Representation Assemble!

Introduction to Multi-Modal Joint-Representation Learning

Jul 12, 2021

Sungguk Cha

MINDs Lab

TL; DR

- We introduce multi-modal joint-representation learning.
- We provide a research trend and applications of the field .

MINDs Lab

Contents

- My Talks
- Introduction
- Multi-Modal Joint-Representation Learning
- CVPR21!
- Conclusion

MINDs Lab

My Talks

- ➔ Zero-Shot Semantic Segmentation
  via Spatial and Multi-Scale Aware Visual Class Embedding
  Nov 23, 2020

- ➔ How Can We Correlate Inter-Domain Knowledge?
  Reviewing Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation
  Dec 21, 2020

- ➔ Representation Learning: How Should Feature Be Learned in Vision?
  Introducing CLIP and an Unsupervised Semantic Segmentation Approach
  Feb 22, 2021

- ➔ Are the Relationships of Class Representations in Vision and Language Similar?
  Introducing an Experiment in SM-VCENet
  Mar 15, 2021

- ➔ **Representation Assemble!**
  Introduction to Multi-Modal Learning
  Jul 12, 2021

MINDs Lab

Introduction

➔ **What** is multi-modal learning?

➔ **Why** multi-modal joint-representation learning?

MINDs Lab

Introduction:

Awesome-multimodal!

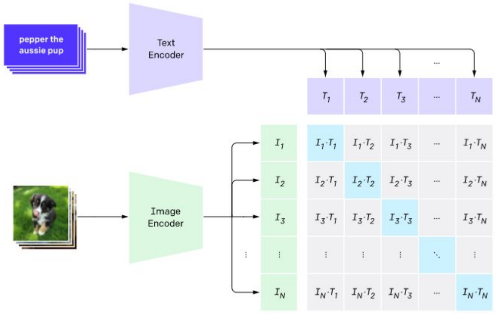
Recently, multimodal learning gains **popularity**.

MINDs Lab

Introduction:

Awesome-multimodal!

Recently, multimodal learning gains **popularity**.



(a) CLIP

(b) DALL E

What color are her eyes?
What is the mustache made of?

(c) VQA

Image sources
OpenAI/CLIP
OpenAI/DALLE
VQA

Introduction:
Multi-modal learning

Multi-modal learning is to **leverage multiple modalities**.
- *e.g., computer vision, language and audio*
- *autonomous driving: image, LiDAR*

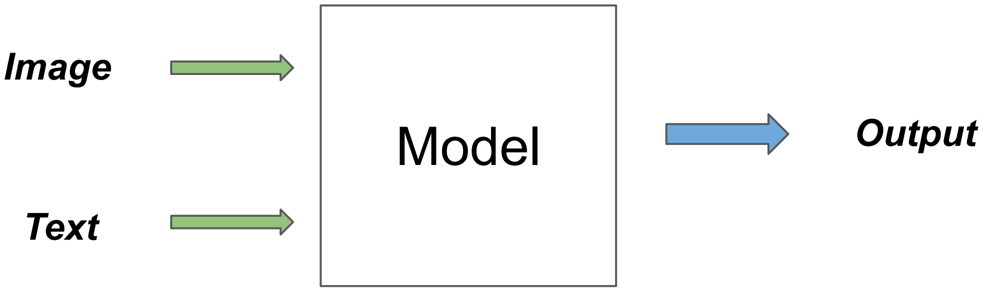


Figure: Multi-modal approach example.

Introduction:

