Parameter Efficient Multimodal Transformers for Video Representation Learning

ICLR 2021






Sangho Lee
SNU

Youngjae Yu
SNU

Gunhee Kim
SNU

Thomas Breuel
NVIDIA

Jan Kautz
NVIDIA

Yale Song
MSR

SEOUL NATIONAL UNIV. VISION & LEARNING

NVIDIA

Microsoft Research

[Goal]

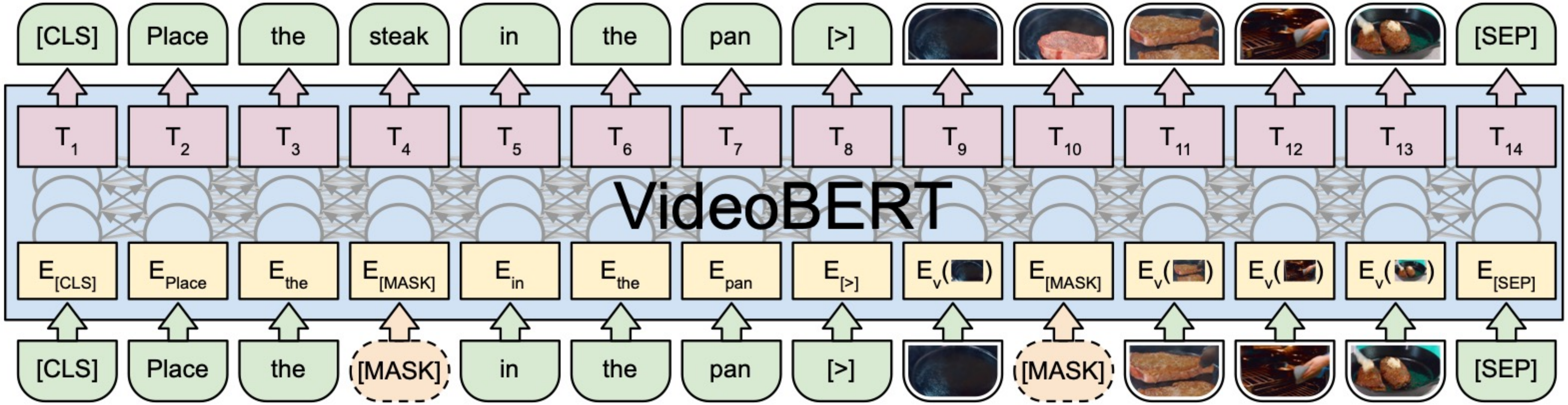
Learn from unlabeled videos by leveraging audio-visual correlations

[Our Solution]

