Learnable Irrelevant Modality Dropout for Multimodal Action Recognition on Modality-Specific Annotated Videos
(*Supplementary Material*)

1. Semantic Audio-Video Label Dictionary (SAVLD)

For building the SAVLD, we use ***BERT-base uncased*** in which the label text is lowercased before tokenizing it. Therefore, as a preprocessing step, we first normalized all textual labels into lowercase form. Notably, the resulting label dictionary SAVLD is not very accurate since the semantic gap between video and audio datasets is still large. It is also because the number of classes on video/audio datasets is still considered small (*i.e*. Kinetics400 has 400 labels and AudioSet has 527 label). However, using SAVLD dictionaries, our framework narrows the input noise by using the audio predictions. Then, it matches the audio scene to its closest similar visual class regardless of the fact that many audio classes do not have large relevance. In the training phase, the labeling process is guided by applying the IOU function between the AST predictions and the corresponding audio labels to the video label in the SAVLD dictionary. Table 3 shows a small part of the Kinetics400-AudioSet label dictionary when $k = 5$. Additionally, we visualize the AudioSet labels' relevance to Kinetics400 and UCF-101 labels in Fig. 2 and Fig. 4, respectively. The most frequent mapped audio labels to the vision-specific dataset labels are shown in Fig. 1 and Fig. 3 for Kinetics400 and UCF-101 datasets, respectively.[1]

2. Further Results and Analysis

2.1. Experiments on HMDB51 and Kinetics-Sounds

We further evaluate our method against relevant methods for video-based action recognition in two more datasets HMDB51 [6] and Kinetics-Sounds [8]. **HMDB51** [6] is split into three overlapped splits. It contains $6,766$ videos of 51 classes with an average length of 3 seconds. **Kinetics-Sounds** [1] is a subset from the original Kinetics400 [2] dataset. It contains 34 classes in which each class videos have a remarkable sound signature. Since Kinetics dataset editions are downloadable from YouTube, its size may vary by time as some videos may get removed. Herein, we use $19,627$ videos for training and $1,344$ videos for evaluation. length of 3 seconds. The average video length in this dataset is 10 seconds.

Table 1. Performance comparison to relevant visual-based methods (RGB + Optical Flow) on HMDB51 dataset.

Model	Top-1 (%)
CoViAR [10]	59.1
Two-stream fusion [5]	65.4
TSN [9]	69.4
I3D [2]	66.4
CoViAR + OF [10]	70.2
IMD-B (ours)	**71.3**

Table 2. Performance comparison to relevant methods on Kinetics-Sounds dataset.

Model	P-train	Top-1 (%)
L3-Net [1]	IN-1K	74.0
SlowFast R101 [4]	IN-1K	77.9
AVSlowFast, R101 [11]	IN-1K	85.0
MBT (AV) [7]	IN-21K	85.0
IMD-B (ours)	IN-21K	**90.48**
IMD-B* (ours)	IN-21K	**91.44**

Table 1 reports the performance of several visual action recognition methods including CoViAR [10], Two-stream fusion [5], TSN [9], I3D [2], and CoViAR + OF [10]. Notably, our method provides a slight performance boost on HMDB51 because our method is a Transformer-based framework that shows better improvement on large datasets

[1] We implemented our framework in Pytorch in which we borrowed several parts from the following codebases:
https://github.com/huggingface/transformers
https://github.com/YuanGongND/ast
https://github.com/facebookresearch/SlowFast
https://github.com/facebookresearch/TimeSformer
https://github.com/rwightman/pytorch-image-models
https://github.com/unixpickle/audioset
https://github.com/ekazakos/temporal-binding-network
https://github.com/marl/l3embedding
https://github.com/johnarevalo/gmu-mmimdb

Table 3. Samples of the most relevant AudioSet labels to video labels retrieved by semantic sentence-based embeddings mapping by BERT when $K = 5$.

