11-775 Large-Scale Multimedia Analysis

Multimodal Self-Supervised Learning

Bernie Huang

March 24, 2021

Slide Courtesy: Andrew Zisserman, Ishan Misra, Carl Doersch

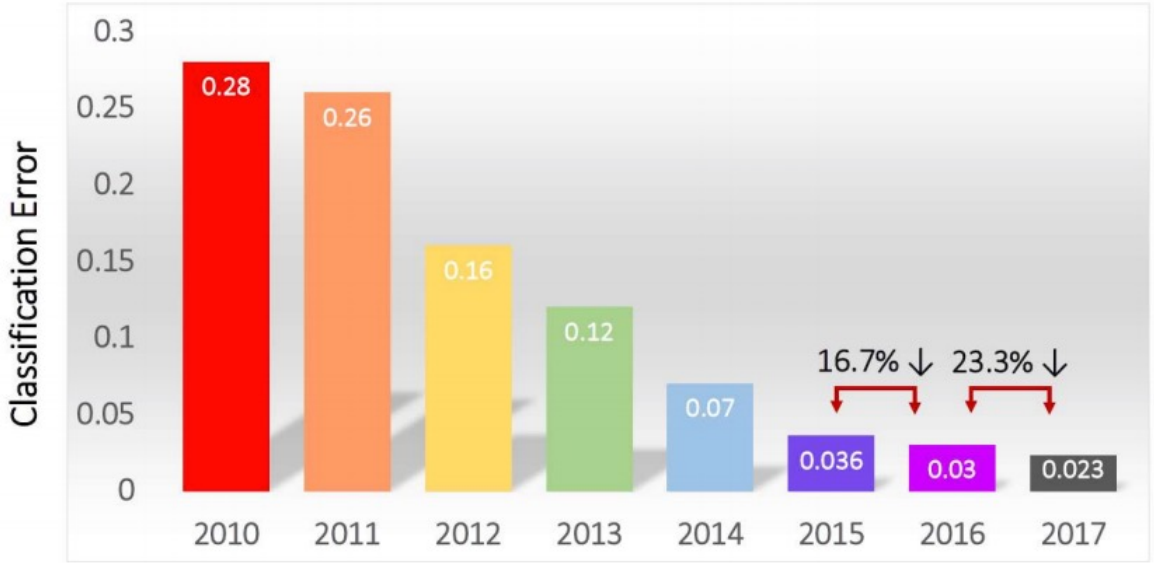
Carnegie Mellon University
School of Computer Science

Language Technologies Institute

The Story of ImageNet

Carnegie Mellon

- IMAGENET
  - Large-scale (1.2 million) image collection
  - Well-annotated 1000 classes
- ImageNet Classification Results year-by-year

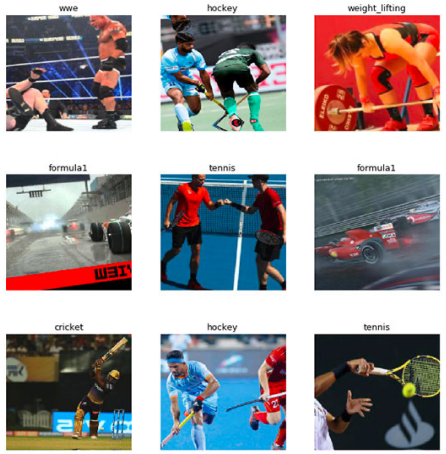


Better than Human now!

The outcome of ImageNet

Carnegie Mellon

- Features from networks trained on ImageNet can be used for other visual tasks, e.g. detection, segmentation, action recognition, fine grained visual classification
- To some extent, any visual task can be solved now by:
  - Construct a large-scale dataset labelled for that task
  - Specify a training loss and modify the ImageNet-pretrained network
  - Train it and deploy



Why Self-Supervision?

Carnegie Mellon

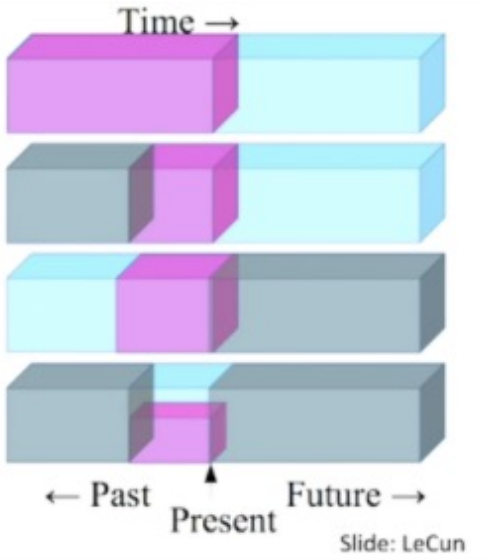
- Expense of producing a new dataset for each new task
- Some areas are data-sparse and supervision-starved, e.g., medical data, which it is hard to obtain and annotate
- Untapped/availability of vast numbers of un-labelled images/videos
  - Facebook: one billion images uploaded per day
  - 300 hours of video are uploaded to YouTube every minute
- How infants may learn …

Your HW is 22 hr videos
~5 sec of YouTube
upload traffic

What is Self-Supervision?

Carnegie Mellon

- A form of unsupervised learning where the data provides the supervision
- How to self-supervise:
  - Withhold/augment part of the data, and task the network with predicting it.
  - The pretext task defines a proxy loss, and the network is forced to learn what we really care about, e.g., a semantic representation, in order to solve it
- Examples:
  - Predict any part of the input from any other part
  - Predict future from the past
  - Predict future from the recent past
  - Predict past from the present
  - Predict top from the bottom
  - Predict the occluded part from the visible



Outline

Carnegie Mellon

- Single-modal self-supervised leaning
  - Image
  - Audio
  - Video
  - Text
- Multi-modal self-supervised learning
  - Audio and Video
  - Video and Text (ours ICLR 2021, NAACL 2021)

Image-based SSL

Carnegie Mellon

- Common workflow
  - A large collection of unlabeled images
  - Define a pretext task
  - Train a model on the pretext tasks with unlabeled images
  - Evaluation:
    - Use one intermediate feature layer of this model to feed a multinomial logistic regression classifier on ImageNet classification..
- Pretext tasks
  - Rotation
  - Colorization
  - Relative Position
  - Generative Models
  - Contrastive Learning

Image-based SSL

Carnegie Mellon

- Rotation
  - Gidaris et al., “Unsupervised Representation Learning by Predicting Image Rotations” ICLR 18

Image-based SSL

Carnegie Mellon

- Rotation
  - Gidaris et al., “Unsupervised Representation Learning by Predicting Image Rotations” ICLR 18

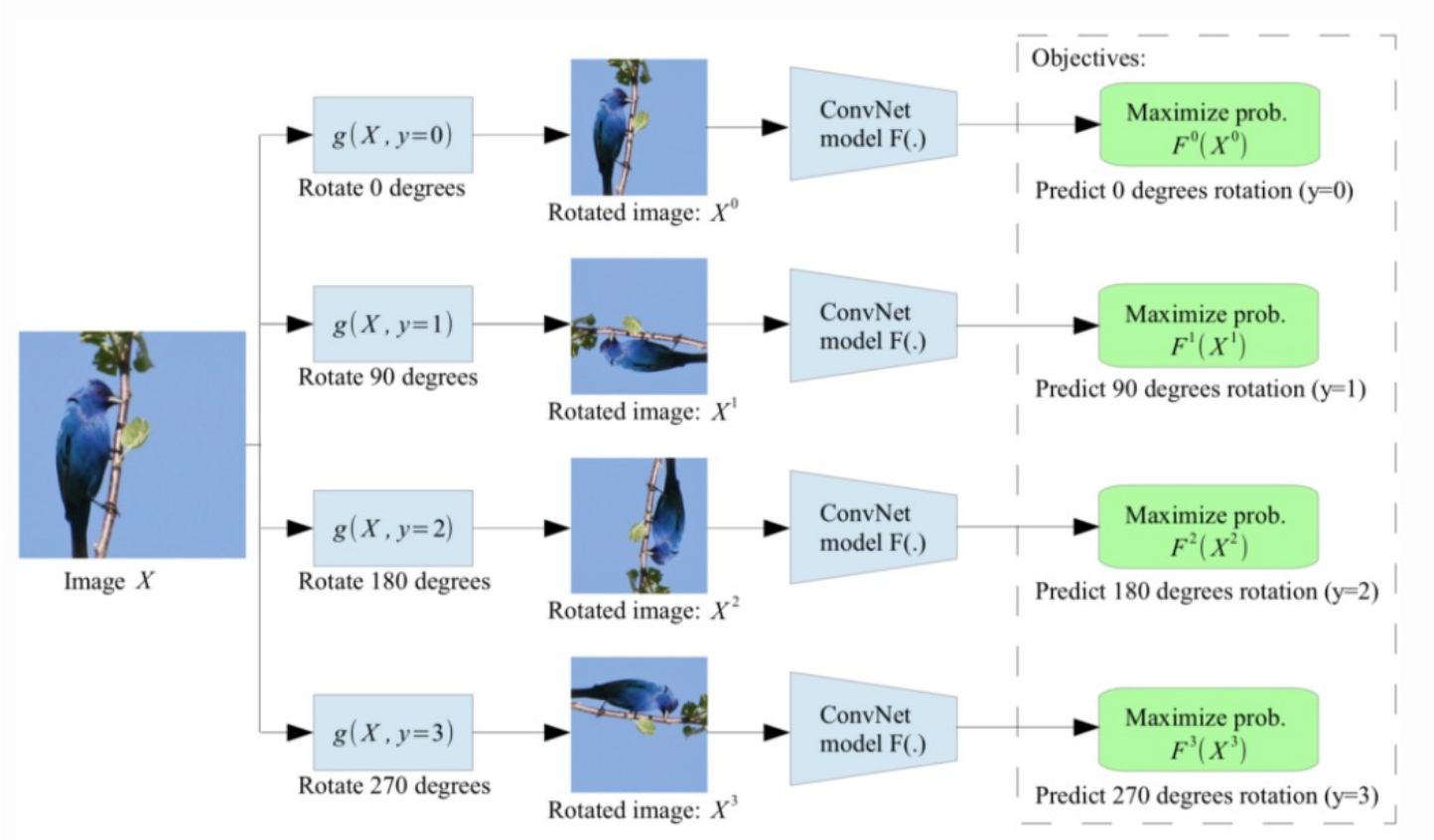


Image-based SSL

Carnegie Mellon

- Colorization
  - Colorful Image Colorization, Zhang et al., ECCV 2016

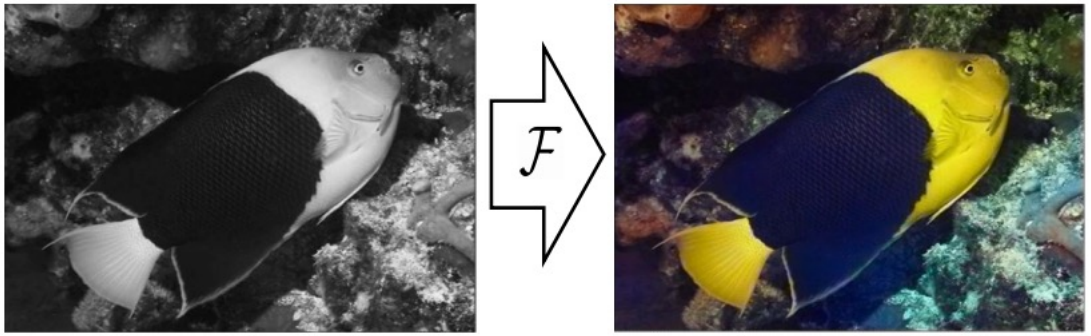
Train network to predict pixel colour from a monochrome input

Image-based SSL

Carnegie Mellon

- Colorization
  - Colorful Image Colorization, Zhang et al., ECCV 2016

Train network to predict pixel colour from a monochrome input



Grayscale image: *L* channel

$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Concatenate (L,ab)

$(\mathbf{X}, \hat{\mathbf{Y}})$

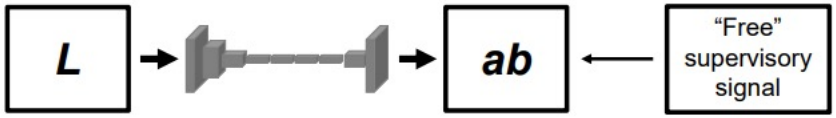


Image-based SSL

Carnegie Mellon

- Relative Position

Doersch, "Unsupervised Visual Representation Learning by Context Prediction" ICCV 15