Train **multimodal Transformers** in a **self-supervised** manner

Previous Work:
Multimodal Transformers

Multimodal Transformers have been widely used in vision-and-language tasks

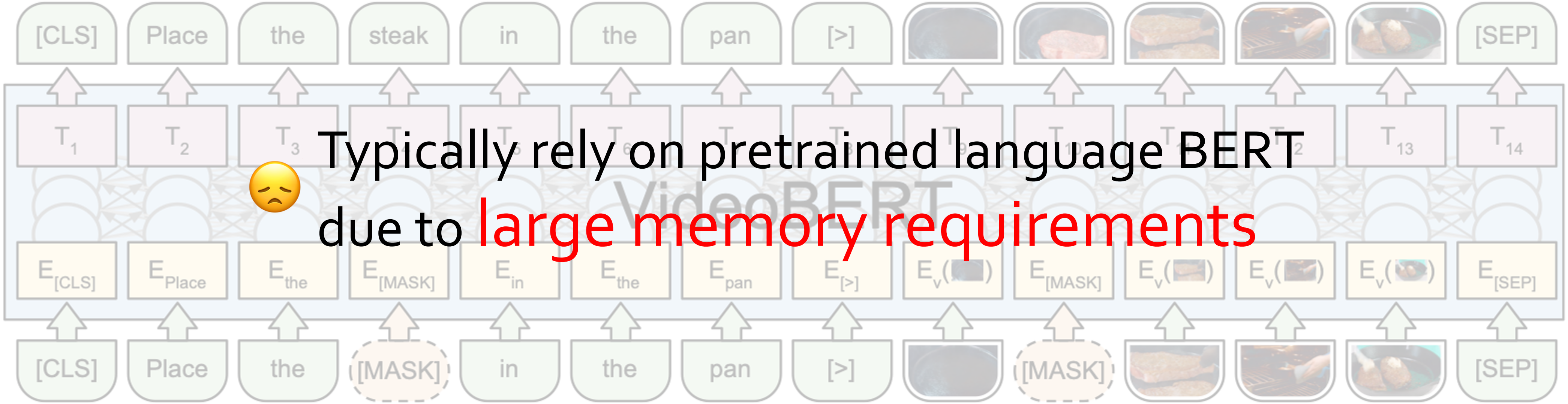


Tan and Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *EMNLP-IJCNLP*
Sun et al. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. *ICCV*
Lu et al. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *NeurIPS*

Previous Work:

Multimodal Transformers

Multimodal Transformers have been widely used in vision-and-language tasks



Typically rely on pretrained language BERT due to large memory requirements

Tan and Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *EMNLP-IJCNLP*
Sun et al. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. *ICCV*
Lu et al. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *NeurIPS*

[Problem]
We do not have pretrained components in the task of audio-visual representation learning

⇨ We need to train models end-to-end

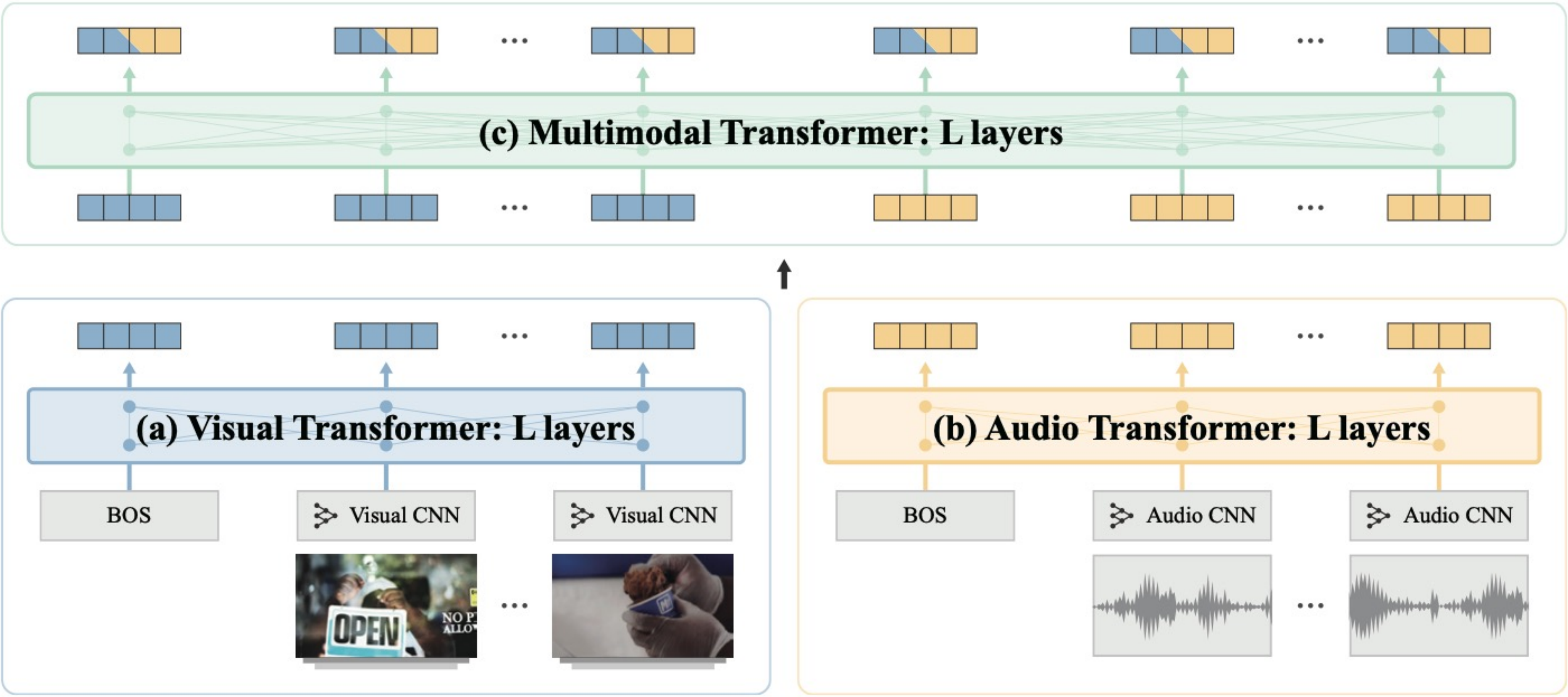
⇨ **We need to reduce the model size!**

Contributions

1. First **end-to-end trainable** audio-visual Transformers
   a. By using a novel **parameter reduction** scheme
2. Novel **content-aware negative sampling** for contrastive learning objectives
3. **Competitive** results on visual-only / audio-only / audio-visual downstream tasks

