On Negative Sampling for Audio-Visual Contrastive Learning from Movies

Mahdi M. Kalayeh Shervin Ardeshir Lingyi Liu* Nagendra Kamath*
Netflix Research
{mkalayeh,shervina,nkamath,lliu}@netflix.com

Ashok Chandrashekar†
WarnerMedia
ashok.chandrashekar@warnermedia.com

1. Introduction

This work aims at exploring audio-visual self-supervised learning from movies that are long-form and uncurated, while examining its differences with the more prevalent scenario of learning from short and curated videos. Datasets used for self-supervised video representation learning often consist of *large* catalogues of *short* videos [5, 7]. The short length of the videos, and the shear number of them has led to an underlying i.i.d assumption of data distribution, based on which many of the prior works have been developed. With the data of such nature, there is often an implicit assumption of within-video semantic consistency [4, 12], which is intuitive as the likelihood of a short video containing a single semantic concept or at least very coherent ones is relatively high. On that basis, prior works [4, 12] treat different clips of a given video as augmentations of the same semantic concept. Hence, minimizing a contrastive objective is set to encourage two clips that are sampled from the same video to become more similar in the latent embedding space, while repelling pairs where clips come from two different video instances. Here we argue that such an assumption is not universal, and in fact is sub-optimal when learning from long-form content like movies. In the following, we identify three main characteristics for clips derived from a collection of long-form contents.

Semantic Diversity. Long-form content often contains a diverse set of semantic concepts, e.g. characters, actions, and environments. Thus, unlike the short-video regime, random clips from the a long-form source are very likely to be semantically dissimilar. This characteristic encourages within-content negative sampling, shown in the Figure 1.

Non-Semantic Consistency. Movies usually have underlying attributes such as color palettes, thematic background music, and other artistic patterns, some also as a result of post production. These *artifacts*, often consistent throughout the content, are independent of the audiovisual semantics that are being depicted. We argue that considering all clips of a long-form video to be semantically correspondent, as it is practiced in the prior works like [4, 12], could lead to the model relying on such irrelevant artifacts and ignoring the semantics of the content.

Reoccurring Concepts. Concepts such as environments, and characters, often re-appear with minute variations throughout a movie. Thus, even though random clips of the same long-form content are likely semantically dissimilar, the possibility of a semantic correspondence even between temporally distant clips still exists. This can theoretically lead to the *class-collision* phenomenon which is naturally the price of negative sampling on unlabeled data. However, due to *semantic diversity*, its likelihood is relatively low, as random pairs are more likely to represent different concepts, as semantic diversity grows.

The characteristics mentioned above, suggest exploring within-content negative sampling, with the possibility of diminishing returns past a certain level of emphasis. We explore such hypothesis, and experiment with the extent to which negative sampling could be helpful in this context.

2. Approach

Notations and Architecture. Our pretraining dataset is denoted by $\mathcal{X} = \{\mathcal{X}_n | n \in [1 \cdots N]\}$, where $\mathcal{X}_n = \{x_{n,m} | m \in [1 \cdots M_n]\}$ contains M_n non-overlapping audiovisual snippets which are temporally segmented from the duration of the n^{th} long-form content (movie) in the dataset. Each snippet includes both audio and video modalities, formally $x_{n,m} = (a_{n,m}, v_{n,m})$. Video and audio are processed through 18-layers deep R(2+1)D and ResNet architectures, respectively referred to as $f : \mathbb{R}^3 \rightarrow \mathbb{R}^{d_f}$ and $g : \mathbb{R}^1 \rightarrow \mathbb{R}^{d_g}$. We use *projection heads*, $h_f : \mathbb{R}^{d_f} \rightarrow \mathbb{R}^d$ and $h_g : \mathbb{R}^{d_g} \rightarrow \mathbb{R}^d$, to map corresponding representations into a common d-dimensional space before computing the

*Equal Contribution.
†The author was with Netflix when this work started.