Multi-modal learning examples

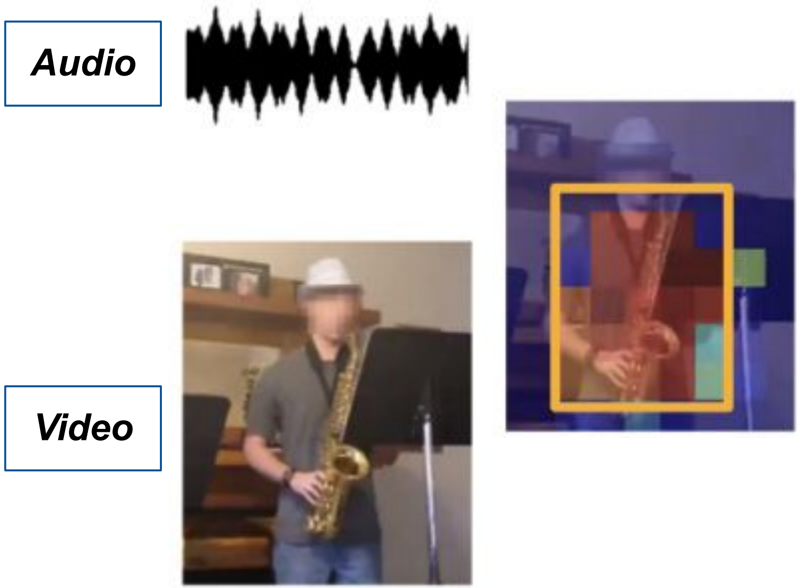


Figure: Visual recognition with audio

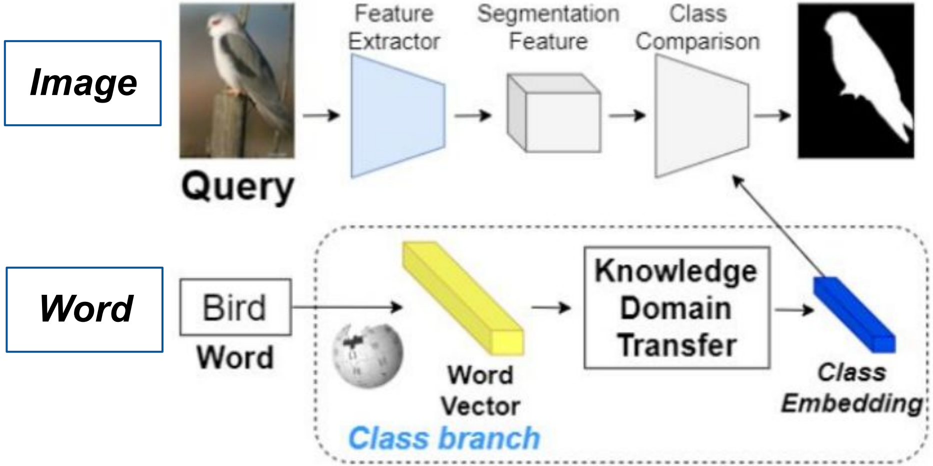


Figure: Zero-shot semantic segmentation

Image source
left figure: Self-supervised object detection from audio-visual correspondence, Afouras and Asano et al.
right figure: Sungguk Cha

Introduction:

Modality interaction is **challenging**

MINDs Lab

Introduction:

Modality interaction is challenging

- Imagine a framework that *image encoder* and *text encoder* are **independently trained**.
- E.g., **ImageNet pretrained ResNet** and **pretrained BERT**

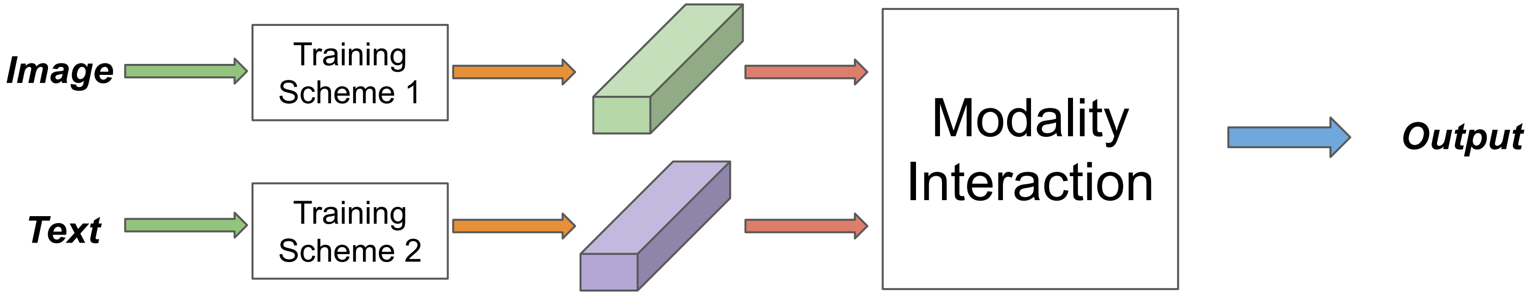


Figure: Multi-modal approach with independent training schemes.

MINDs Lab

Introduction:
Modality interaction is challenging

- It is like "***suddenly** marrying two people from different cultures*".

- They will suffer from
    - different culture (semantic)
    - different language (representation)
    - different knowledge
    - and so on.

MINDs Lab

Introduction:

Modality interaction is challenging

- It will be like suddenly marrying two people from different cultures.
- They will suffer from
    - different culture (semantic)
    - different language (representation)
    - different knowledge
    - and so on.

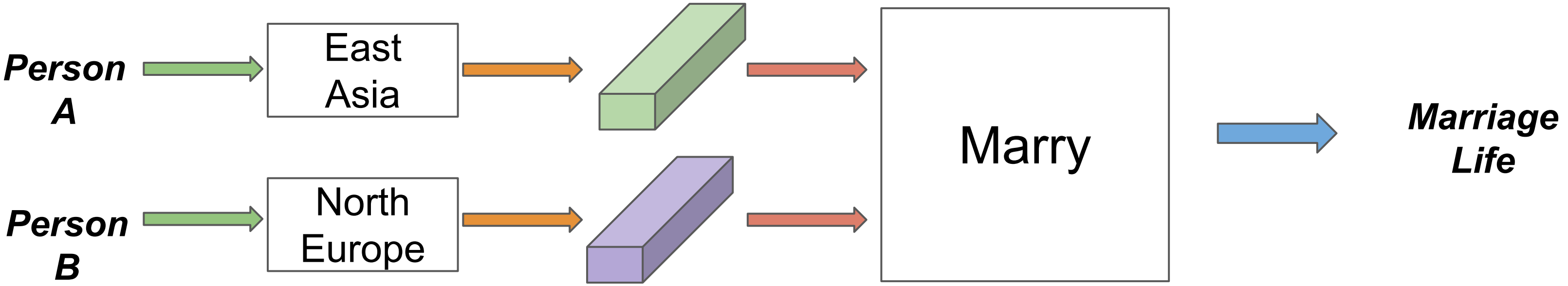


Figure: Multi-culture marriage example

Introduction:

Modality interaction is challenging

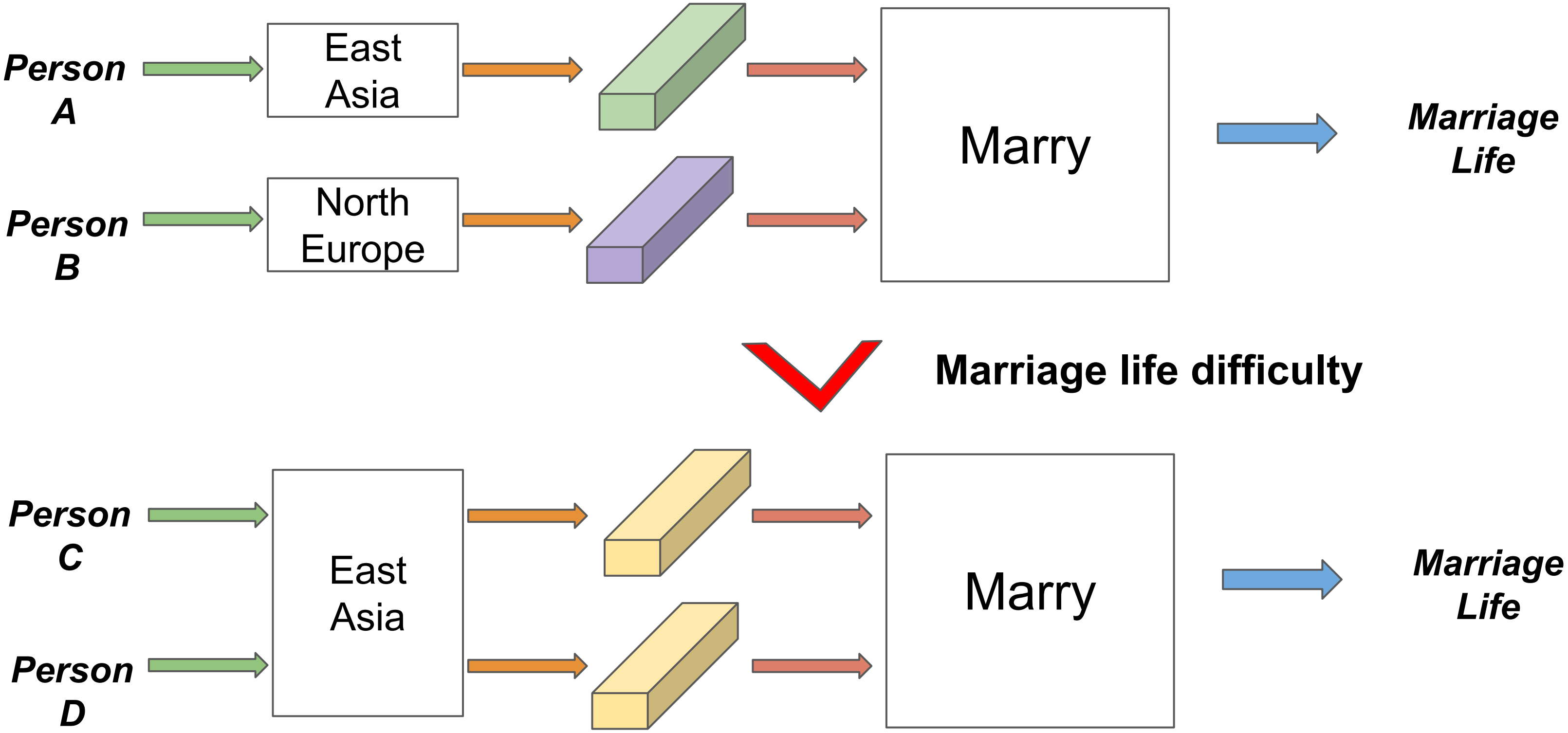


Figure: Marriage life difficulty comparison between multicultural and single-culture marriage. If they share the same background (culture), the marriage life will be easier.

MINDs Lab

Introduction:

Modality interaction is challenging

Having different training scheme results

- **different** representation, semantic and knowledge.