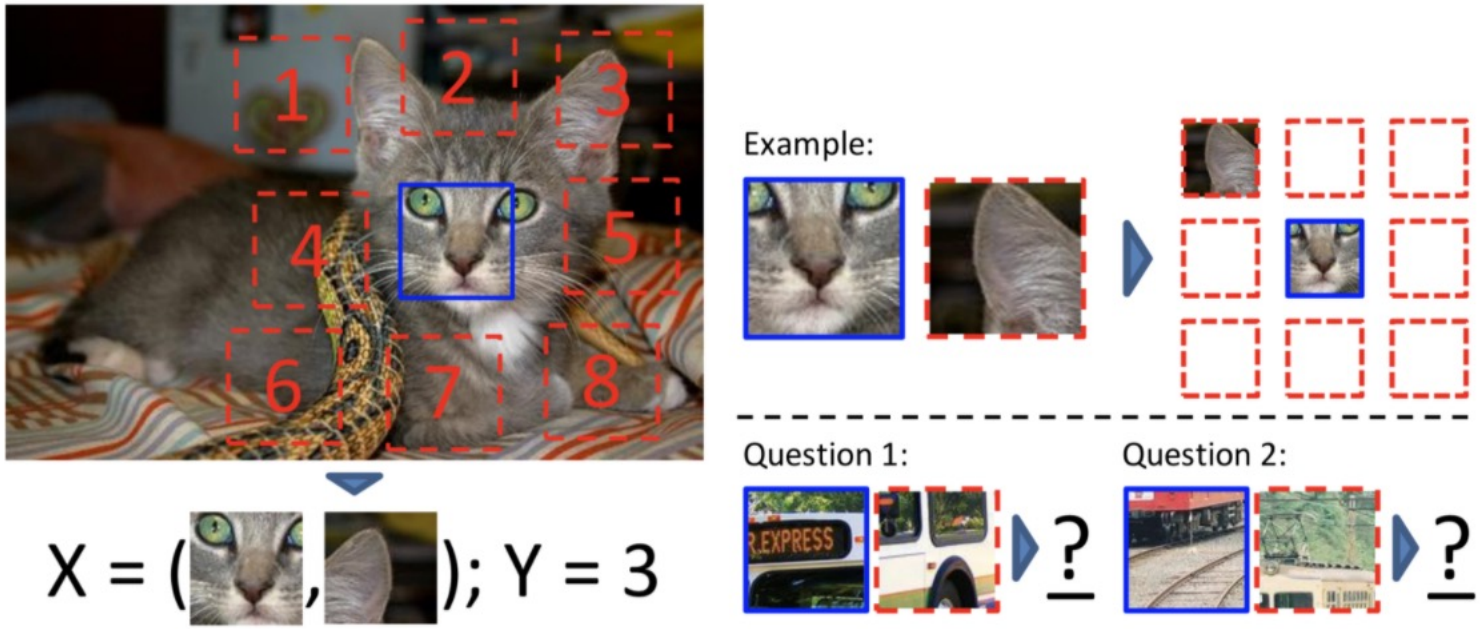


Image-based SSL (distortion)

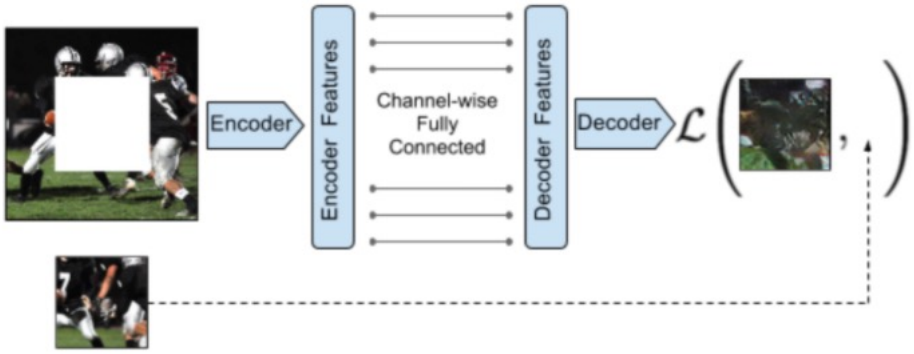
Carnegie Mellon

- Results:
  - Closes gap between ImageNet and self-supervision

	PASCAL VOC Detection mAP
Random	43.4
Rel. Pos.	51.1
Colour	46.9
Rotation	54.4
ImageNet Labels	56.8

Image-based SSL

- Generative models (Inpainting)
  - Pathak et al. “Context Encoders: Feature Learning by Inpainting” CVPR 2016



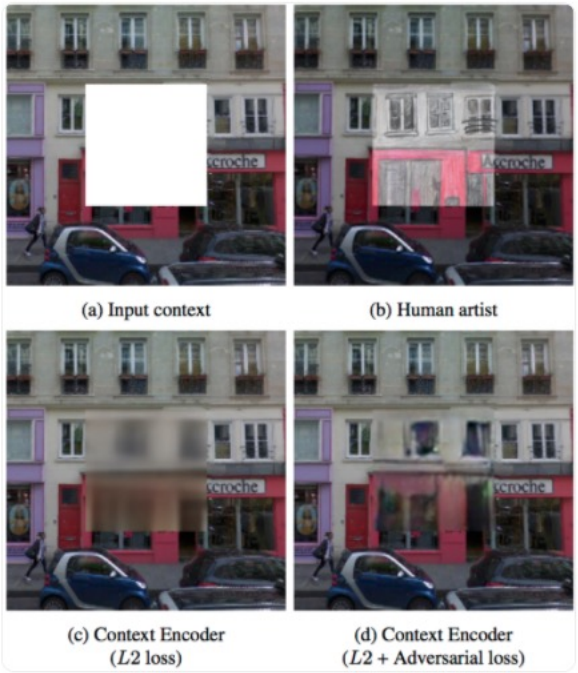
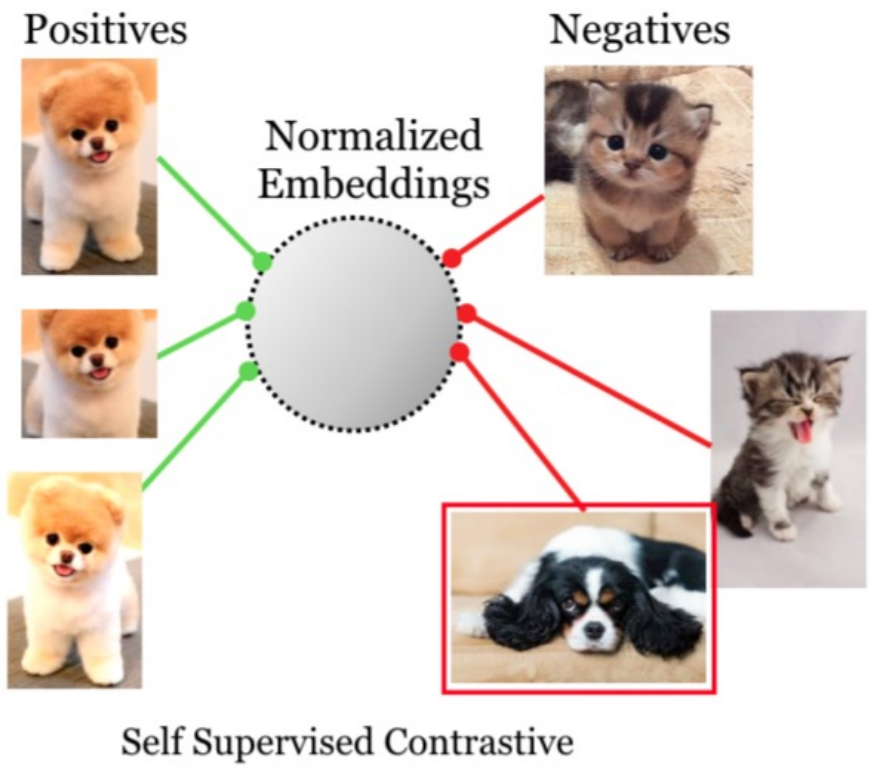


Image-based SSL

Carnegie Mellon

- Contrastive Learning



Self Supervised Contrastive

Generative / Predictive

Loss measured in the output space

Examples: Colorization, Auto-Encoders

Contrastive

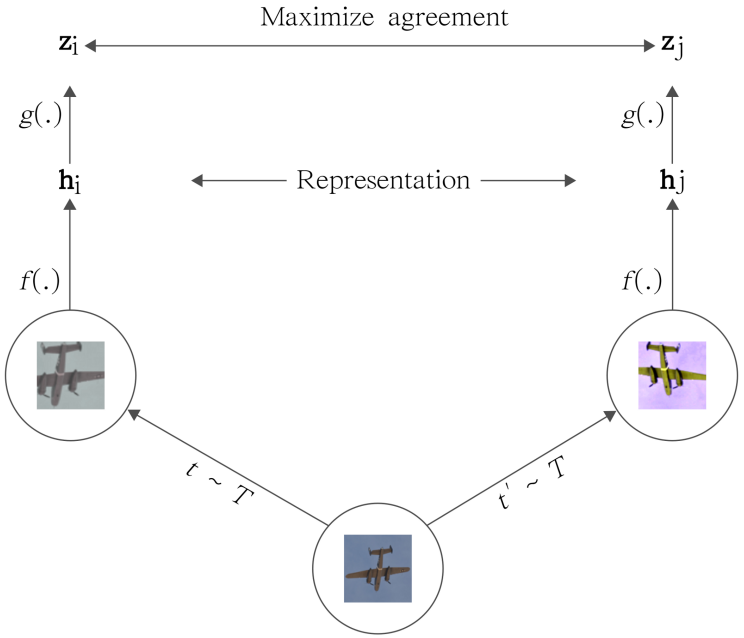
Loss measured in the representation space

Examples: TCN, CPC, Deep-InfoMax

Image-based SSL

Carnegie Mellon

- Example: SimCLR
  - Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations" ICML 2020



$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

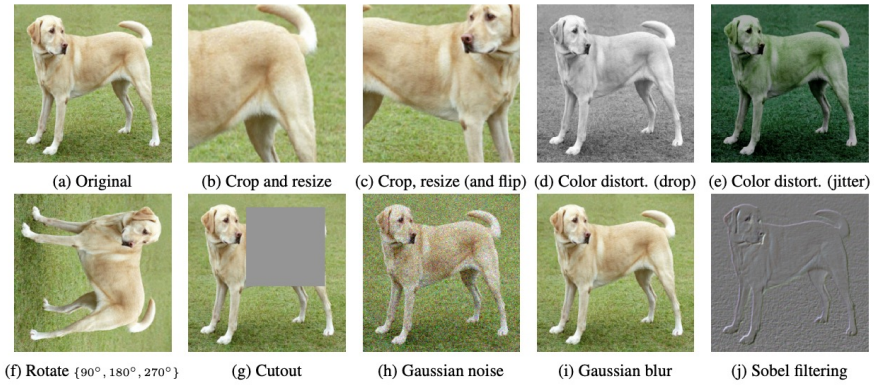
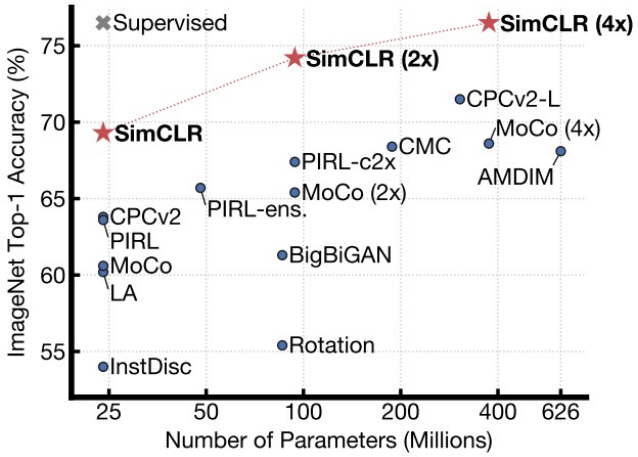


Image-based SSL

Carnegie Mellon

- Example: SimCLR
  - GAP to the supervised results is closed!



A Simple Framework for Contrastive Learning of Visual Representations

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

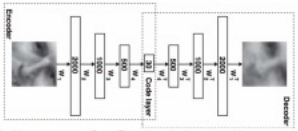
Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 $(4\times)$ models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B.8 for experimental details and results with standard ResNet-50.

Image-based SSL

Carnegie Mellon

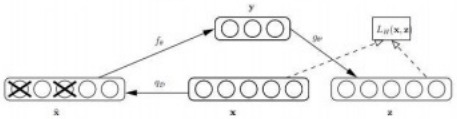
Others

Autoencoders



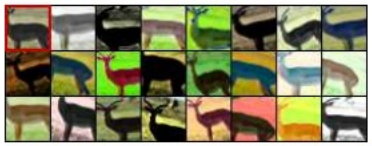
Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders



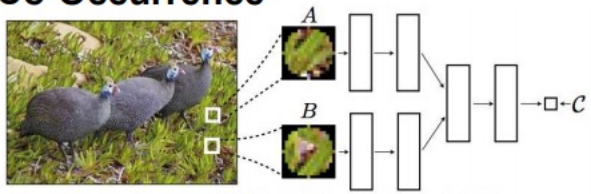
Vincent *et al.* ICML 2008.