Dataset	Label	Relevant AudioSet Labels
Kinetics400	playing guitar	bass guitar;acoustic guitar;guitar;chopping food;electric guitar
	applauding	speech;applause;whistling;chime;clapping
	belly dancing	rapping;yodeling;synthetic singing;child singing;frying food
	canoeing or kayaking	rowboat, canoe, kayak;motorboat, speedboat;skateboard;folk music;sailboat, sailing ship
	clean and jerk	fill with liquid;pump liquid;filing rasp;rumble;rustle
	country line dancing	female singing;dance music;male singing;salsa music;drum roll
	driving car	emergency vehicle;motor vehicle road;filing rasp;car;engine starting
	feeding birds	wild animals;insect;mosquito;bird;patter
	gargling	gargling;gurgling;snoring;reversing beeps;yodeling
	kissing	whispering;typing;cheering;laughter;breathing
	pumping gas	frying food;pump liquid;sawing;sanding;filing rasp
	recording music	vocal music;music;soundtrack music;wedding music;jingle music
	scrambling eggs	singing bowl;spray;thunder;wheeze;tools
	sniffing	whimper;growling;cheering;whispering;rattle
	sneezing	gurgling;snoring;babbling;gargling;rapping
	tickling	whispering;rustle;cheering;growling;screaming
	yawning	babbling;rapping;frying food;gurgling;snoring
	writing	writing;speech;typing;chatter;mechanisms
	whistling	whistling;humming;whistle;whip;siren
	welding	gears;scissors;drill;boiling;bicycle

in terms of number of videos per class. However, our framework provides top-1 of 71.3% which is better compared with CoViAR + OF [10] with a performance boost of $\sim$1.1%. We compare our method with several methods on the visual-audio annotated dataset Kinetics-Sounds. This dataset was first used in [1] as a subset of the main Kinetics400 dataset [2]. Our framework provides a significant boost in this dataset, where it provides top-1 of 90.48% and 91.44% with and without intra-class cross-modality augmentation, respectively. Since this dataset is an audio-video annotated in which audio and video are mostly relevant in each video, the cross-modality augmentation does not improve the performance. This interprets our finding regarding this augmentation method, where it provides most performance boosts on datasets with low audio-video relevance. Therefore, in our method, cross-modality augmentation is particularly applied for improving the video-based human activity recognition on vision-specific videos.

2.2. Visual Two-Stream Transformer Variants

In this part, we report the performance of three Transformer variants of the proposed two-stream visual Transformer. In order to leverage the pretrained ImageNet knowledge, we have adopted several parts from ViT [3] in terms of number of Transformer encoder blocks, embedding size, number of self-attention heads, input dimensions. Our Transformer scales properties are almost similar to the [3], *i.e. small*, *base*, and *large*, except the spatiotemporal encoder blocks. We used one spatiotemporal block on the *small* encoder instance as it involves 8 blocks, whereas we add 2 and 4 spatiotemporal blocks for *base* and *large* instances as they involve 12 and 24 blocks, respectively. Table 4 reports the recognition performance on Kinetics400 dataset as well as the Transformer instances' costs in terms of number of parameters and GFLOPs.

Table 4. Performance of the proposed visual two-stream Transformer with different size. Three Transformer instances are trained, where each of which is initialized with its ViT corresponding ImageNet-21K weights. The performance is reported on Kinetics400. Number of parameters is reported in million.

T. Size	Top-1 (%)	Top-5 (%)	Params	GFLOPs
Small	78.8	92.8	47.9×2	$2,871$
Base	81.1	94.3	88.6×2	$4,464$
Large	82.6	95.2	173.7×2	$8,232$

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, Listen and Learn. In *ICCV*, volume Octob, pages 609–617, oct 2017.