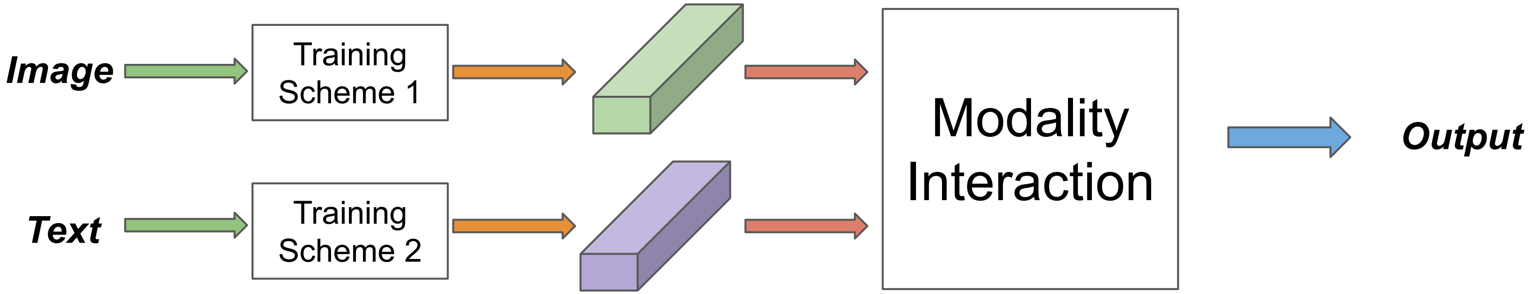


Figure: Multi-modal approach with independent training schemes.

Introduction:

Modality interaction is challenging

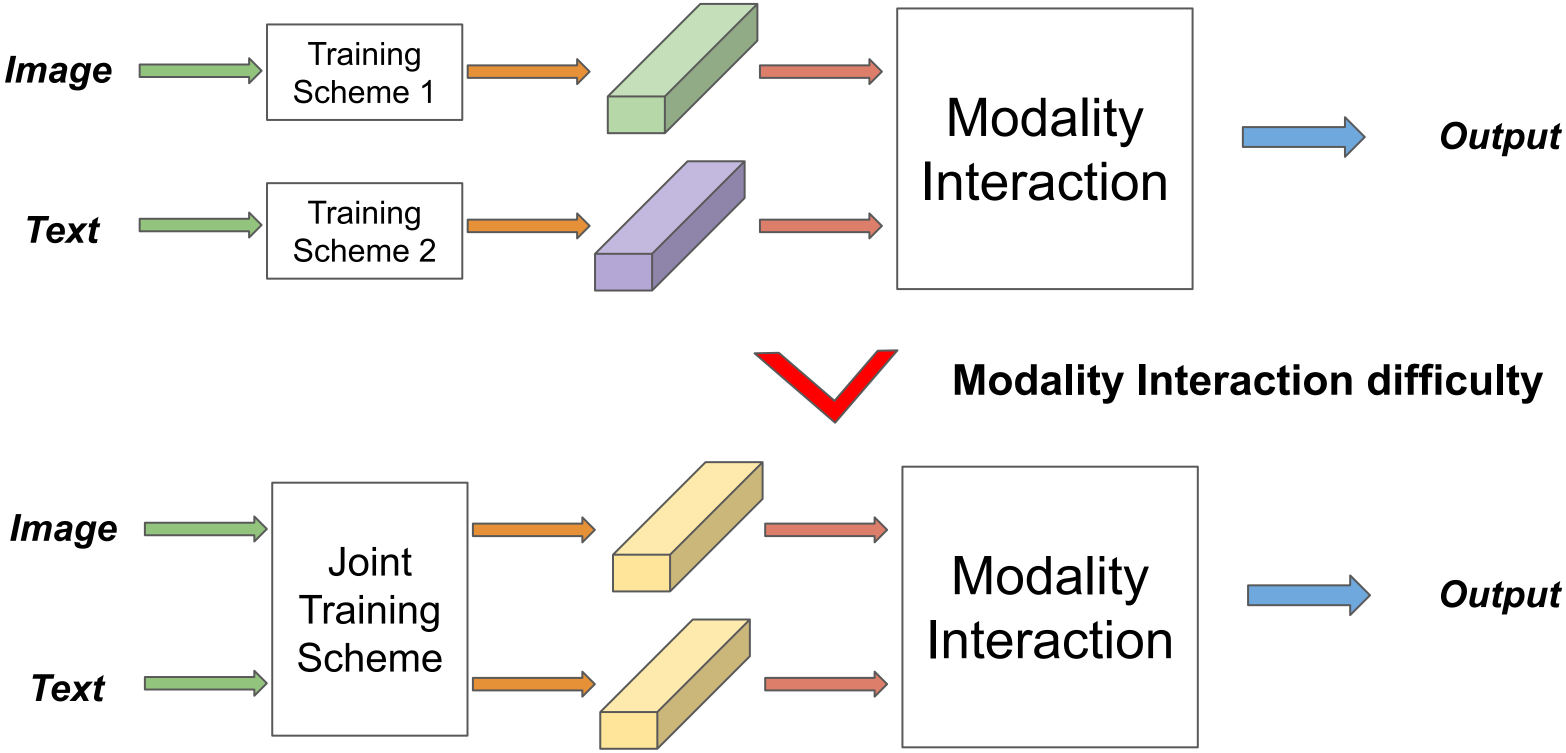


Figure: Modality interaction difficulty comparison between joint training scheme and independent training scheme.
If they share the same training scheme, modality interaction will be easier.