Exemplar networks



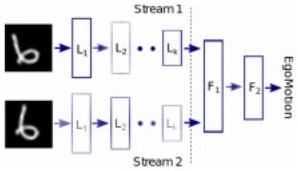
Dosovitskiy *et al.*, NIPS 2014

Co-Occurrence



Isola *et al.* ICLR Workshop 2016.

Egomotion



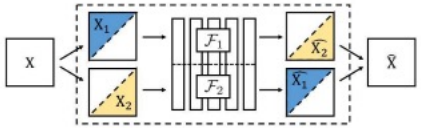
Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

Context

Noroozi et al 2016

Pathak *et al.* CVPR 2016

Split-brain auto-encoders



Zhang *et al.* CVPR 2017

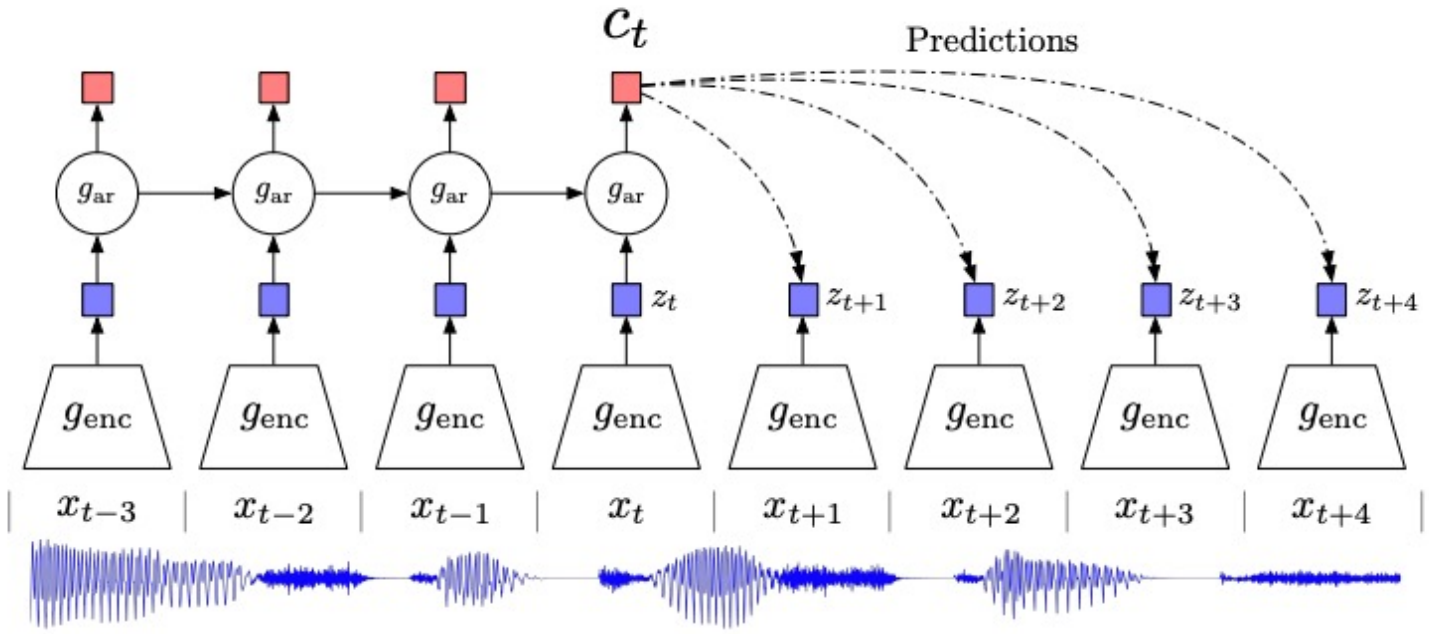
Outline

- Single-modal self-supervised leaning
  - Image
  - Audio
  - Video
  - Text
- Multi-modal self-supervised learning
  - Audio and Video
  - Video and Text

Audio-based SSL

Carnegie Mellon

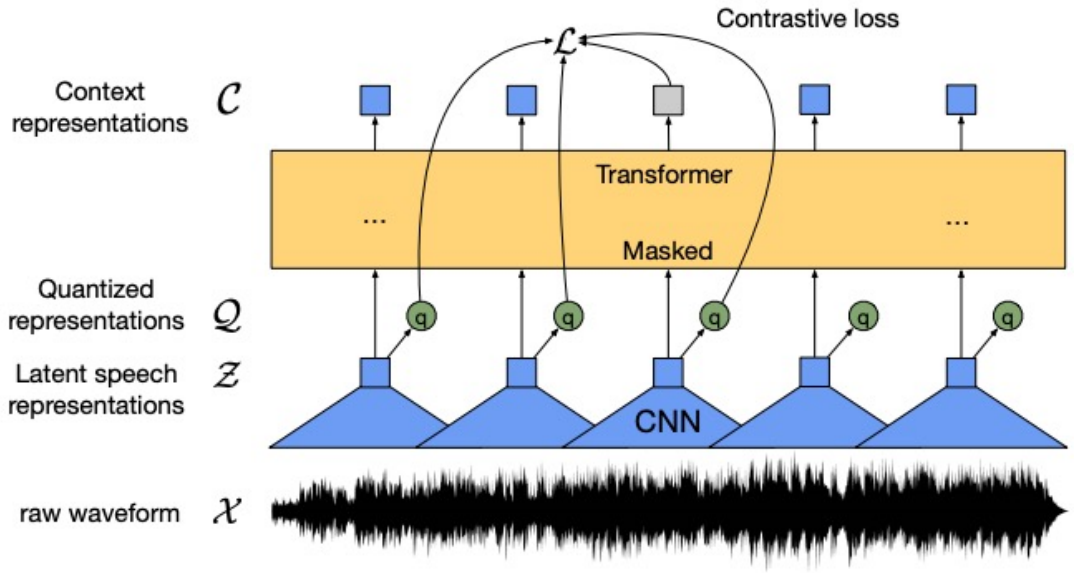
- Contrastive Predictive Coding (CPC)
  - Oord et al., “Representation Learning with Contrastive Predictive Coding” NeurIPS 2019



Audio-based SSL

Carnegie Mellon

- Wav2vec 2.0
  - Baevski et al, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations” NeurIPS 2020



Outline

Carnegie Mellon

- Single-modal self-supervised leaning
  - Image
  - Audio
  - Video
  - Text
- Multi-modal self-supervised learning
  - Audio and Video
  - Video and Text

Video-based SSL

Carnegie Mellon

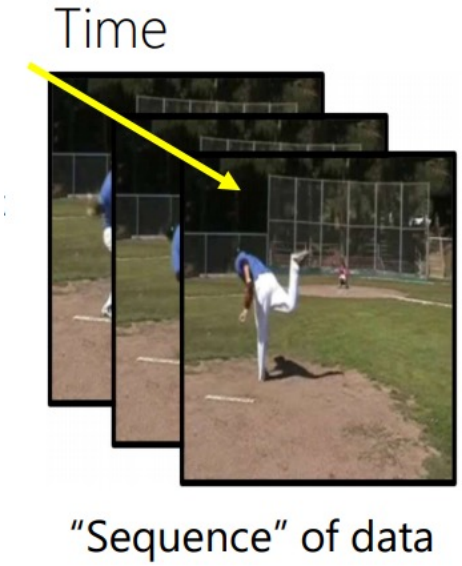
- Video: A temporal sequence of frames

- What can we use to define a pretext task/ proxy loss?
  - Nearby (in time) frames are strongly correlated, further away may not be
  - Temporal order of the frames
  - Motion of objects (via optical flow)
  - …etc

Video-based SSL

Carnegie Mellon

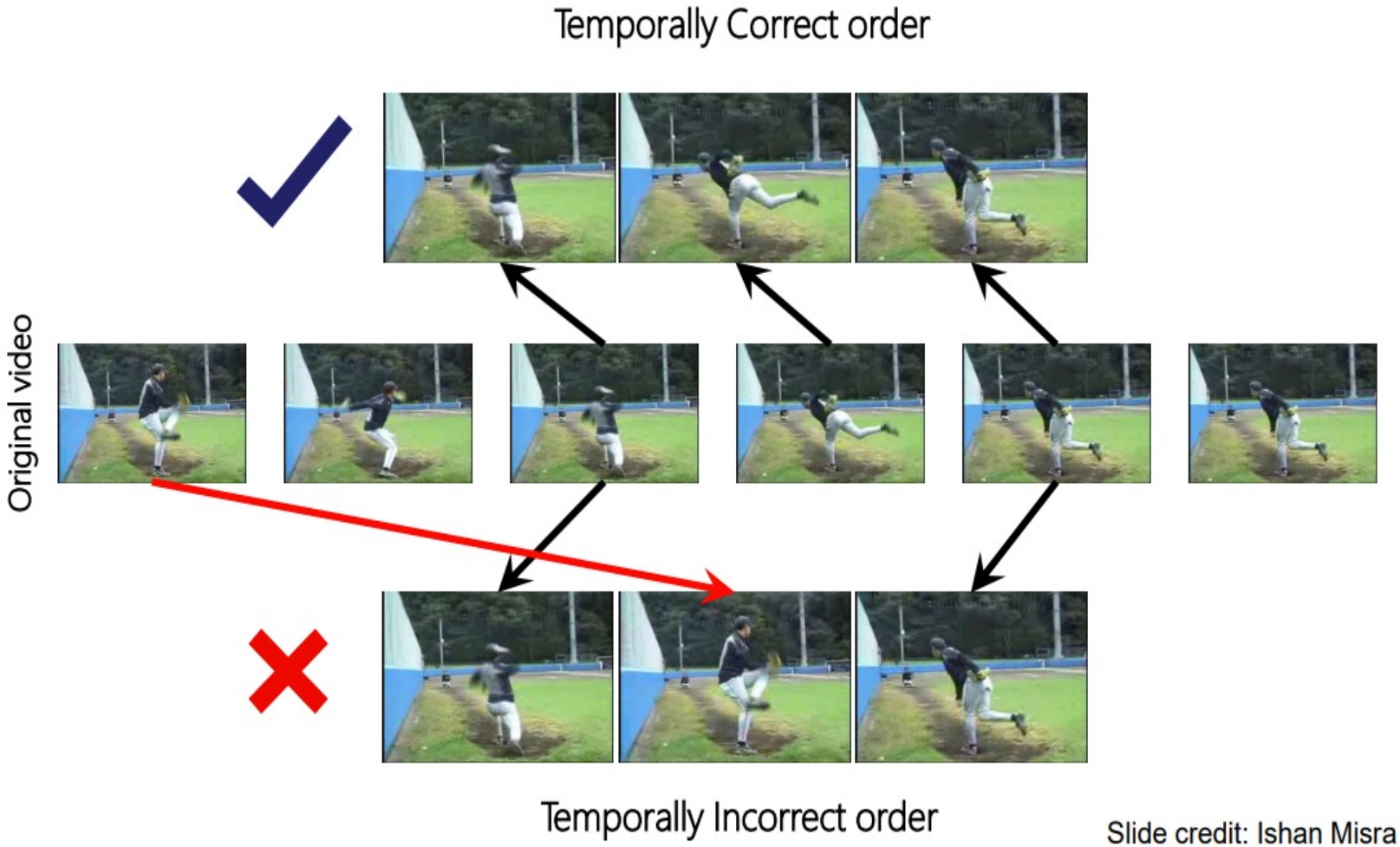
- Temporal structure in videos: (Video sequence order)
  - Misra et al, “Shuffle and Learn: Unsupervised Learning using Temporal Order Verification” ECCV 16



- Is this a valid sequence?