[2] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In

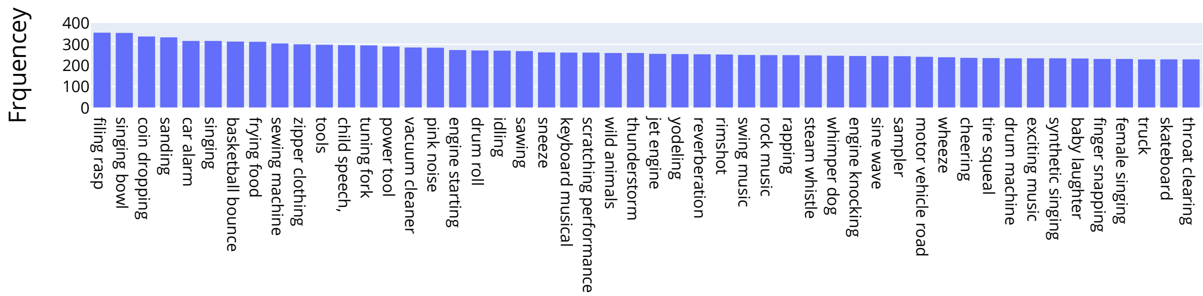

Figure 1. The most 50 frequent audio labels in AudioSet mapped to Kinetics400 labels when $k = 50$.