MINDs Lab

Introduction

Instead of solving modality interaction problem,

we choose joint-training scheme.

Introduction:

In this talk,

➔ We introduce multi-modal joint-representation learning.

➔ We provide a research trend and applications of the field .

MINDs Lab

Multi-Modal Joint-Representation Learning

Joint-representation learning is based on **contrastive learning**.

MINDs Lab

Multi-Modal Joint-Representation Learning

Joint-representation learning is based on contrastive learning.

Supervising an encoder to encode

the same things into **the same representations,**

different things into **different representations**.

Multi-modal Joint-Representation Learning:

Example, classification model

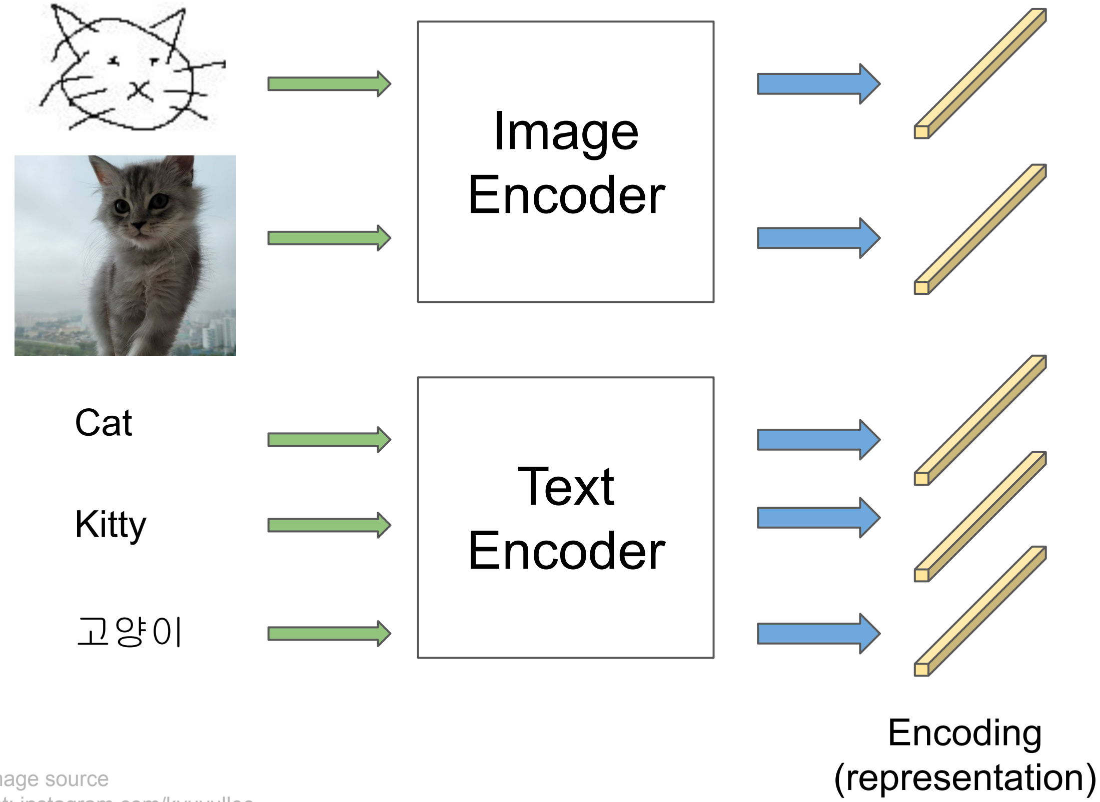
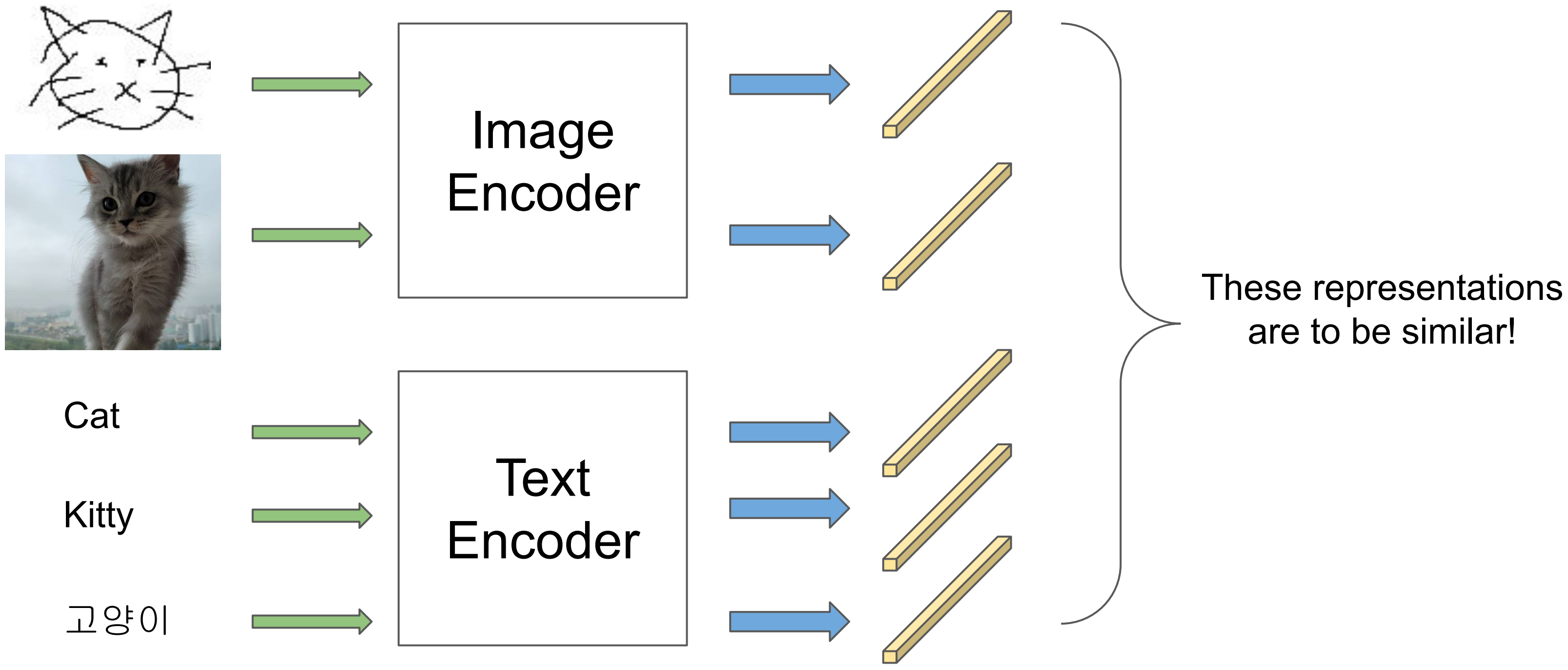