Video-based SSL

Carnegie Mellon

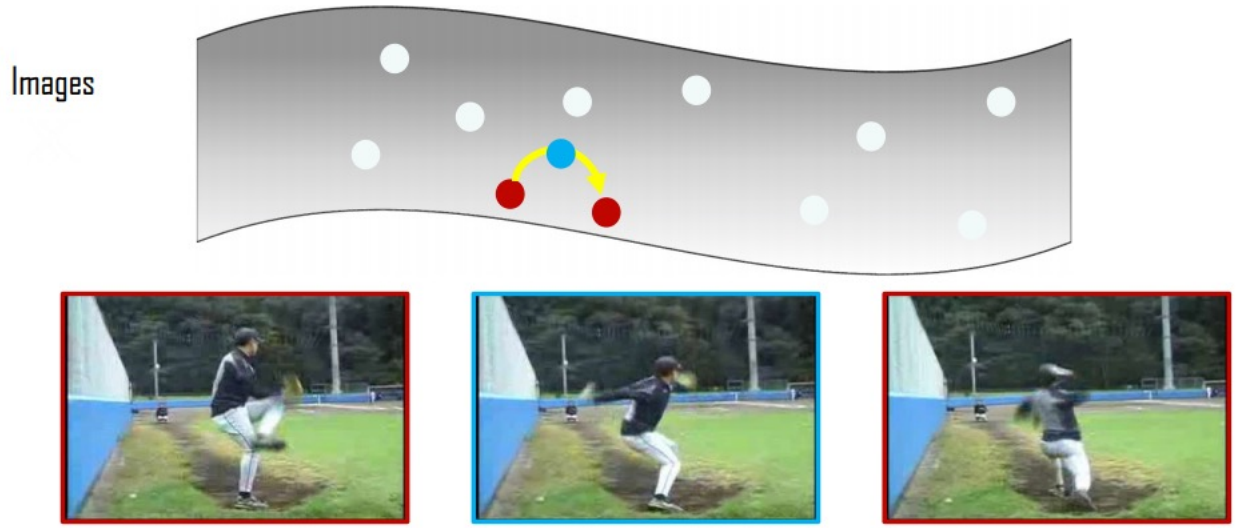


Slide credit: Ishan Misra

Video-based SSL

Carnegie Mellon

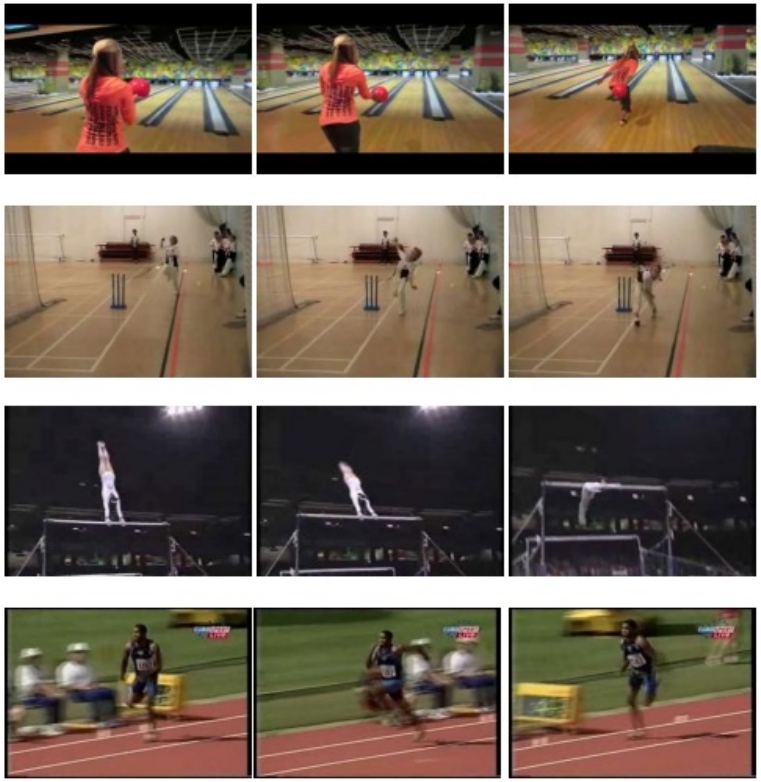
- Geometric View



Video-based SSL

Carnegie Mellon

Positive Tuples

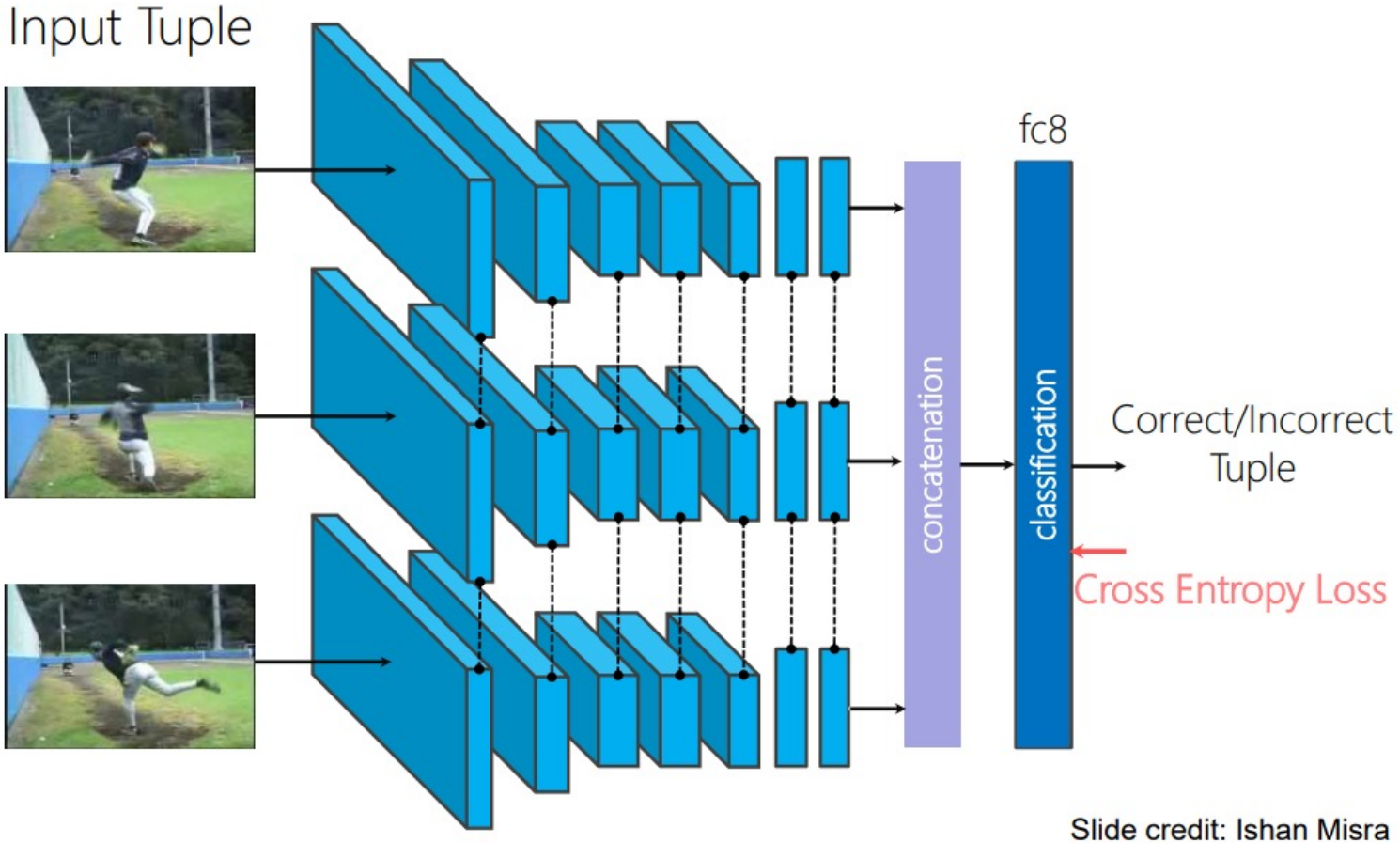


Negative Tuples

~900k tuples from UCF-101 dataset (Soomro et al., 2012)

Video-based SSL

Carnegie Mellon

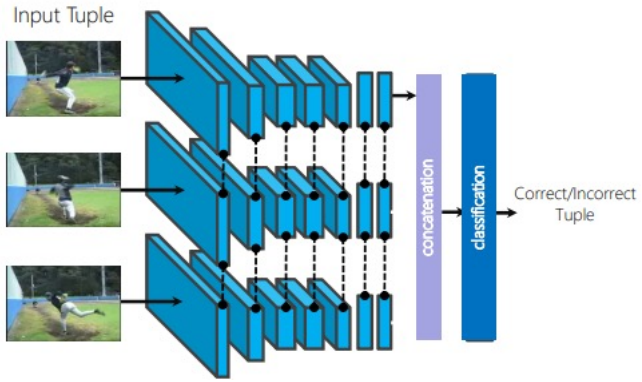


Slide credit: Ishan Misra

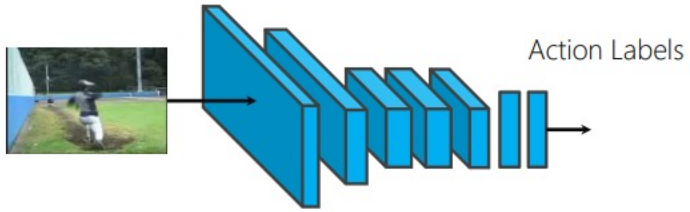
Video-based SSL

Carnegie Mellon

Self-supervised Pre-train



Test -> Finetune



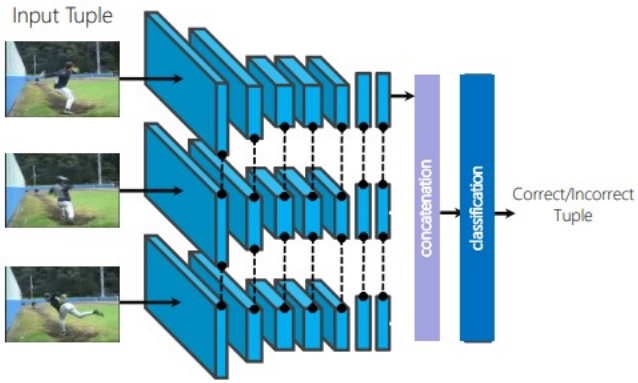
Dataset	Initialization	Mean Classification Accuracy
UCF101	Random	38.6
	Shuffle & Learn	50.2
	ImageNet pre-trained	**67.1**

Video-based SSL

Carnegie Mellon

- Human Pose Estimation

Self-supervised Pre-train

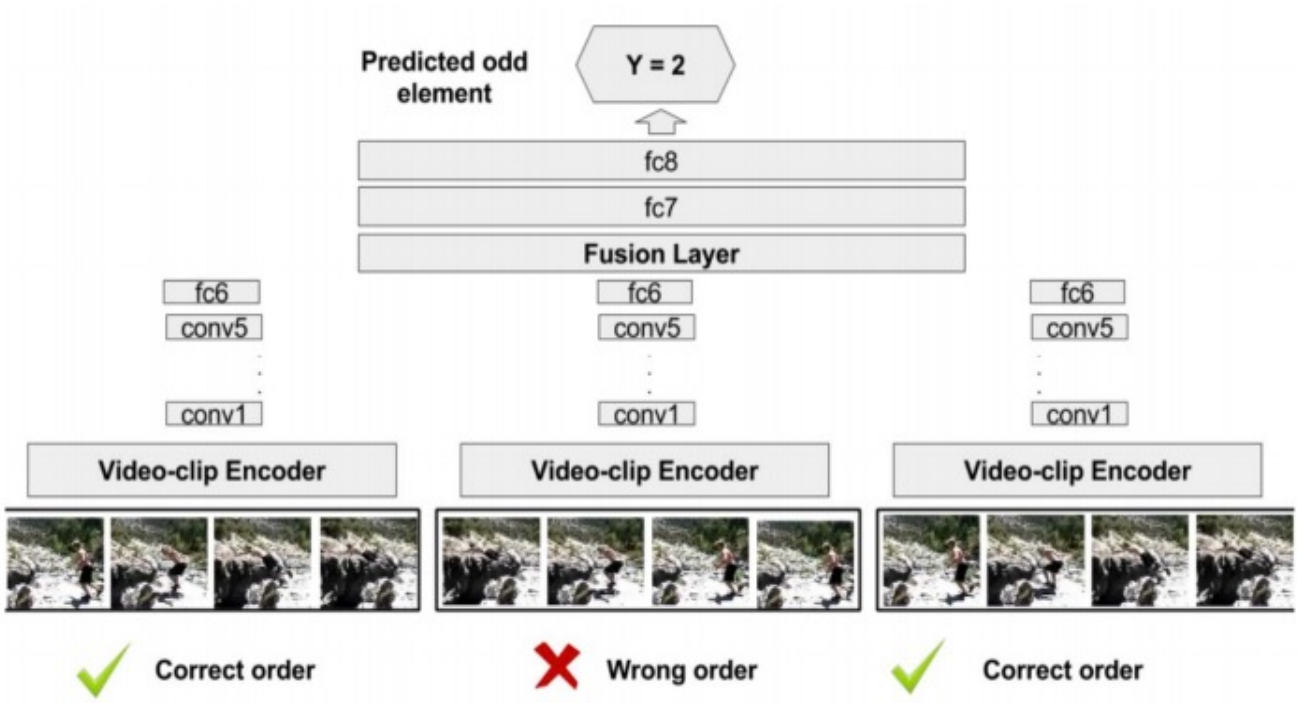


Initialization	FLIC Dataset		MPII Dataset	
	Mean PCK	AUC PCK	Mean PCKh@0.5	AUC PCKh@0.5
Shuffle & Learn	84.9	49.6	**87.7**	**47.6**
ImageNet pre-train	**85.8**	**51.3**	85.1	47.2