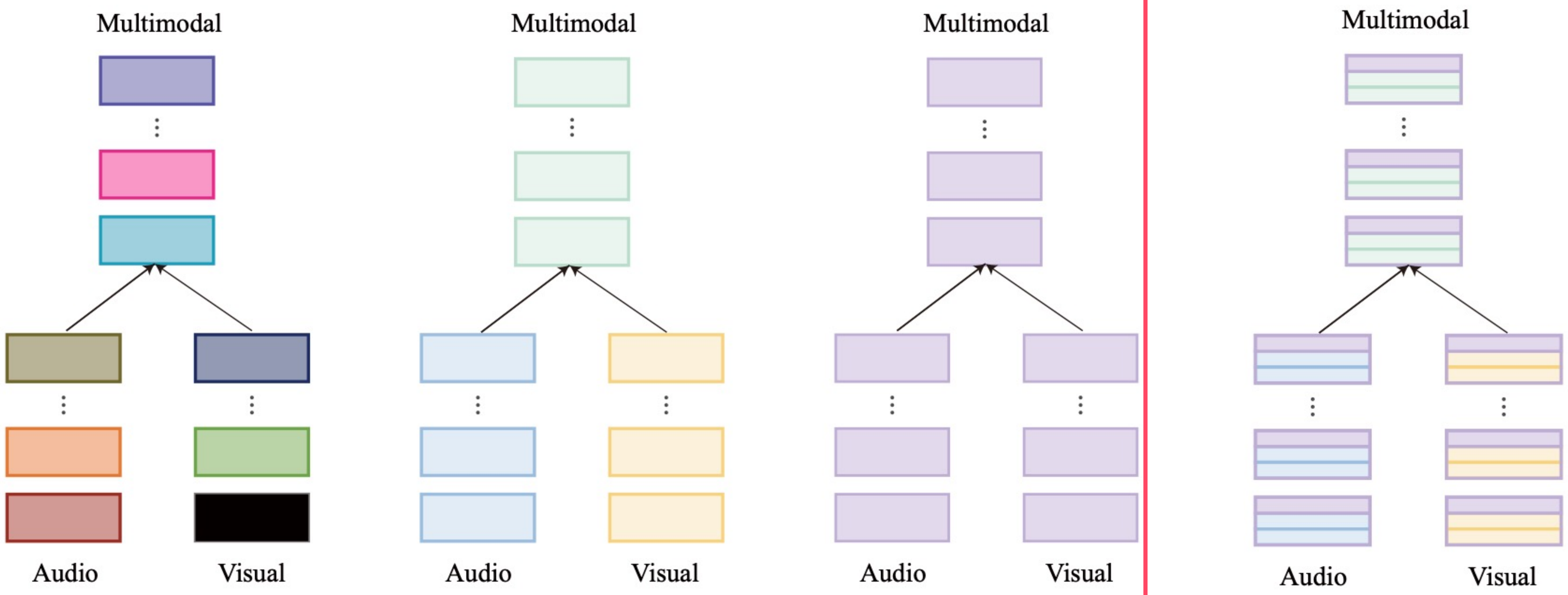
Our Architecture

Mid-fusion multimodal Transformers (**no pretrained** components)



Parameter Sharing Schemes

No Sharing | Cross-Layers Sharing | All Sharing | **Ours: Partial Sharing + Low-Rank Factorization**



128M parameters | 21M | 7M | 4M (**97%** reduction)

Lan et al. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ICLR*
Jaegle et al. 2021. Perceiver: General Perception with Iterative Attention. arXiv preprint