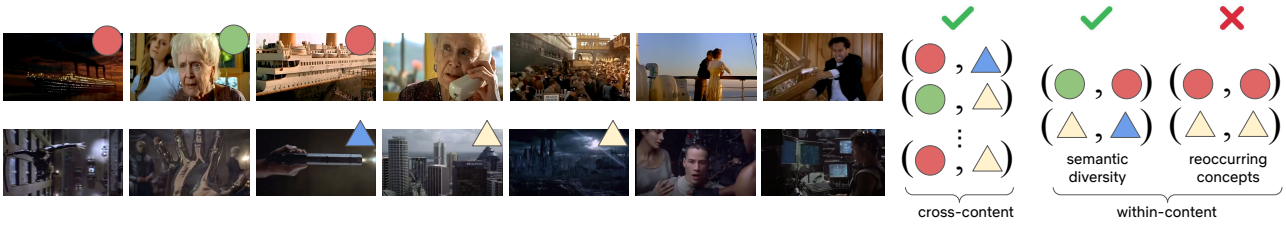

Figure 1. The non-i.i.d nature of data distribution in long-form content: Left shows sampled frames from two different movies, one in each row, where we can observe *non-semantic consistency*, in form of color pallet driven artifacts (bottom row is generally darker), across different clips of each movie. Due to *semantic diversity*, negative sampling from within a movie is safe as it would mostly result in semantically non-correspondent pairs of clips. However, because of *reoccurring semantic concepts*, there still exists a possibility of constructing a negative pair from semantically similar clips which will not be ideal.

contrastive loss. We discard the projection heads and use f and g for transfer learning on respective downstream tasks.

Loss Function. With a slight abuse of notation , $\mathcal{B} = \{x_i = (a_i, v_i) | i \in [1 \cdots B]\}$ represents a minibatch of size B, where video and audio modalities associated with the i^{th} instance, x_i, are denoted by v_i and a_i. We use $z_v^i = h_f(f(v_i))$ and $z_a^i = h_g(g(a_i))$ to represent the associated embeddings generated by projection heads, and optimize the noise contrastive estimation (NCE) loss shown in Equation 1 in order to maximize the symmetric joint probability between a corresponding audio and video. For the i^{th} element in the minibatch, (z_v^i, z_a^i) serves as the positive pair, while assuming negative pairs for both modalities, $\mathcal{N}_i = \{(z_v^i, z_a^j), (z_v^j, z_a^i) | j \in [1 \cdots B], i \neq j\}$ constitutes the set of negative pairs.

$$\mathcal{L} = -\sum_{i=1}^{B} \log \left(\frac{e^{(z_v^i)^\intercal (z_a^i)}}{e^{(z_v^i)^\intercal (z_a^i)} + \sum\limits_{(z_v', z_a') \in \mathcal{N}_i} e^{(z_v')^\intercal (z_a')}} \right) \quad (1)$$

Sampling Policy. Equation 1, is computed over B training instances, each in the form of an audiovisual snippet. A naive sampling policy may ignore the fact that snippets comprising the pretraining dataset are in fact temporal segments trimmed from longer-form contents, *i.e.* movies. Such an assumption treats our training data as independent and identically distributed random variables from $\bigcup_{n=1}^{N} \mathcal{X}_n$, which constitutes the default sampling policy that is commonly used in the general deep learning literature. However, as detailed in Section 1, the underlying *artifacts* (within-content *non-semantic consistency*), in addition to commonalities and correlations along the temporal axis of a long-form content (*reoccurring semantic concepts*), contribute to breaking the previously discussed i.i.d assumption on the training data. We hypothesize that during training, model gradually discovers previously mentioned content-exclusive artifacts, and latches onto those to quickly minimize Equation 1 leading to sub-optimal generalization.

The reason being $B \ll N$, hence for $n \sim \mathbb{U}(1, N)$ and $m \neq m'$, $\mathsf{P}(x_{n,m} \in \mathcal{B} \wedge x_{n,m'} \in \mathcal{B})$ is very low. In other words, if a naive random sampling policy is adopted, the set of negative pairs in Equation 1 would mainly include audio-video pairs from two different movies. As shown in Figure 1, this results in easy cross-content negatives. To prevent the model's reliance on artifacts, our approach emphasizes the within-content negative pairs by dividing the minibatch budget of B, across B/k randomly chosen long-form contents, where we sample k snippets from each. It is worth reiterating that the prior works [4, 12] encourage temporally distant segments of the same video to be similar (positive pair) in the latent embedding space. In contrast, we treat such instances as a negative pair and aim for the optimization to push them apart from one another. We first uniformly sample a long-form content, $n \sim \mathbb{U}(1, N)$ and then draw k distinct snippets from $\mathcal{X}_n$, creating $\{x_{n,m} | m \in \mathcal{M}_n\}$, where $\mathcal{M}_n \subset [1 \cdots M_n]$ and $|\mathcal{M}_n| = k$. This ensures that for $x_i \in \mathcal{B}$, $\mathcal{N}_i$ always includes $2k - 2$ pairs sampled from the same movie to which x_i belongs. By putting constraints on $\mathcal{M}_n$, specifically how temporally far from each other the k samples are drawn, we may go one step further and to some extent control the audiovisual similarity between snippets. This serves as an additional knob to tune for hard negative sampling, as temporally nearby snippets tend to share more commonalities.

