features are encoded by an MLP. Text tokens are mapped to embeddings. As seen in Figure 2, audio, video and text features are then encoded by transformer encoders f_a, f_v, and f_t, each with 4 layers and a hidden size of 1024. Missing modalities in a batch are masked out during loss calculation. We use a batch size of 32, 0.07 temperature, $1e^{-5}$ learning rate, Adam optimizer, and a cosine-learning schedule. Pre-training is done for 25 epochs on a single Titan X GPU.

UCF-101 Downstream The UCF-101 dataset has 13k videos across 101 action categories. We train a linear classifier on top of the frozen LAVA video encoder using a 1e-4 learning rate and a batch size of 32. Also, we compare the downstream performance of the classifier when using the frozen video encoder vs. frozen audio and video encoders. Logits are averaged across multiple clips per test set video. Top-1 accuracy is averaged across all 3 splits.

Pre-train	Downstream	UCF-101	HMDB-51
AV	Video Only	68.6	41.5
AV	Audio+Video	72.4	-
AV+VT	Video Only	74.91	49.4
AV+VT	Audio+Video	79.36	-
AV+VT+AVT	Video Only	81.1	51.8
AV+VT+AVT	Audio+Video	84.2	-

Table 2. All results are split-1 top-1 accuracy. Pre-train denotes which pre-training losses are used.

HMDB-51 Downstream The HMDB-51 dataset has 7k videos across 51 action categories. We follow the UCF-101 evaluation procedure, except we do not evaluate with audio+video mode as most HMDB-51 videos have no audio. Otherwise, our procedure follows UCF-101.

4. Results

Downstream Tasks Table 1 compares LAVA linear evaluation performance on UCF-101 and HMDB-51 to various self-supervised benchmarks. LAVA outperforms all methods trained on datasets comparable to Kinetics-700 size besides Li & Wang (2020); Qian et al. (2021). This can be attributed to LAVA's increased data efficiency as it makes use of audio, video, and text for each sample, whereas these benchmarks rely on at most two modalities. Notably, LAVA significantly outperforms Sun et al. (2019a), which omits audio and uses ASR-extracted captions as text, whereas LAVA is pre-trained on audio and video titles. LAVA also slightly outperforms Patrick et al. (2021), which has a R(2+1)-18+transformer video encoder and ResNet audio encoder, but does not pre-train on text. LAVA performs competitively with Li & Wang (2020), which also uses video titles as text, and is outperformed by Qian et al. (2021), which uses the video modality using extensive spatio-temporal augmentations and a large batch size of 1024. However, both methods Li & Wang (2020); Qian et al. (2021) use ResNet3D-50 as their backbone, which is more performant when trained from scratch than transformer encoders trained with the same amount of data Bertasius et al. (2021). See Section 5 for plans to directly compare for to these methods.

Interestingly, Miech et al. (2020), despite having been pre-trained on over 150X samples than Kinetics-700, only slightly outperforms LAVA on UCF-101 and HMDB-51. We believe this is because Miech et al. (2020) uses caption-based text and does not use audio, whereas LAVA is pre-trained on audio and titles-based text. Akbari et al. (2021); Alayrac et al. (2020) outperform LAVA using audio, video and caption-based text, while Stroud et al. (2020) uses video and titles-based text. While this performance gap can likely be attributed to LAVA's reliance on less than 1% of the data used in these benchmarks, it does seem using audio and titles-based text can increase data efficiency as LAVA's performance rivals that of Miech et al. (2020).

Given that LAVA uses Kinetics-700 video titles as text, it can be argued that this is a form of weak supervision as per Li & Wang (2020); Stroud et al. (2020). We found that while all HMDB-51 classes are fully covered by Kinetics-700 text, the following UCF-101 classes are not covered: IceDancing, PizzaTossing, PommelHorse, SkiJet, StillRings. The 3-split average top-1 accuracy for these classes is 79.9%, which is comparable to LAVA's overall 81.3% for UCF-101, indicating that LAVA's representations transfer well to unseen classes downstream.

Since LAVA is a multi-modal model, we use our novel evaluation strategy to quantify downstream performance improvements when using embeddings from audio and video modalities. Table 1 shows that audio-video LAVA outperforms video-only by 3%, indicating that audio embeddings may provide additional information for action recognition. See Section 5 for more plans in this direction.

Lastly, pre-training LAVA takes 408 GPU hours, which is over 70% faster than all other benchmarks documented in Table 1. Notably, LAVA also completes all pre-training using just 1 GPU, whereas all other benchmarks use few as 4 and as many as 256 accelerators. This combination of 70% fewer accelerators and 70% fewer GPU hours means LAVA is significantly more efficient in terms of computational cost, in addition to label and data-efficiency.

Ablations As shown in Table 3, we ablate pre-training duration by evaluating LAVA at 1, 10, and 25 epochs of pre-training; video-only increases from 73.1% to 80.0% to 81.3%, while audio+video increases from 76.0% to 82.3% to 84.3%. Additionally, we ablate pre-training losses as seen in Table 2. As expected, LAVA pre-trained without text and centroid contrastive loss performs the worst. LAVA pre-trained with L_{AV} and L_{VT}, but no centroid loss performs worse than LAVA pre-trained with the centroid loss. This indicates that the text modality and the centroid loss, both of which increase data efficiency by making greater use of the same number of samples, significantly improve downstream video action recognition. The performance boost from using audio+video embeddings during linear evaluation is also seen in Table 2.