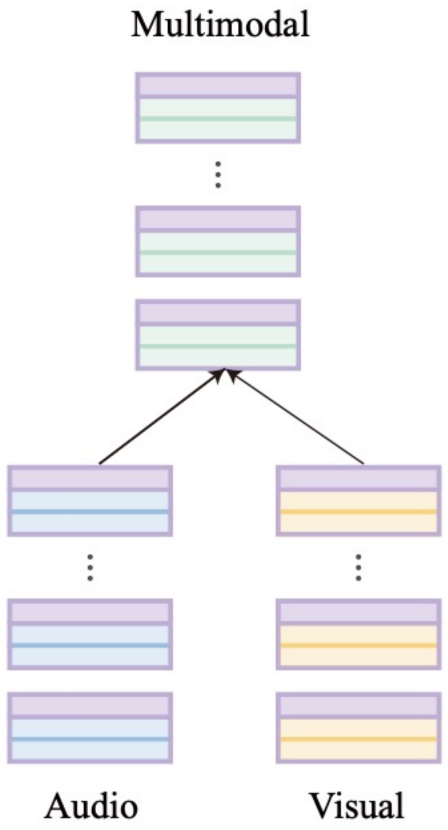
Weight Sharing via Low-Rank Factorization

Perform low-rank factorization of $W \in \mathbb{R}^{M\times N} (O \ll M, N)$

$$W = \underbrace{U}_{shared} \overbrace{\Sigma V^\top}^{modality\ specific}$$

$$U \in \mathbb{R}^{M\times O}, \Sigma \in \mathbb{R}^{O\times O}, V \in \mathbb{R}^{N\times O}$$

1. Reduce # of parameters: (M + N + O)O $\ll MN$
2. Able to model dynamics of each modality (Σ, V)



128M → **4M**
(**97%** reduction)

Experiments: Low-Rank Factorization

Results on Kinetics-Sounds (audio-visual classification benchmark)

Multi-6: Mid-fusion model, each Transformer of which has 6 layers
X.-L: Cross-Layers sharing
X.-T: Cross-Transformers sharing (All: all sharing, **Part**: low-rank factorization (**ours**))

Model	X.-L	X.-T	Params	top-1/5
`Multi-6`	✗	✗	128M	- / -
`Multi-6`	✓	✗	21M	65.7 / 89.9
`Multi-6`	✓	✓(`All`)	7M	67.1 / 92.3
`Multi-6`	✓	✓(`Part`)	**4M**	**67.5 / 92.3**

Arandjelovic and Zisserman. 2017. Look, Listen and Learn. *ICCV*

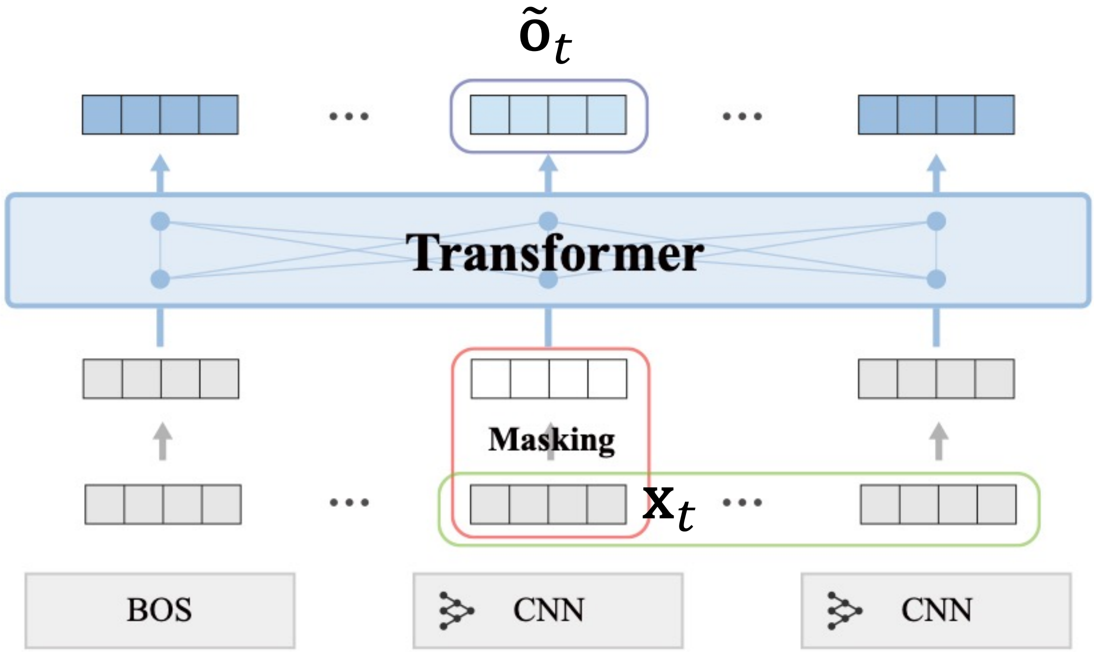
Self-Supervised Learning Task

Masked **E**mbedding **P**rediction (MEP)