$k \leq \texttt{max}[\mathcal{M}_n] - \texttt{min}[\mathcal{M}_n] + 1 \leq w \leq M_n$ defines the bounds on our sampling policy, where w, standing for a sampling *window*, determines the farthest two out of k samples drawn from $\mathcal{X}_n$ can be. Accordingly, $w = k$ represents the case where all k samples are temporally adjacent, hence the expected audiovisual similarity is maximized due to temporal continuity in content. In our preliminary studies, we observed that having such level of hard negatives, even with a small k, prevents proper training and results in performance degradation. On the other hand, $w = M_n$ indicates random sampling where no temporal constraint is imposed on $\mathcal{M}_n$, thus samples are less likely to be drawn from adjacent time-stamps. The rest of the spectrum pro-

vides middle grounds where two samples drawn from $\mathcal{X}_n$ can at most be $w+1$ snippets apart.

3. Experiments

Downstream Evaluation. We follow recent works [1, 2, 9, 10, 13] and perform transfer learning on UCF101 [14] and HMDB51 [8] for action recognition, along with ESC50 [11] for audio classification. We further evaluate our models on datasets which are larger in scale, namely Kinetics-400 [7] and VGGSound [3]. For implementation details please refer to the extended version [6]. Illustrated in figure 2, we observed that increasing k beyond 1 results in harder pretraining objectives as more within-content negative samples are contributing to the denominator of the Equation 1. Meanwhile, increasing the difficulty of the self-supervised pretext task is leading to better downstream performance, in a linear evaluation regime on different downstream tasks. With an effective batch size of 96, spanning k across the full spectrum allows us to study how self-supervised pretraining is influenced by different amounts of video-level diversity.

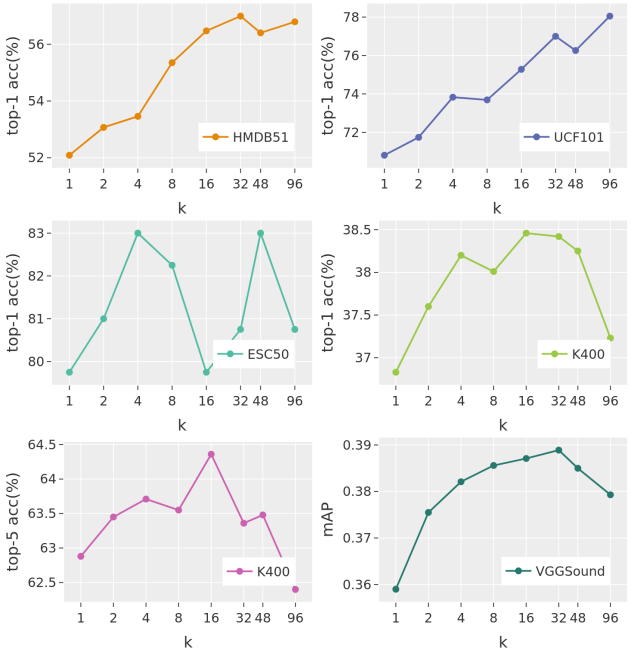

Figure 2. Effect of emphasizing on within-content negative sampling through increasing k during pretraining. Downstream transfer learning performances are measured in a linear evaluation regime. For HMDB51, UCF101 and ESC50, numbers are reported on the split-1 of the corresponding datasets. For Kinetics-400 (K400) and VGGsound, we use their validation sets. Sampling window (w) is set to 4 times as k. By emphasis on within-content negative sampling, compared to $k=1$ baseline, we achieve gains of **4.90%** on HMDB51, **7.24%** on UCF101, **3.25%** on ESC50, **1.63%** on Kinetics-400, and **0.04** mAP on VGG-Sound.

method	pretraining	HMDB51	UCF101	ESC50
protocol: linear evaluation on split-1				
XDC [2]	IG-Random	49.9	80.7	84.5
XDC [2]	IG-Kinetics	56.0	85.3	84.3
ours	movies	63.5	79.8	82.5
protocol: finetuning on split-1				
XDC [2]	IG-Random	61.2	88.8	86.3
XDC [2]	IG-Kinetics	63.1	91.5	84.8
ours	movies	73.0	89.7	88.7

Table 1. Learning from movies as a source of semantically *uncurated* pretraining data.