Image source
cat: instagram.com/kyuyullee

Multi-modal Joint-Representation Learning:

Example, classification model

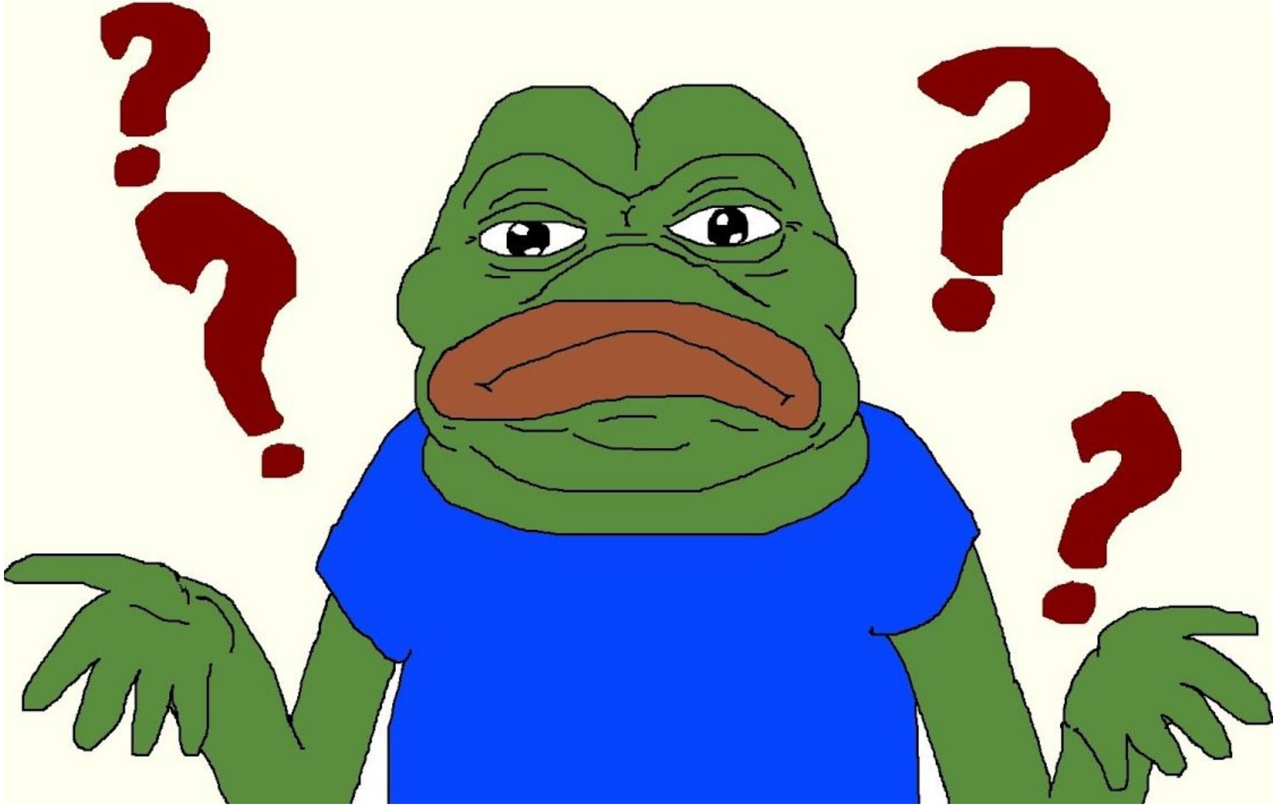


Encoding (representation)

Image source
cat: instagram.com/kyuyullee

MINDs Lab

How???