Video-based SSL

Carnegie Mellon

- More on temporal sequence
  - Fernando et al, “Odd-One-Out Networks” ICCV 2017

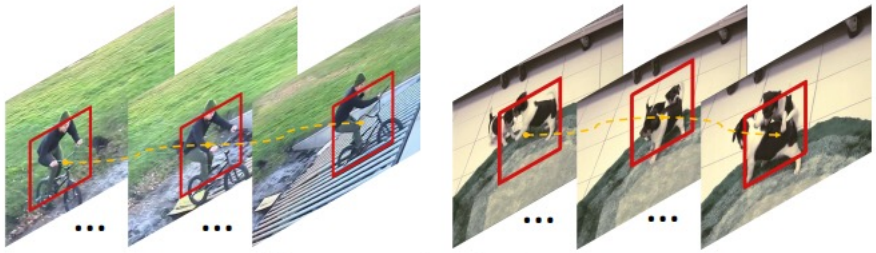


Initialization	Mean Classification Accuracy
Random	38.6
Shuffle and Learn	50.2
Odd-One-Out	60.3
ImageNet pre-trained	**67.1**

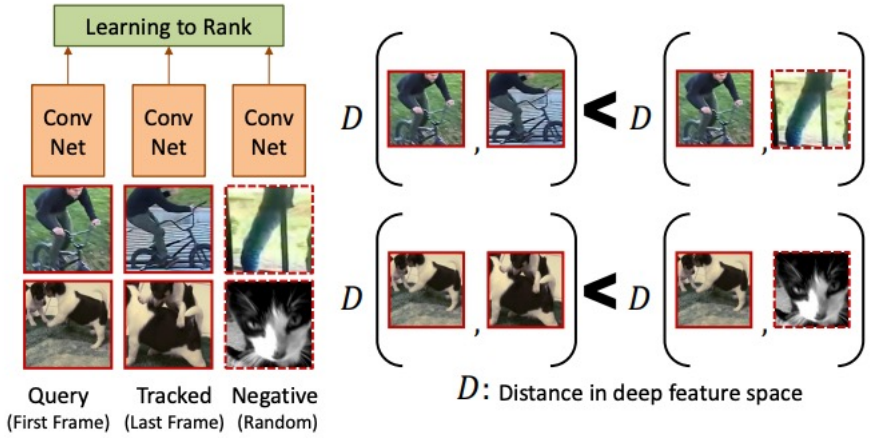
Video-based SSL

Carnegie Mellon

- Tracking moving objects:
  - Wang et al., "Unsupervised Learning of Visual Representations using Videos" CVPR 15

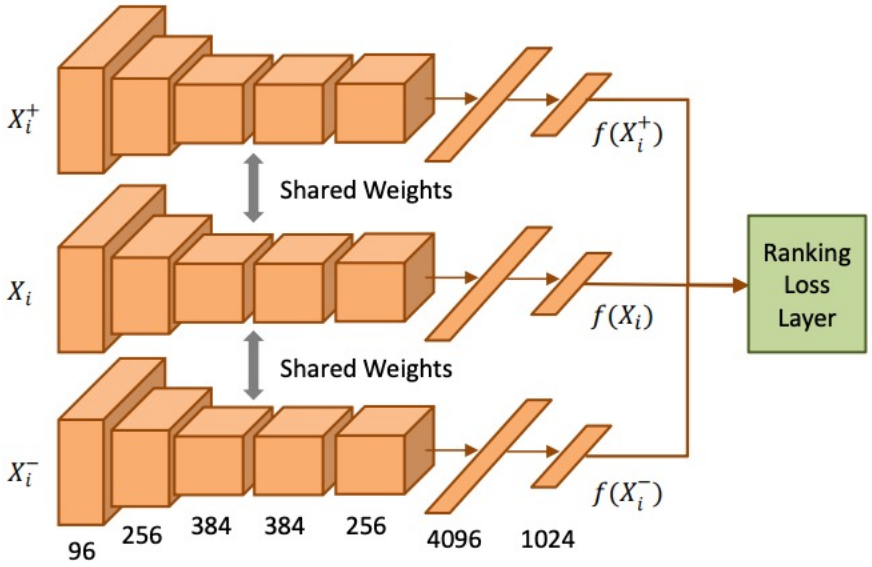


(a) Unsupervised Tracking in Videos



(b) Siamese-triplet Network

(c) Ranking Objective



Video-based SSL

Carnegie Mellon

- Summary: lessons so far:
  - Important to select informative data in training
    - Hard negatives and positives
    - Otherwise, most data is too easy or has no information and the network will not learn
    - Often use heuristics for this, e.g. motion energy
  - Consider how the network can possibly solve the task (without cheating)
    - This determines what it must learn, e.g. human key points in “shuffle and learn”
  - Choose the proxy task to encourage learning the features of interest

Outline

Carnegie Mellon

- Single-modal self-supervised leaning
  - Image
  - Audio
  - Video
  - Text
- Multi-modal self-supervised learning
  - Audio and Video
  - Video and Text

Text-based SSL

Carnegie Mellon

- Language as a sequence of tokens. Can we also design pretext tasks? What does the model learn?
- Contextual assupmtion:
  - Center word prediction: (continuous BoW)

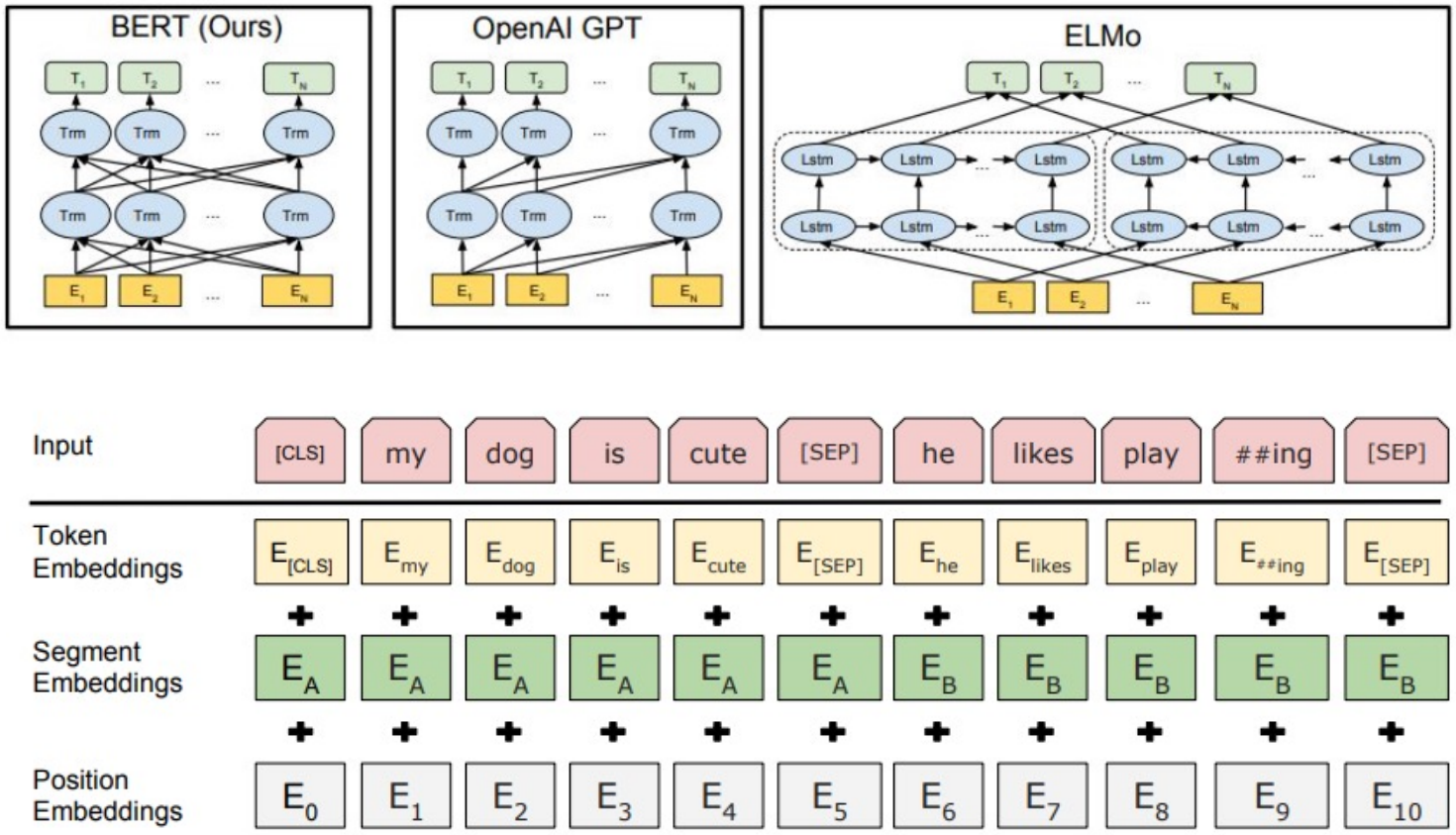
    A quick brown fox jumps over the lazy dog
  - Neighbor Word Prediction (skip-gram)

    A quick brown fox jumps over the lazy dog
  - Auto-regressive language modeling (e.g. n-gram)

    Nothing is _____
  - Masked language modeling

Text-based SSL

Carnegie Mellon



- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Devlin et al.
  - SoTA in 13 NLP tasks at that time with one pre-trained model

Outline

Carnegie Mellon

- Single-modal self-supervised leaning
  - Image
  - Audio
  - Video
  - Text
- Multi-modal self-supervised learning
  - Audio and Video
  - Video and Text

Multimodal SSL: A+V

Carnegie Mellon

- Audio-Visual Co-supervision

- Sound and frames are:
  - Semantically consistent
  - Synchronized

Multimodal SSL: A+V

Carnegie Mellon

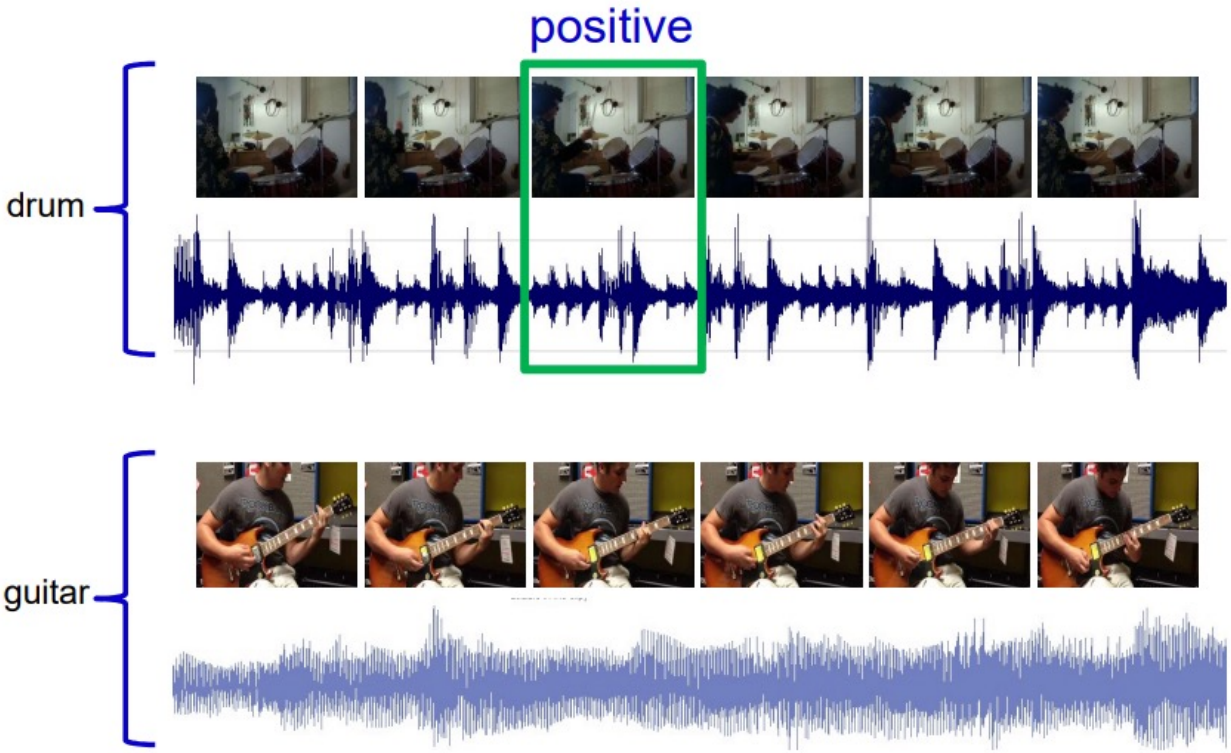
- Objective: use vision and sound to learn from each other

- Two types of proxy task:
  - Predict audio-visual correspondence
  - Predict audio-visual synchronization