Pretraining on Uncurated Data: To the best of our knowledge, the only *uncurated* dataset used in literature for audio-visual self-supervised learning is IG-Random [2] which has 65M training videos. It is an *uncurated* version of the weakly-supervised collected IG-Kinetics [5] where videos were retrieved by tags relevant to the categories in the Kinetics dataset [7]. Alwassel *et.al* [2] argue that self-supervised pretraining on likes of IG-Kinetics [5], and other supervised datasets for that matter, introduces additional privileges since even without using labels, training videos are still biased due to the sampling distribution (*e.g.*, taxonomy of the curated dataset). In this work, from a large catalogue of movies, we've randomly selected ∼3.6K films, an equivalent of 0.7 years worth of content (30 times smaller than IG-Random [2]), as our pretraining dataset, which as discussed in Section 1, like IG-Random [2] is semantically *uncurated*. Table 1 compares our approach against XDC [2]. In the finetuning regime, our model trained on a collection of movies consistently outperforms XDC [2] with a large margin across three different tasks. Although, the gap reduces in the linear evaluation protocol.

Comparison with the state-of-the-art: Table 2 compares our proposed approach against the best performing audiovisual self-supervised learning methods. For fairness, we included specifics of architectures and pretraining datasets used in each method. In general, we achieve very competitive results on HMDB51 [8], however on UCF101 [14], our numbers do fall behind. Please note the pretraining datasets used by other methods are all curated, with the exception of IG-Random [2], and are often significantly larger than our pretraining data. On ESC50 [11] and Kinetics-400 [7], we achieve comparable results with state-of-the-art. For instance, on Kinetics-400 [7], while using the same backbone architecture, our model performs on par with AVID [9] despite it has been pretrained on the same Kinetics-400 [7] dataset. Finally, we did experiment with VGGSound [3] and obtained **0.38** and **0.48 mAP**, respectively in linear and finetuning evaluation regimes. For a more thorough com-

protocol: finetuning				
Method	Arch.	Data	HMDB51	UCF101
GDT [10]	R(2+1)D-18	IG-K	72.8	95.2
AVID [9]	R(2+1)D-18	AS	64.7	91.5
XDC [2]	R(2+1)D-18	IG-K	68.9	95.5
MMV [1]	R(2+1)D-18	AS	70.1	91.5
CVRL [12]	R3D-50	K400	66.7	92.2
BraVe [13]	TSM-50	AS	75.3	95.6
protocol: linear evaluation				
MMV [1]	R(2+1)D-18	AS	60.0	83.9
CVRL [12]	R3D-50	K400	57.3	89.2
BraVe [13]	TSM-50	AS	69.1	93.4
Ours	R(2+1)D-18	Movies	64.6	80.8
protocol: linear Evaluation				
Method	Arch.	Data	ESC50	K400
AVID [9]	R(2+1)D-18	K400	79.1	44.5
XDC [2]	R(2+1)D-18	K400	78.5	–
AVID [9]	R(2+1)D-18	AS	89.1	48.9
MMV [1]	R(2+1)D-18	AS	85.6	–
CVRL [12]	R3D-50	K400	–	66.1
BraVe [13]	TSM-50	AS	92.1	–
Ours	R(2+1)D-18	Movies	83.6	43.6

Table 2. Comparison with the state-of-the-art. Column "Data" indicates the pretraining dataset with abbreviations as follows: **K**inetics-**400**, **A**udio**S**et, **Y**outube-**8M**, **IG-K**inetics, and **IG-R**andom. For HMDB51, UCF101, and ESC50, we report the average results on all the folds.

parison please refer to [6].

4. Conclusion

We studied self-supervised pretraining on semantically uncurated long-form content (*i.e.* movies). We identified characteristics specific to movies, and explored how within-content negative sampling harnesses them to improve representation learning. Our experiments show that pretraining on such data, even at a comparatively smaller scale to the curated and supervised alternatives, can give rise to representations capable of competing with the state-of-the-art.

References

[1] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*, 2020. 3, 4

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems*, volume 33, pages 9758–9770, 2020. 3, 4

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 3

[4] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 1, 2

[5] 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. 1, 3

[6] Mahdi M Kalayeh, Shervin Ardeshir, Lingyi Liu, Nagendra Kamath, and Ashok Chandrashekar. On negative sampling for audio-visual contrastive learning from movies. *arXiv preprint arXiv:2205.00073*, 2022. 3, 4

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 3

[8] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 3

[9] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *arXiv preprint arXiv:2004.12943*, 2020. 3, 4

[10] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9587, 2021. 3, 4

[11] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 3

[12] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 1, 2, 4

[13] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, Jean-Bastien Grill, Aäron van den Oord, and Andrew Zisserman. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1255–1265, October 2021. 3, 4

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3