Identify the correct masked input compared to a set of negative samples

mutual information

$$\mathcal{L}_{\text{MEP}}(\mathbf{x}, \tilde{\mathbf{o}}) = -\mathbb{E}_{\mathbf{x}} \left[\sum_t \log \frac{\mathrm{I}(\mathbf{x}_t, \tilde{\mathbf{o}}_t)}{\mathrm{I}(\mathbf{x}_t, \tilde{\mathbf{o}}_t) + \sum_{j \in \text{neg}(t)} \mathrm{I}(\mathbf{x}_j, \tilde{\mathbf{o}}_t)} \right]$$



Content-Aware Negative Sampling (CANS)

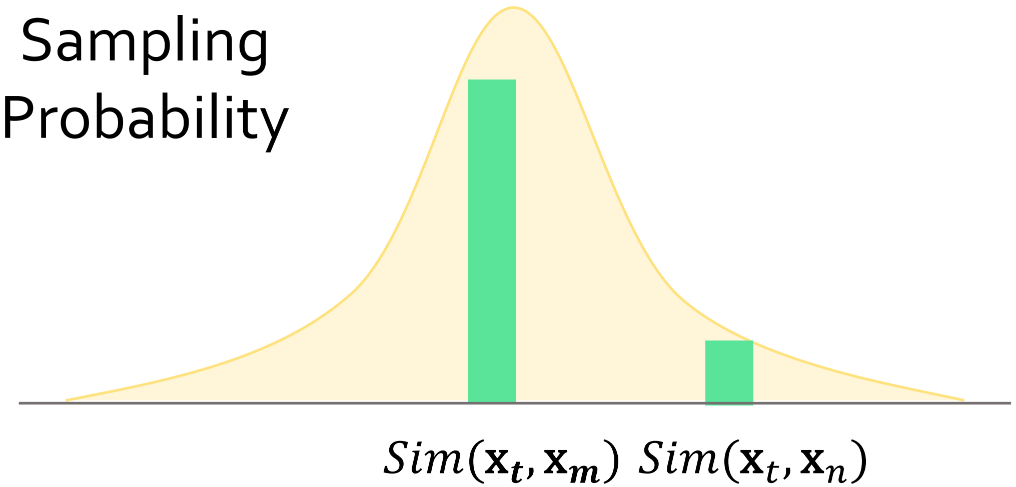
Stochastic sampling based on $Sim(\mathbf{x}_t, \mathbf{x}_{\boldsymbol{j}})$

- favors negatives sufficiently similar to x_t

$$\mathcal{L}_{\mathrm{MEP}}(\mathbf{x}, \tilde{\mathbf{o}}) = -\mathbb{E}_{\mathbf{x}}\left[\sum_t \log \frac{\mathrm{I}(\mathbf{x}_t, \tilde{\mathbf{o}}_t)}{\mathrm{I}(\mathbf{x}_t, \tilde{\mathbf{o}}_t) + \sum_{j \in \mathbf{neg}(t)} \mathrm{I}(\mathbf{x}_j, \tilde{\mathbf{o}}_t)}\right]$$

Ensure diversity, but favor **hard** negatives

$\longrightarrow$ make the MEP task effective!



$\mathbf{x}_t$ =  $\mathbf{x}_m$ =  $\mathbf{x}_n$ = 

Experiments: CANS

Results on Kinetics-Sounds (audio-visual classification benchmark)

Current-Sequence: Negative sampling *from the same sequence* (only **hard**)
Current-Minibatch: Negative sampling *from the same mini-batch* (too many **easy**)
CANS-Similar: **C**ontent-**A**ware **N**egative **S**ampling (**Ours)**