Joint-Representation Learning Supervisions

There are several approaches.

Joint-Representation Learning Supervisions

There are several approaches.

Mostly self-supervised.

- Labeling costs much.
- Datasets are coarse.

Joint-Representation Learning Supervisions

There are two major approaches.

- Clustering based
- Transformer supervisions

MINDs Lab

Clustering Based Learning

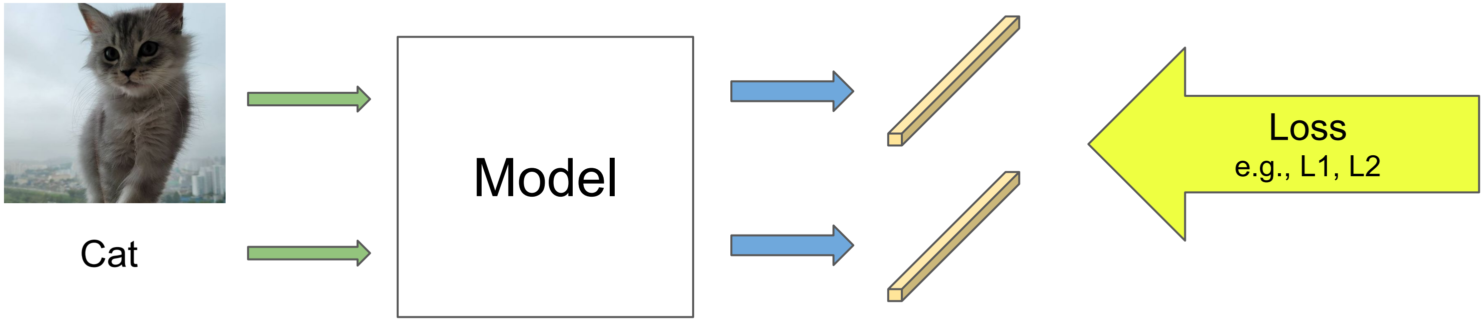
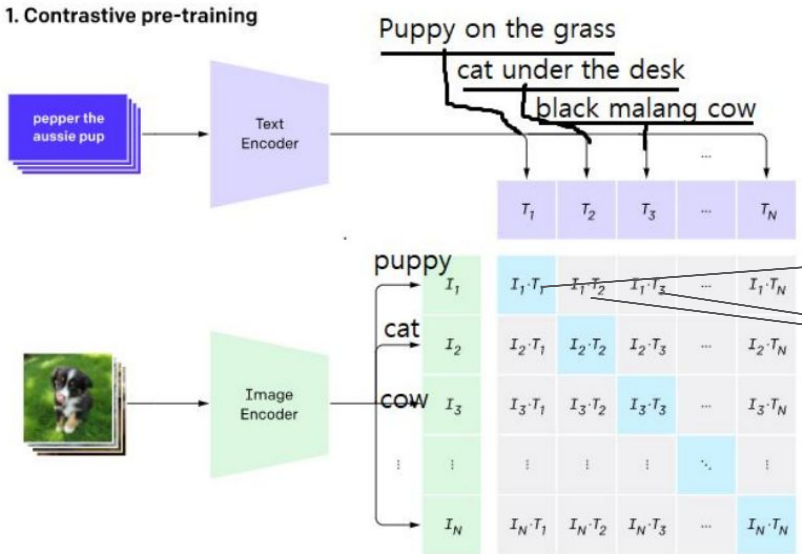


Figure: Joint Contrastive Learning:
Making representations the same.