Multimodal SSL: A+V

Carnegie Mellon

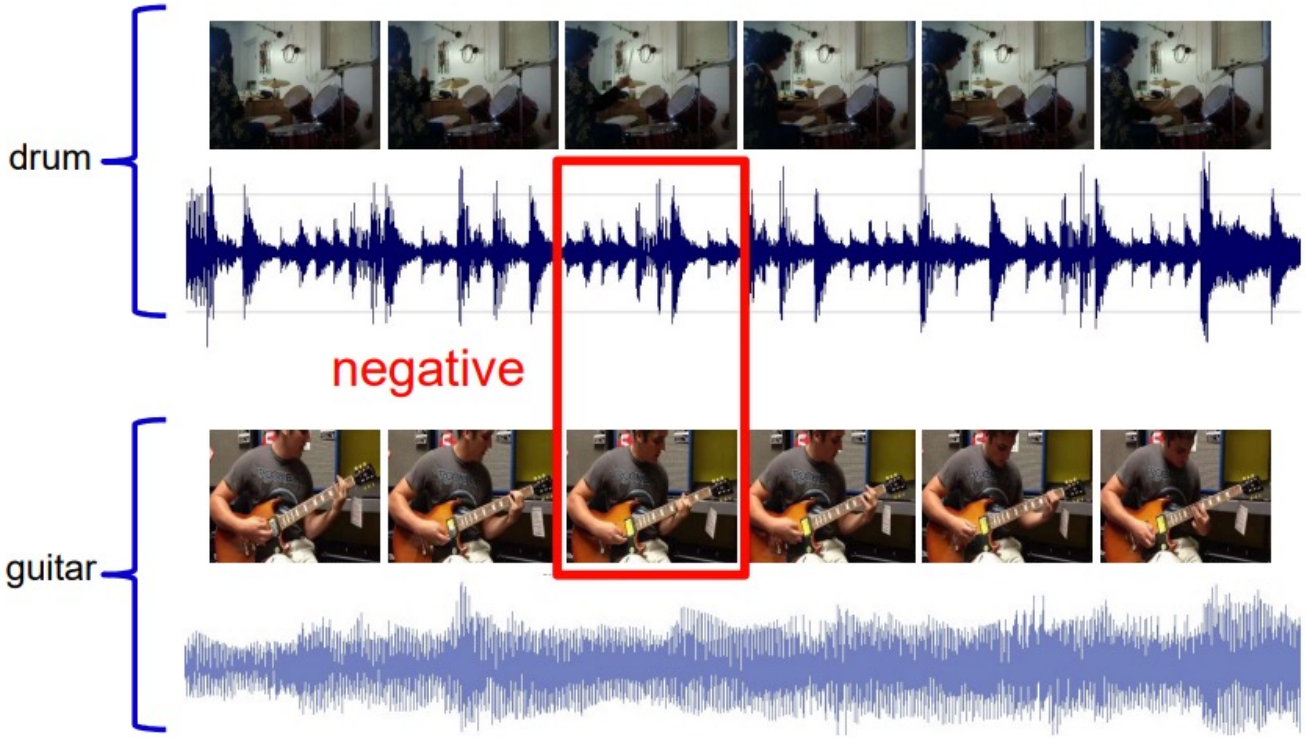
- Train a network to predict if image and audio clip correspond
  - Arandjelović and Zisserman, “Objects that Sound” ECCV 18



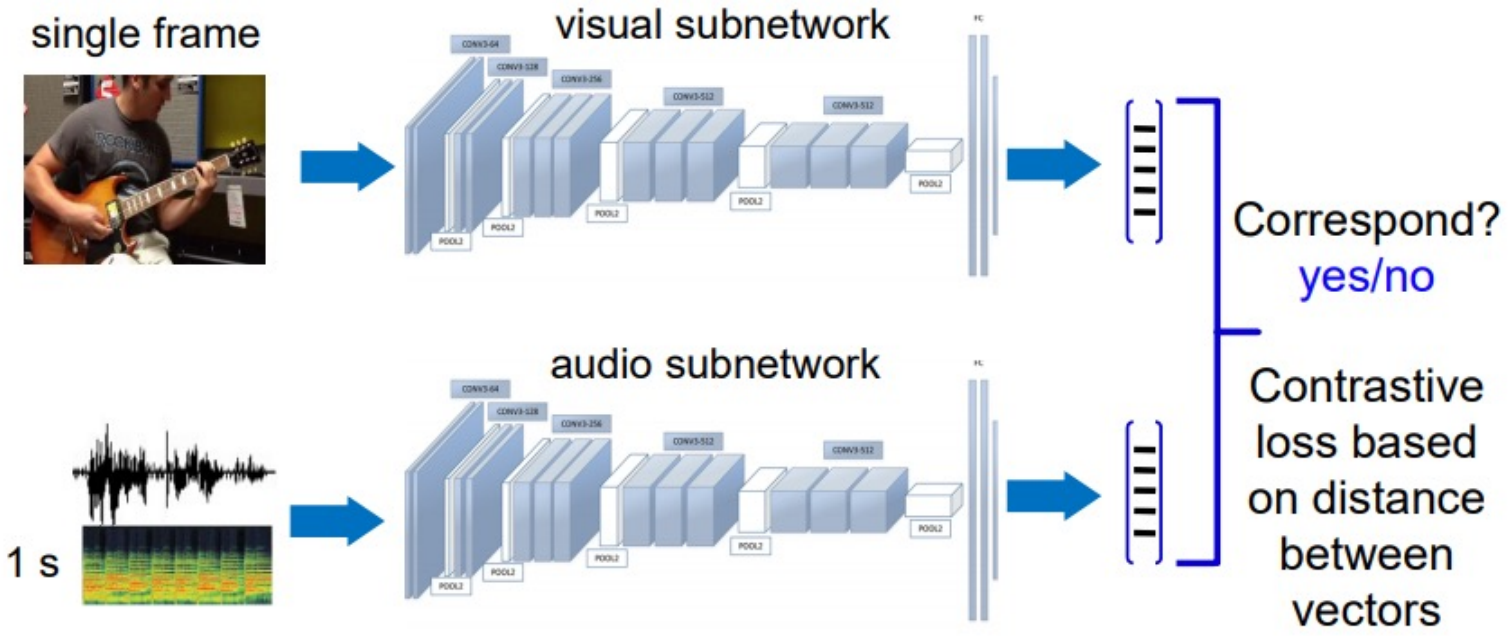
Multimodal SSL: A+V

Carnegie Mellon

- Train a network to predict if image and audio clip correspond
  - Arandjelović and Zisserman, “Objects that Sound” ECCV 18



Multimodal SSL: A+V



Distance between audio and visual vectors:

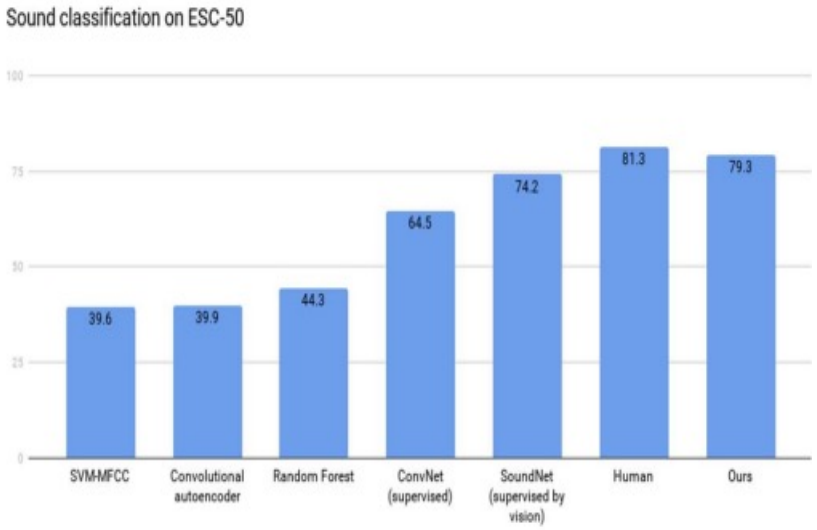
- **Small:** AV from the same place in a video (**Positives**)
- **Large:** AV from different videos (**Negatives**)

Train network from scratch

Multimodal SSL: A+V

Carnegie Mellon

- Sound classification on the ESC-50 dataset



  - Environmental sound classification
  - Use the net to extract features
  - Linear SVM

- ImageNet classification

Method	Top 1 accuracy
Random	18.3%
Pathak *et al.* [21]	22.3%
Krähenbühl *et al.* [14]	24.5%
Donahue *et al.* [7]	31.0%
Doersch *et al.* [6]	31.7%
Zhang *et al.* [34] (init: [14])	32.6%
Noroozi and Favaro [18]	34.7%
Ours random	12.9%
Ours	32.3%

Carnegie Mellon

Localizing objects with sound

Input: audio and video frame

Output: localization heatmap on frame

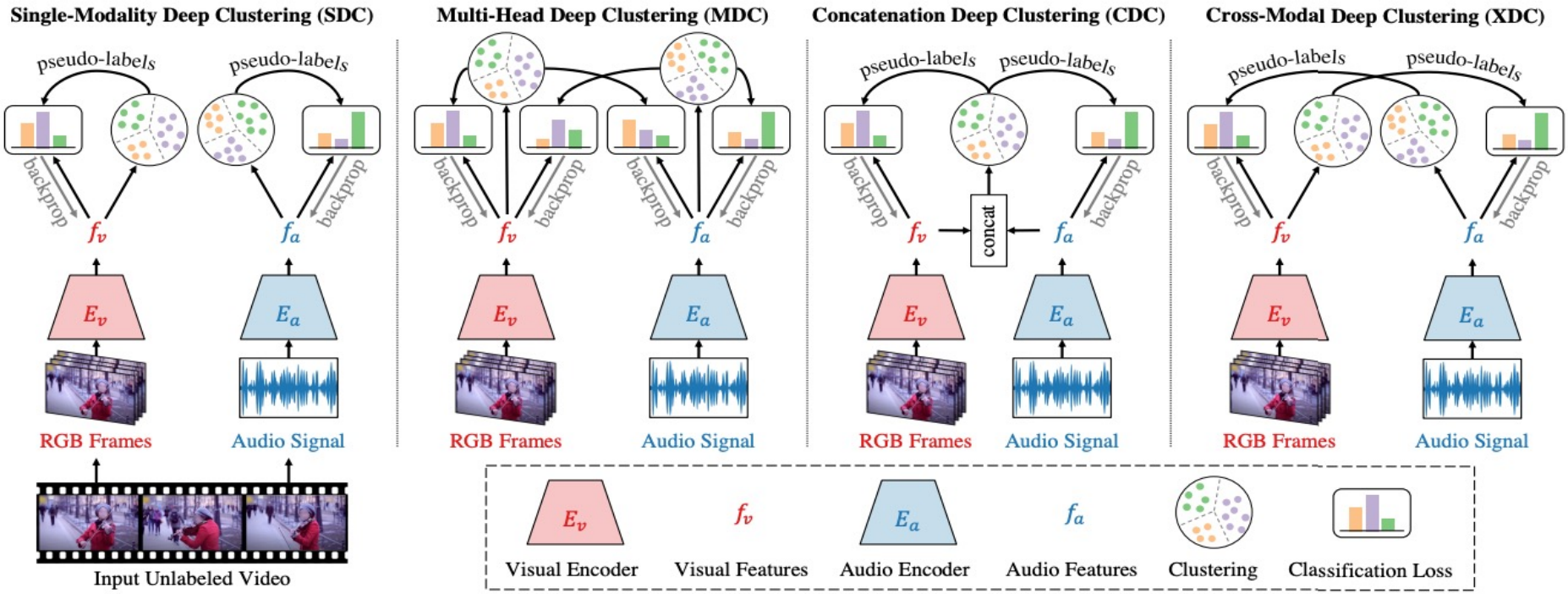
What would make this sound?

Note, no video (motion) information is used

Multimodal SSL: A+V

Carnegie Mellon

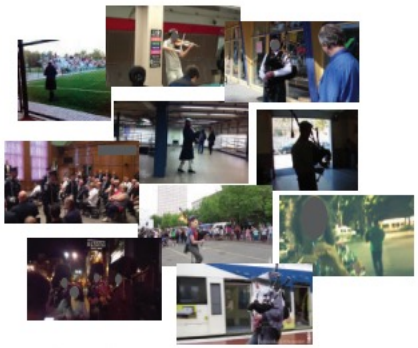
- Self-supervised learning with a global view by clustering:
  - Alwassel et al, “Self-Supervised Learning by Cross-Modal Audio-Video Clustering” NeurIPS 2020



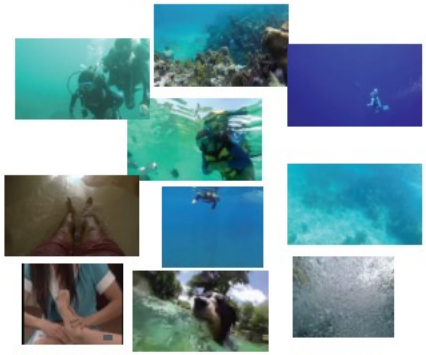
Multimodal SSL: A+V

Carnegie Mellon

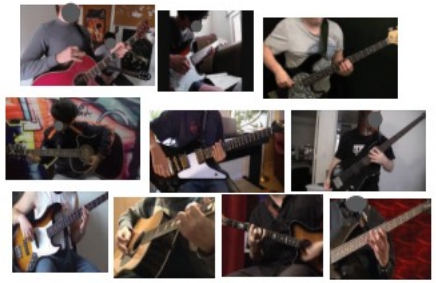
- Self-supervised learning with a global view by clustering:
  - Alwassel et al, “Self-Supervised Learning by Cross-Modal Audio-Video Clustering” NeurIPS 2020



audio cluster #125, purity: 0.70



audio cluster #105, purity: 0.33



video cluster #48, purity: 0.37

video cluster #27, purity: 0.36

Multimodal SSL: A+V

Carnegie Mellon

(a) Video action recognition.