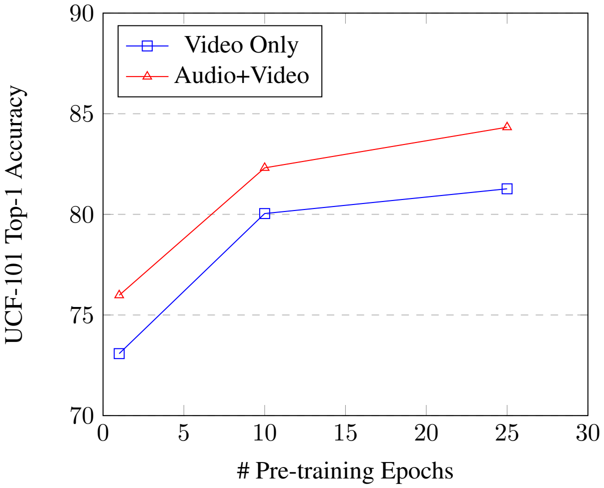

Figure 3. UCF-101 accuracy as a function of pre-training duration.

5. Conclusion

We present LAVA, a novel self-supervised pre-training method for learning audio, visual, and text representations from unlabeled video data using transformers. By using multiple modalities and introducing novel cross-modal and centroid contrastive objectives, LAVA increases data efficiency while performing competitively with self and weakly supervised benchmarks. As LAVA's pre-training effectively combines multi-modal data and generic transformer encoders, we believe it can scale both in terms of the amount on unlabelled data and the number of modalities in the data.

Future Work To more directly compare LAVA performance to Li & Wang (2020); Qian et al. (2021), we plan to pre-train the transformer vision encoder on ImageNet first or use the ResNet3D-50 encoder from scratch, as well as larger batch sizes for contrastive pre-training and more extensive video augmentation. We also plan to pre-train LAVA on larger datasets with non-title-based text (e.g. HowTo100M) to compare more directly with Akbari et al. (2021); Alayrac et al. (2020). Since contrastive pre-training with audio may dilute information in the video embeddings, we plan to compare downstream performance between video embeddings from VT-pretrained LAVA and those of current LAVA. Additionally, we plan to ablate how different coefficients between various LAVA losses will affect downstream performance. Lastly, we intend to explore pre-training on more/different modalities and new downstream tasks (e.g. retrieval, captioning, video/audio generation) using LAVA.

References

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., and Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., and Zisserman, A. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, 2(3):4, 2021.

Chen, B., Rouditchenko, A., Duarte, K., Kuehne, H., Thomas, S., Boggust, A., Panda, R., Kingsbury, B., Feris, R., Harwath, D., et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8012–8021, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/gutmann10a.html.

Han, T., Xie, W., and Zisserman, A. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pp. 312–329. Springer, 2020a.

Han, T., Xie, W., and Zisserman, A. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020b.

Jing, L., Yang, X., Liu, J., and Tian, Y. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018.

Korbar, B., Tran, D., and Torresani, L. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

Li, T. and Wang, L. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*, 2020.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., and Zisserman, A. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9879–9889, 2020.

Morgado, P., Li, Y., and Nvasconcelos, N. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33: 4733–4744, 2020a.

Morgado, P., Vasconcelos, N., and Misra, I. Audio-visual instance discrimination with cross-modal agreement. *arXiv preprint arXiv:2004.12943*, 2020b.

Patrick, M., Asano, Y. M., Kuznetsova, P., Fong, R., Henriques, J. F., Zweig, G., and Vedaldi, A. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020a.

Patrick, M., Huang, P.-Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., and Vedaldi, A. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020b.

Patrick, M., Huang, P.-Y., Misra, I., Metze, F., Vedaldi, A., Asano, Y. M., and Henriques, J. F. Space-time crop & attend: Improving cross-modal video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10560–10572, 2021.

Piergiovanni, A., Angelova, A., and Ryoo, M. S. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 133–142, 2020.

Qian, R., Meng, T., Gong, B., Yang, M.-H., Wang, H., Belongie, S., and Cui, Y. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6964–6974, 2021.

Recasens, A., Luc, P., Alayrac, J.-B., Wang, L., Strub, F., Tallec, C., Malinowski, M., Pătrăucean, V., Altché, F., Valko, M., et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1255–1265, 2021.

Stroud, J. C., Lu, Z., Sun, C., Deng, J., Sukthankar, R., Schmid, C., and Ross, D. A. Learning video representations from textual web supervision. *arXiv preprint arXiv:2007.14937*, 2020.

Sun, C., Baradel, F., Murphy, K., and Schmid, C. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019a.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7464–7473, 2019b.

Xiao, F., Lee, Y. J., Grauman, K., Malik, J., and Feichtenhofer, C. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.