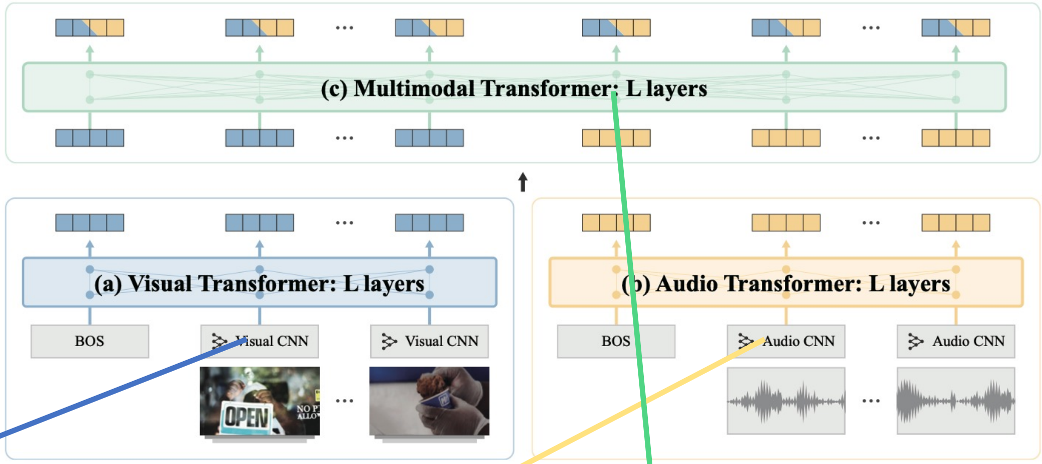
Sampling Method	top-1	top-5
`Current-Sequence`	64.6	89.8
`Current-MiniBatch`	65.5	90.8
`CANS-Similar`	**67.5**	**92.3**

Arandjelovic and Zisserman. 2017. Look, Listen and Learn. *ICCV*

Experiments: Downstream Tasks

Versatility

competitive results on several downstream tasks



a) Model	Net	Data	UCF
ST-Puzzle	3D-R18	K400	65.8
ClipOrder	R(2+1)D	UCF	72.4
DPC	3D-R34	K400	75.7
CBT	S3D	K600	79.5
MultiSens	3D-R18	AS	82.1
AVTS	MC3-18	K400	85.8
AVTS	MC3-18	AS	**89.0**
V-CNN†	SlowFast	K700	85.2
V-CNN†	SlowFast	AS	86.1

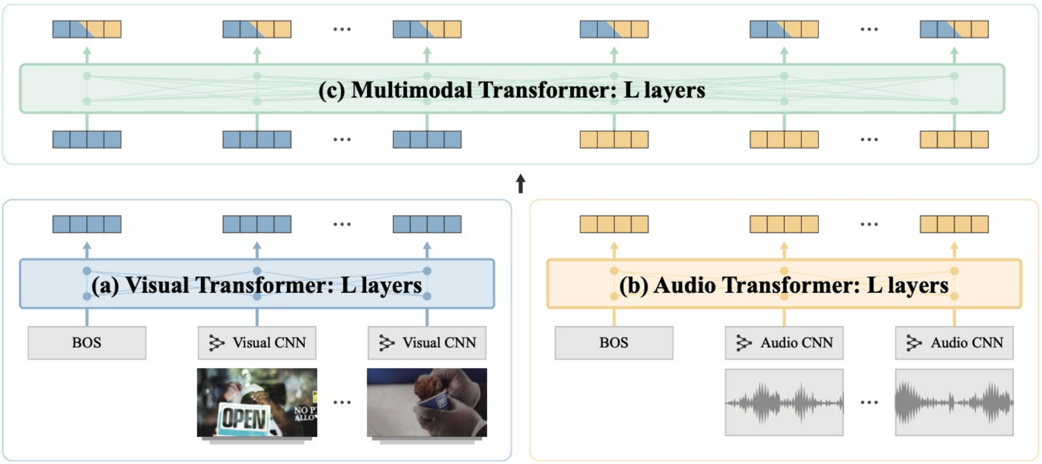
b) Model	Net	Data	ESC
SVM	MLP	-	39.6
ConvAE	CNN-4		39.9
RF	MLP	-	44.3
ConvNet	CNN-4	-	64.5
SoundNet	CNN-8	FS	74.2
L^3-Net	CNN-8	FS	79.3
DMC	VGG-ish	FS	79.8
AVTS	VGG-M	AS	80.6
A-CNN†	R50	AS	**81.5**