Method	Pretraining Architecture	Pretraining Dataset	Evaluation UCF101	Evaluation HMDB51
ClipOrder [79]	R(2+1)D-18	UCF101	72.4	30.9
MotionPred [72]	C3D	Kinetics	61.2	33.4
ST-Puzzle [28]	3D-ResNet18	Kinetics	65.8	33.7
DPC [18]	3D-ResNet34	Kinetics	75.7	35.7
CBT [64]	S3D	Kinetics	79.5	44.6
SpeedNet [4]	S3D	Kinetics	81.1	48.8
AVTS [29]*	MC3-18	Kinetics	84.1	52.5
AVTS [29]†	R(2+1)D-18	Kinetics	86.2	52.3
XDC (ours)	R(2+1)D-18	Kinetics	86.8	52.6
AVTS [29]*	MC3-18	AudioSet	87.7	57.3
AVTS [29]†	R(2+1)D-18	AudioSet	89.1	58.1
XDC (ours)	R(2+1)D-18	AudioSet	93.0	63.7
MIL-NCE [38]	S3D	HowTo100M	91.3	61.0
ELo [50]	R(2+1)D-50	YouTube-8M	93.8	67.4
XDC (ours)	R(2+1)D-18	IG-Random	94.6	66.5
XDC (ours)	R(2+1)D-18	IG-Kinetics	**95.5**	**68.9**
Fully supervised	R(2+1)D-18	ImageNet	84.0	48.1
Fully supervised	R(2+1)D-18	Kinetics	94.2	65.1

(b) Audio event classification.

Method	ESC50
Random Forest [49]	44.3
Piczak ConvNet [48]	64.5
SoundNet [2]	74.2
L^3-Net [1]	79.3
AVTS [29]	82.3
ConvRBM [56]	**86.5**
XDC (AudioSet)	84.8
XDC (IG-Random)	85.4

Method	DCASE
RG [52]	69
LTT [35]	72
RNH [54]	77
Ensemble [61]	78
SoundNet [2]	88
L^3-Net [1]	93
AVTS [29]	94
XDC (AudioSet)	**95**
XDC (IG-Random)	**95**

Outline

Carnegie Mellon

- Single-modal self-supervised leaning
  - Image
  - Audio
  - Video
  - Text
- Multi-modal self-supervised learning
  - Audio and Video
  - Video and Text (ours ICLR 2021, NAACL 2021)

Multimodal SSL: V+T

- Learning from Video-Text association

By contrastive learning objectives:

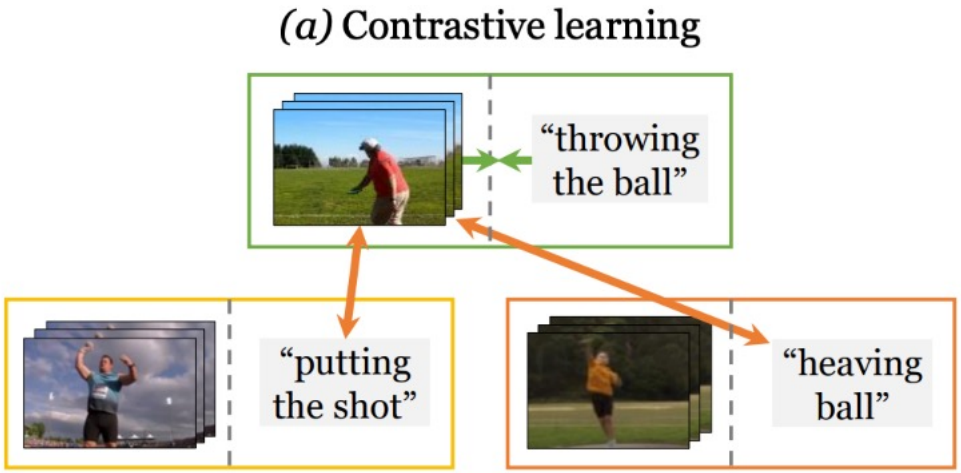
- Triplet
- NCE

HowTo100M
[Miech et al., ICCV 2019]

Multimodal SSL: V+T

Carnegie Mellon

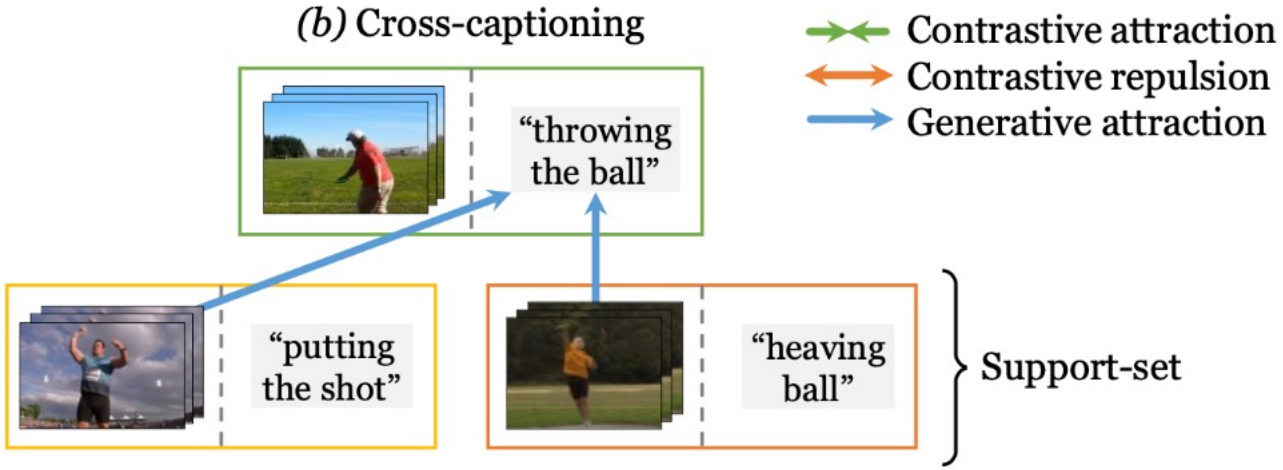
- Curse of Contrastive Learning: Instance Discrimination
  - All other instances in batch are considered negatives
  - Doesn't account for concept sharing or the similarities between instances.



Multimodal SSL: V+T

Carnegie Mellon

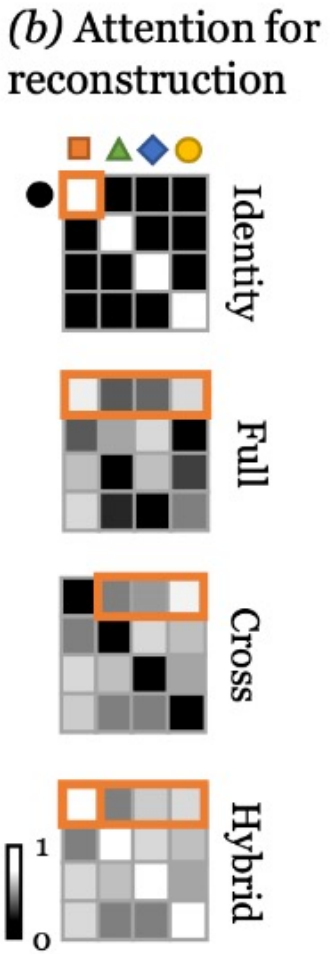
- Our idea: Cross-Instance Captioning from Support-Set
  - Generative objective: learn to reconstruct caption as weighted combination of videos in a support-set.
  - Implicitly pulls together videos with similar captions



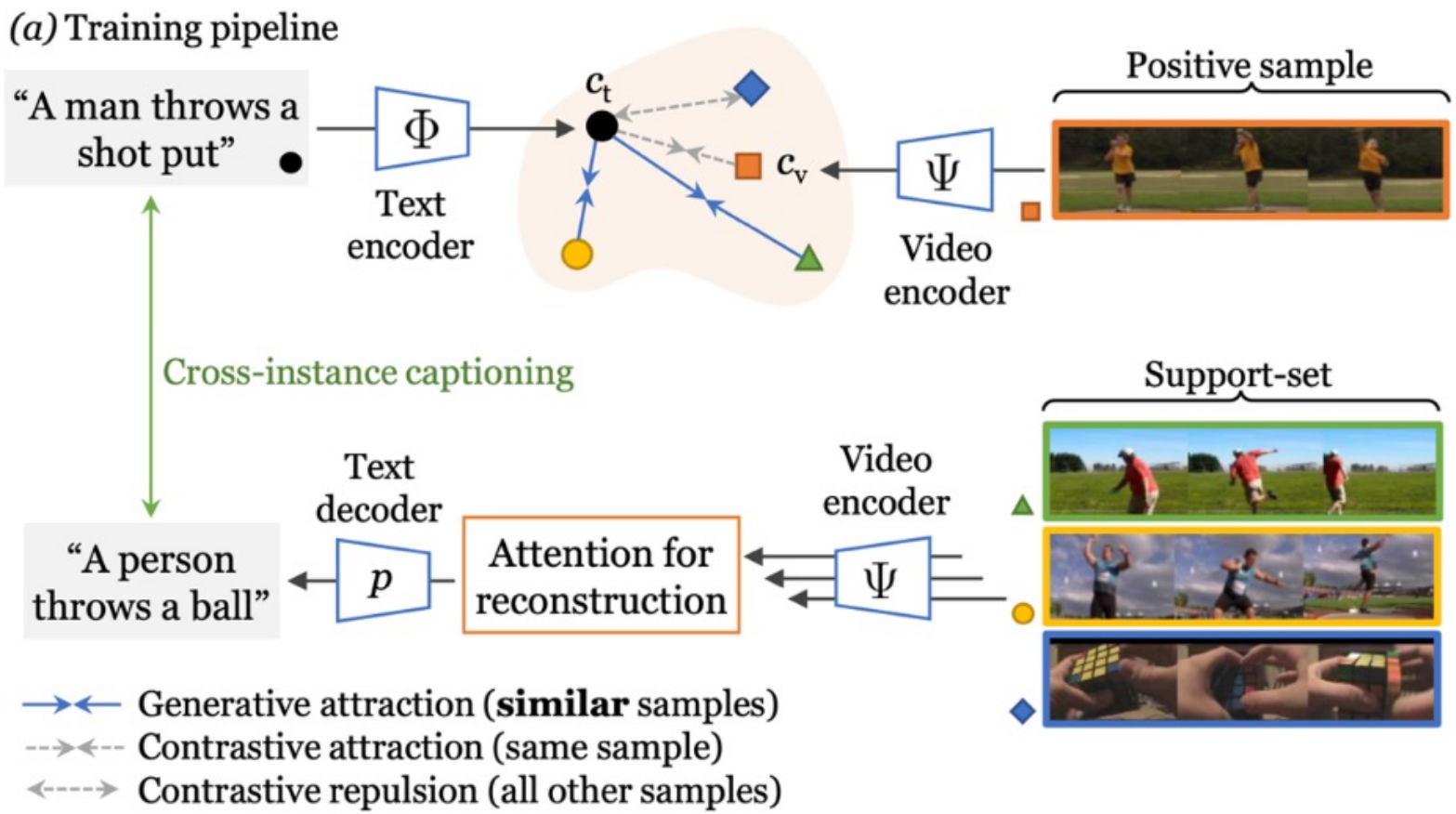
Multimodal SSL: V+T

Carnegie Mellon

- Attention for Support Set
  - None: No generative objective
  - Identity: generate caption from corresponding video
  - Full: generate caption from all videos in the support-set
  - Cross: generate caption from all but the video that one wishes to caption
  - Hybrid: average of full and identity captioning