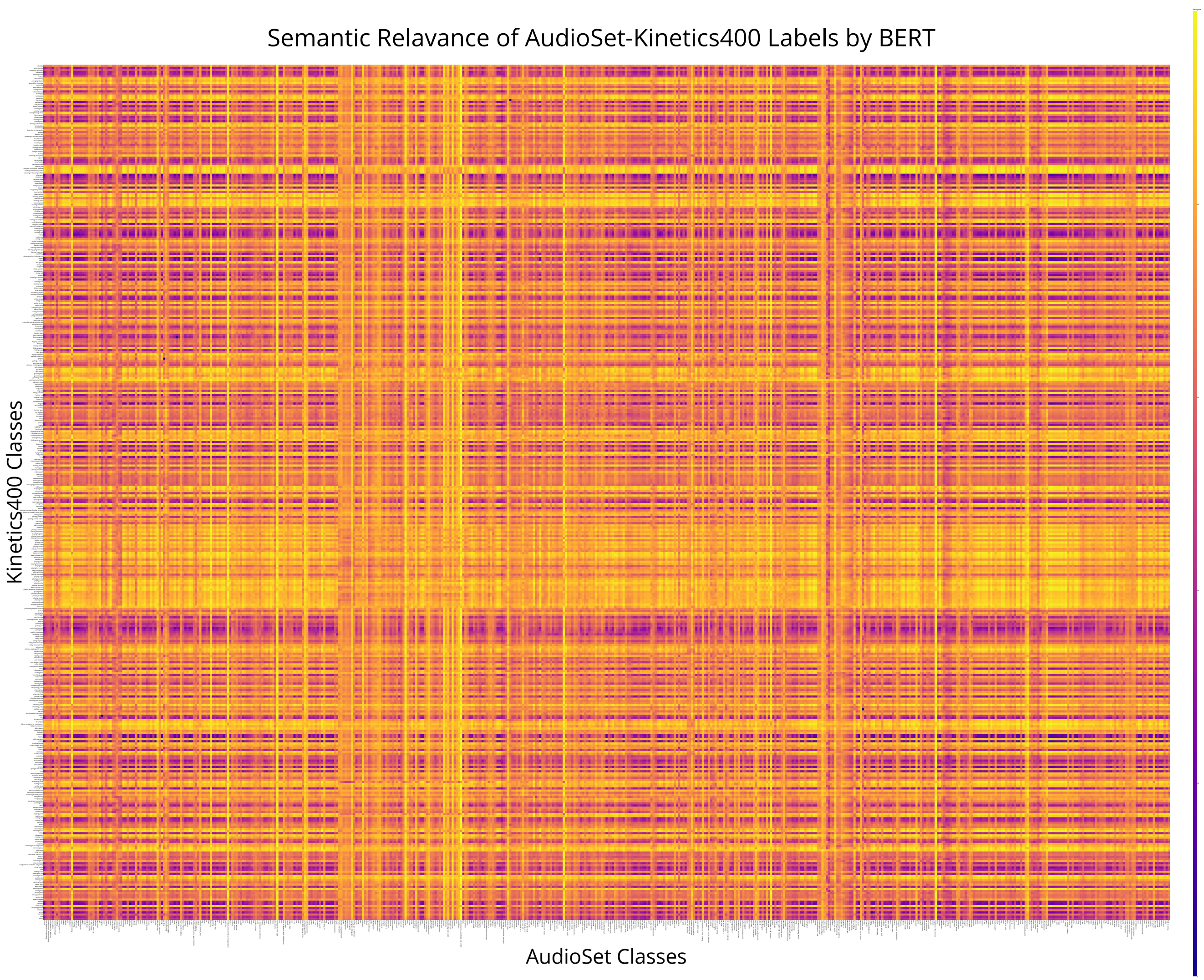

Figure 2. The heatmap of the semantic relevance estimated by BERT between Kinetics400 labels and AudioSet labels. The darker the more relevant.

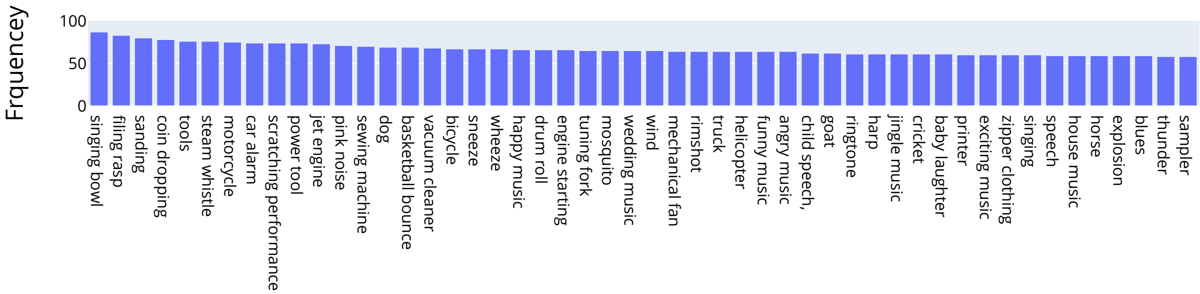

Figure 3. The most 50 frequent audio labels in AudioSet mapped to UCF-101 labels when $k = 50$.

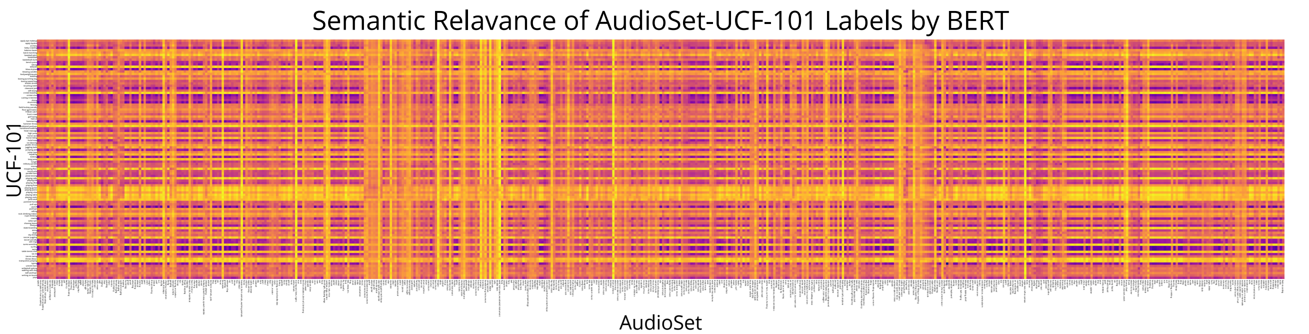

Figure 4. The heatmap of the semantic relevance estimated by BERT between UCF-101 labels and AudioSet labels. The darker the more relevant.

CVPR, pages 4724–4733. IEEE, jul 2017.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[5] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem(i):1933–1941, 2016.

[6] Hilde Kuehne, Hueihan Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, nov 2011.

[7] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[8] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv*, dec 2012.

[9] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV 2016*, pages 20–36. 2016.

[10] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krahenbuhl. Compressed Video Action Recognition. In *CVPR*, pages 6026–6035. IEEE, jun 2018.

[11] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual SlowFast Networks for Video Recognition. *arXiv*, 2020.