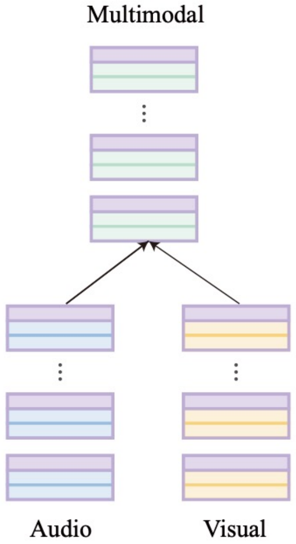
c) Model	Charades	KS
Random	5.9	- / -
ATF	18.3	- / -
ATF (OF)	22.4	- / -
V-CNN	18.7	45.8 / 73.3
A-CNN	18.9	49.4 / 76.9
M-CNN	23.1	59.4 / 83.6
V-BERT	26.0	49.5 / 78.9
A-BERT	27.4	58.9 / 85.7
M-BERT†	**29.5**	**75.6 / 94.6**

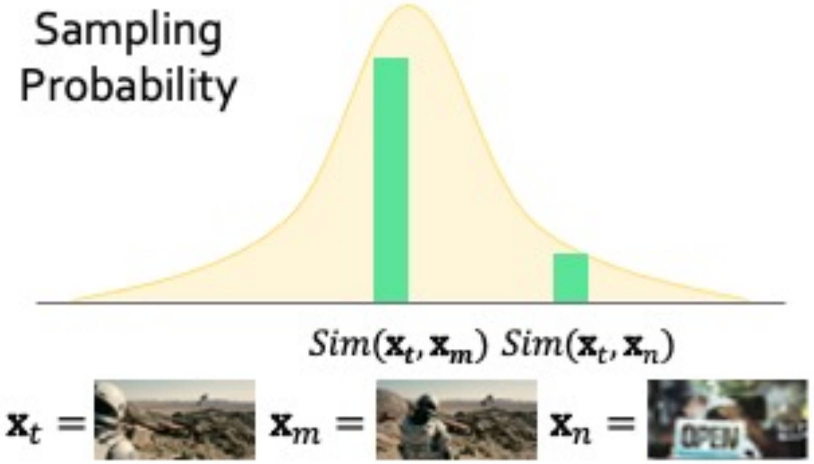
Datasets. K: Kinetics, AS: AudioSet, FS: Flicker-SoundNet, KS: Kinetics-Sounds.

Soomro et al. 2012. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. *CRCV-TR-12-01*
Piczak. 2015. ESC: Dataset for Environmental Sound Classification. *ACM-MM*
Sigurdsson et al. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. *ECCV*
Arandjelovic and Zisserman. 2017. Look, Listen and Learn. *ICCV*

Conclusion







First **end-to-end trainable** audio-visual Transformers / low-rank factorization

Content-aware negative sampling

a) Model	Net	Data	UCF
ST-Puzzle	3D-R18	K400	65.8
ClipOrder	R(2+1)D	UCF	72.4
DPC	3D-R34	K400	75.7
CBT	S3D	K600	79.5
MultiSens	3D-R18	AS	82.1
AVTS	MC3-18	K400	85.8
AVTS	MC3-18	AS	**89.0**
V-CNN†	SlowFast	K700	85.2
V-CNN†	SlowFast	AS	86.1

b) Model	Net	Data	ESC
SVM	MLP	-	39.6
ConvAE	CNN-4		39.9
RF	MLP	-	44.3
ConvNet	CNN-4	-	64.5
SoundNet	CNN-8	FS	74.2
L^3-Net	CNN-8	FS	79.3
DMC	VGG-ish	FS	79.8
AVTS	VGG-M	AS	80.6
A-CNN†	R50	AS	**81.5**

c) Model	Charades	KS
Random	5.9	- / -
ATF	18.3	- / -
ATF (OF)	22.4	- / -
V-CNN	18.7	45.8 / 73.3
A-CNN	18.9	49.4 / 76.9
M-CNN	23.1	59.4 / 83.6
V-BERT	26.0	49.5 / 78.9
A-BERT	27.4	58.9 / 85.7
M-BERT†	**29.5**	**75.6 / 94.6**

Datasets. K: Kinetics, AS: AudioSet, FS: Flicker-SoundNet, KS: Kinetics-Sounds.

Competitive results on downstream tasks

Paper: https://openreview.net/forum?id=6UdQLhqJyFD

Project page: https://vision.snu.ac.kr/projects/avbert