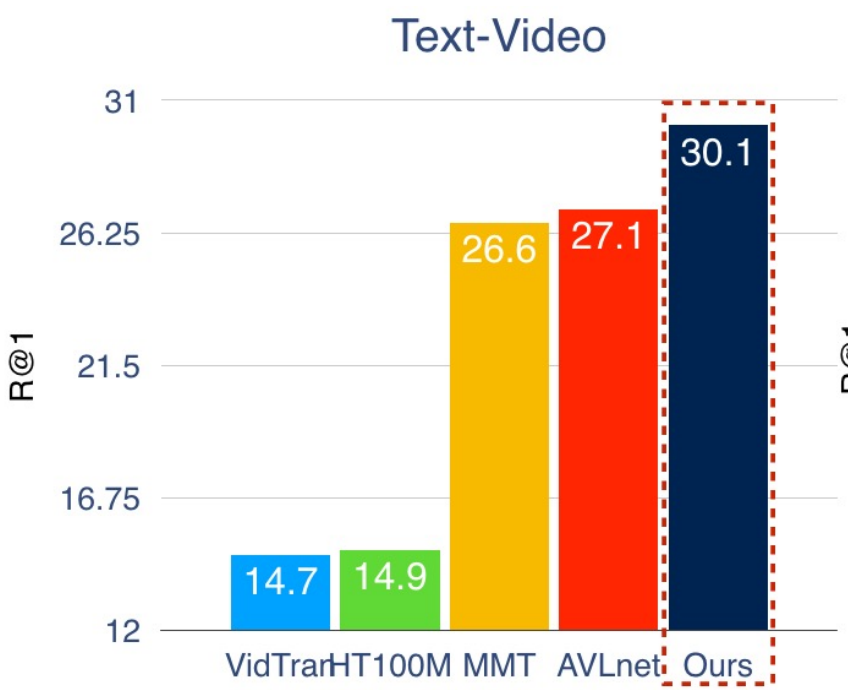
Multimodal SSL: V+T

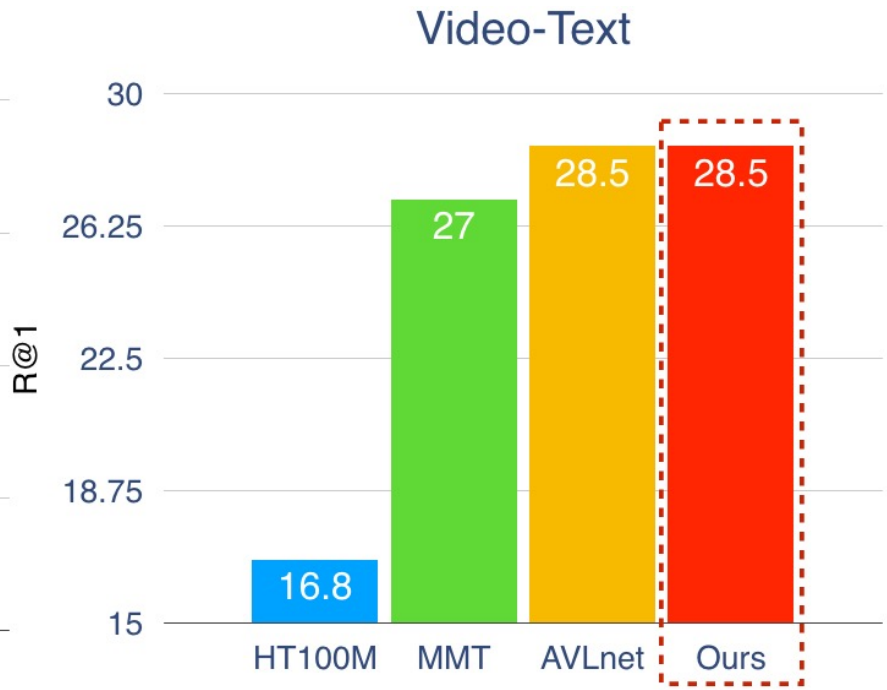


Multimodal SSL: V+T

Carnegie Mellon

- VTT

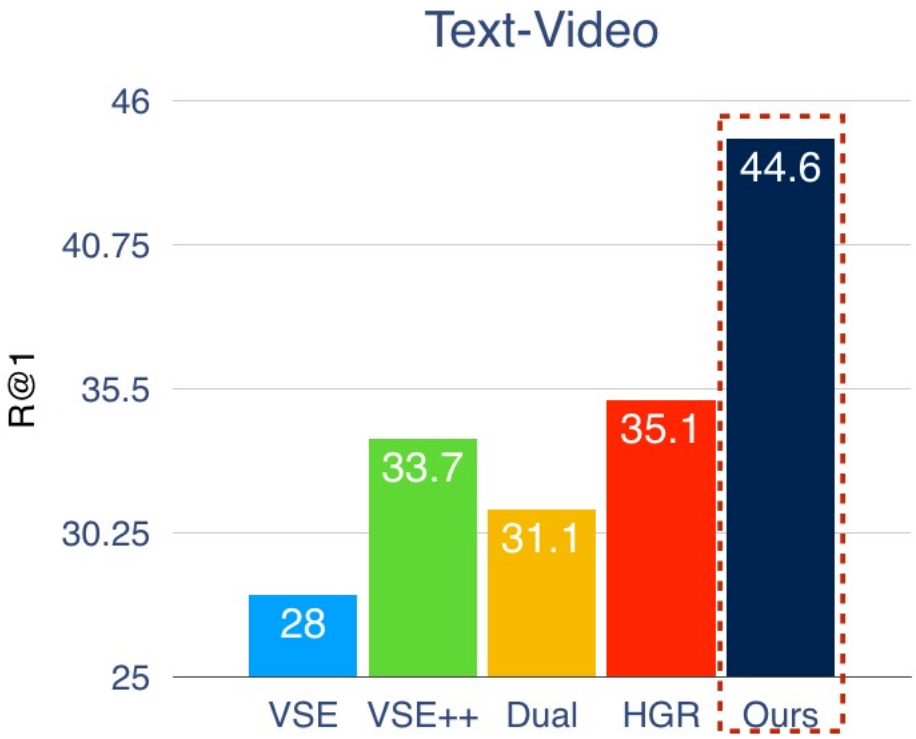




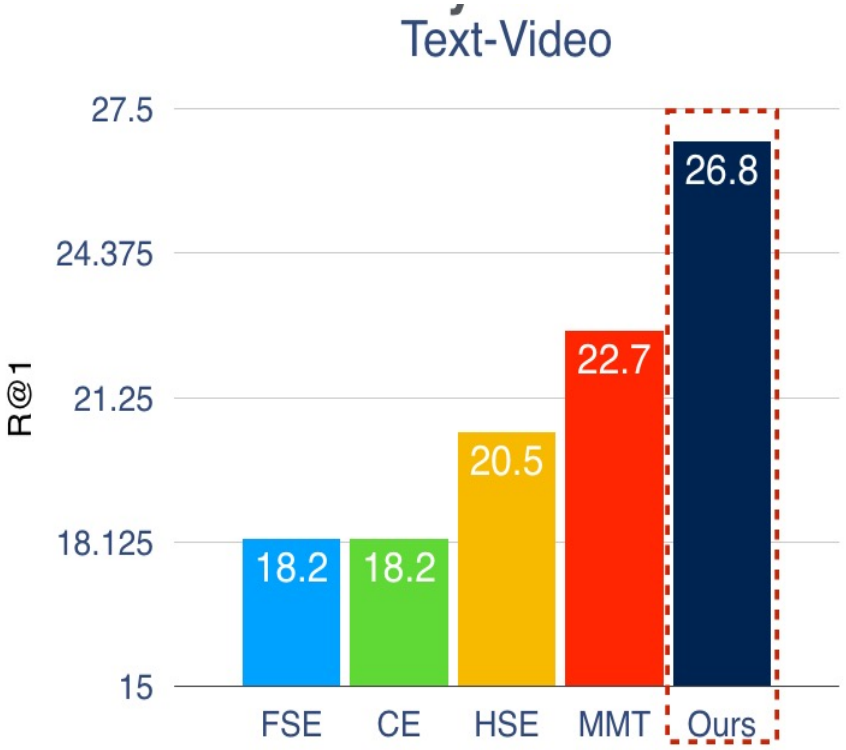
Multimodal SSL: V+T

Carnegie Mellon

- VATEX



ActivityNet



Multimodal SSL: V+T

Carnegie Mellon

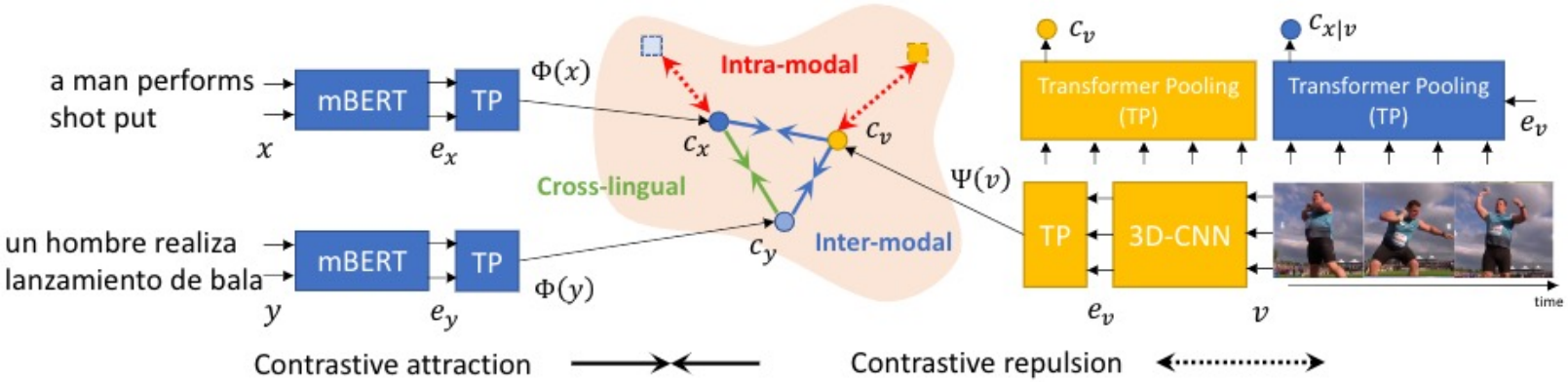
- More than 2 views?
  - Video + English + Chinese + Spanish + German + French ….
  - Multilingual Text-Video dataset (Multi-HowTo100M)

Multimodal SSL: V+T

Carnegie Mellon

- Multimodal NCE
  - Intra-modal NCE
  - Inter-modal NCE
  - Cross-lingual NCE

Multilingual BERT



Multimodal SSL: V+T

Carnegie Mellon

- Zero-Shot Performance on multilingual VTT

Model	*en*	*de*	*fr*	*cs*	*zh*	*ru*	*vi*	*sw*	*es*	Avg↑
mBERT	19.9	11.1	11.6	8.2	6.9	7.9	2.7	1.4	12.0	9.1
mBERT-MP	20.6	11.3	11.9	8.0	7.1	7.7	2.5	1.1	12.5	9.2
mBERT-MMP	21.8	15.0	15.8	11.2	8.4	11.0	3.7	3.4	15.1	11.7
XLM-R	21.0	16.3	17.4	16.0	14.9	15.4	7.7	5.7	17.3	14.7
XLM-R-MP	23.3	17.4	18.5	17.1	16.3	17.0	8.1	6.2	18.5	15.8
XLM-R-MMP	**23.8**	**19.4**	**20.7**	**19.3**	**18.2**	**19.1**	**8.2**	**8.4**	**20.4**	**17.5**
mBERT + translated VTT	19.6	18.2	18.0	16.9	16.2	16.5	8.4	13.0	18.5	16.1
mBERT-MMP + translated VTT	21.5	19.1	19.8	18.3	17.3	18.3	8.9	14.1	20.0	17.4
XLM-R + translated VTT	21.5	19.6	20.1	19.3	18.9	19.1	10.3	12.5	18.9	17.8
XLM-R-MMP + translated VTT	**23.1**	**21.1**	**21.8**	**20.7**	**20.0**	**20.5**	**10.9**	**14.4**	**21.9**	**19.4**

Multimodal SSL: V+T

Carnegie Mellon

a soccer team walking out on the field

Rank

1 

2 

3 

человек жонглирует палками на вершине заснеженной горы

1 

2 

3 

Drei Kinder singen zusammen auf der Stimme

1 

2 

3 

une personne nage dans des rapides d'eau vive

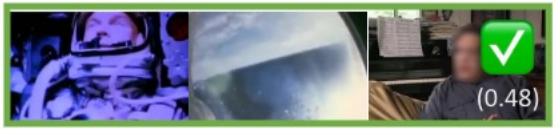
1 

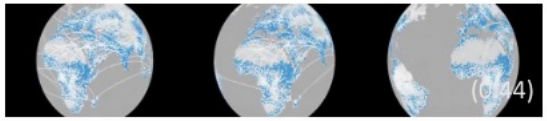
2 

3 

một người đàn ông đang nói về dự án không gian adam

1 

2 

3 

一个男人在麦克风说话

1 

2 

3 

Conlcusion

Carnegie Mellon

- Self-supervised learning
  - It's important to pick informative data for training
    - What are the positives? What are the negatives?
  - Design pretext task/ proxy loss wisely
- The trend of ML
  - -2010: unsupervised/ supervised learning
  - 2010-2015: supervised learning
  - 2015-2018: semi-supervised learning/ weakly labeled
  - 2018-2020: self-supervised learning
  - 2021- ??

Carnegie Mellon

Thank you