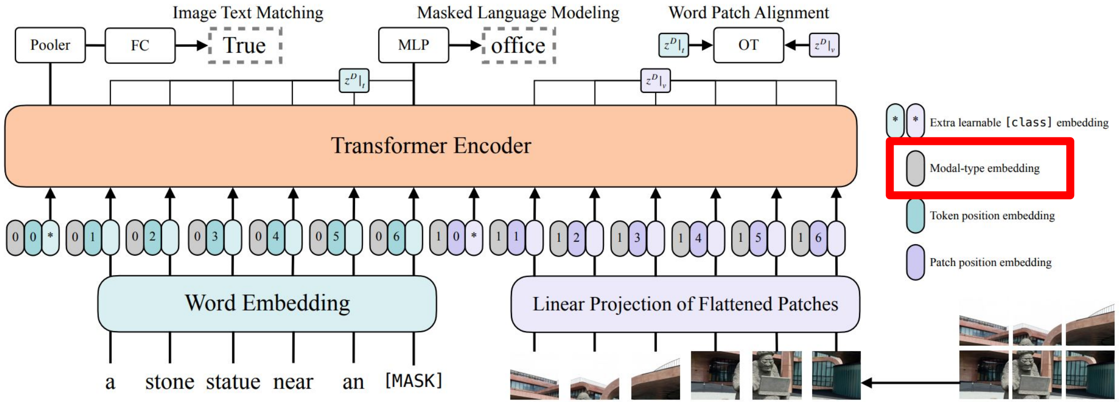
Clustering Based Learning



- Maximize a cosine similarity between the pair representations
- Minimize the similarity between the independent representations

Figure: Joint Contrastive Learning:
supervision with similarity
Note, it is very popular

Transformer Supervisions



- Uni-framework
- Image Text Matching loss
- Masked Language Modeling
- Word Patch Alignment

Image source
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, Wonjae Kim, Bokyung Son and Ildoo Kim, ICML2021

CVPR21 Papers!

- Vx2Text: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs
- Cross-Modal Contrastive Learning for Text-to-Image Generation
- Audio-Visual Instance Discrimination with Cross-Modal Agreement
- M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-Training
- Multimodal Contrastive Training for Visual Representation Learning

MINDs Lab

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
FACEBOOK AI
Georgia Tech
Dartmouth

VX2TEXT: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs

Xudong Lin[1], Gedas Bertasius[2], Jue Wang[2], Shih-Fu Chang[1], Devi Parikh[2,3], Lorenzo Torresani[2,4]
[1]Columbia University [2]Facebook AI [3]Georgia Tech [4]Dartmouth

CVPR VIRTUAL JUNE 19-25

Motivation

- Build an AI for *"video+x to text"* tasks:
  - Effectively extract and fuse information from video and other modalities;
  - Generate texts to interact with humans.

	Representation of Video and Other Modalities	Multimodal Fusion
HERO, MTN, etc.	Continuous features	Fusion modules (pretrained by multimodal pretext tasks)
VX2TEXT (Ours)	Symbolic text tokens	Text transformers

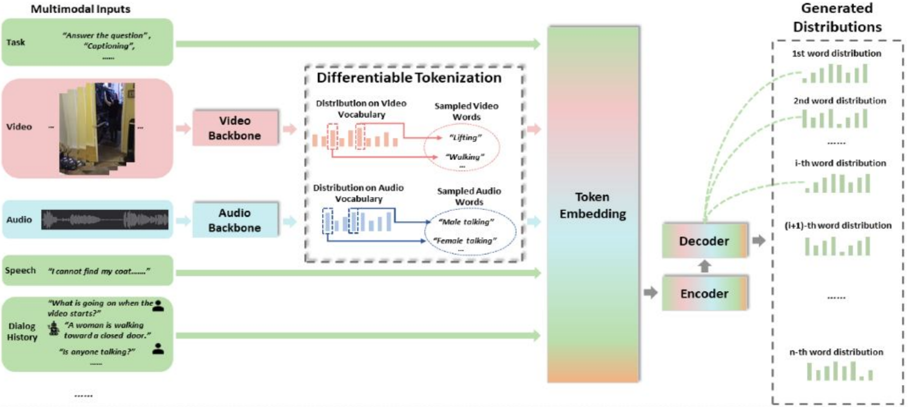
Our Insights

- Symbolic text tokens can effectively describe key information in video and other modalities.
- Powerful pretrained language models can fuse symbolic multimodal tokens and generate desired texts.

Our Contribution

- Differentiable Tokenization that addresses the non-differentiability of tokenization on continuous inputs (e.g., video or audio) and enables end-to-end training.
- State-of-the-art on three video-based text-generation tasks:
  - Question answering
  - Audio-visual scene-aware dialog
  - Captioning

Technical Approach



- Given video and other modalities as input,
  - Use modality-specific pretrained networks to classify them into predefined categories;
  - Differentiable Tokenization uses Gumbel-Softmax trick to enable end-to-end training;

$$G \approx \nabla_{\mathbf{w}_m} \frac{\exp(\log p_m(c|\mathbf{x}) + g_m(c))}{\sum_{c'}^{|C_m|} \exp(\log p_m(c'|\mathbf{x}) + g_m(c'))}$$

- An encoder-decoder architecture performs multimodal fusion in the language space and generates desired texts.

Experimental Results

- Video Question Answering: TVQA

Models	# Samples for Multimodal Pretext	Val	Test
HERO [29]	7.6M	74.8	73.6
TVQA [26]	0	67.7	68.5
STAGE [27]	0	70.5	70.2
HERO [29]	0	70.7	70.3
MSAN [20]	0	71.6	71.1
BERT QA [52]	0	72.4	72.7
VX2TEXT (Ours)	0	**74.9**	**75.0**

- Audio-visual scene-aware dialog: AVSD

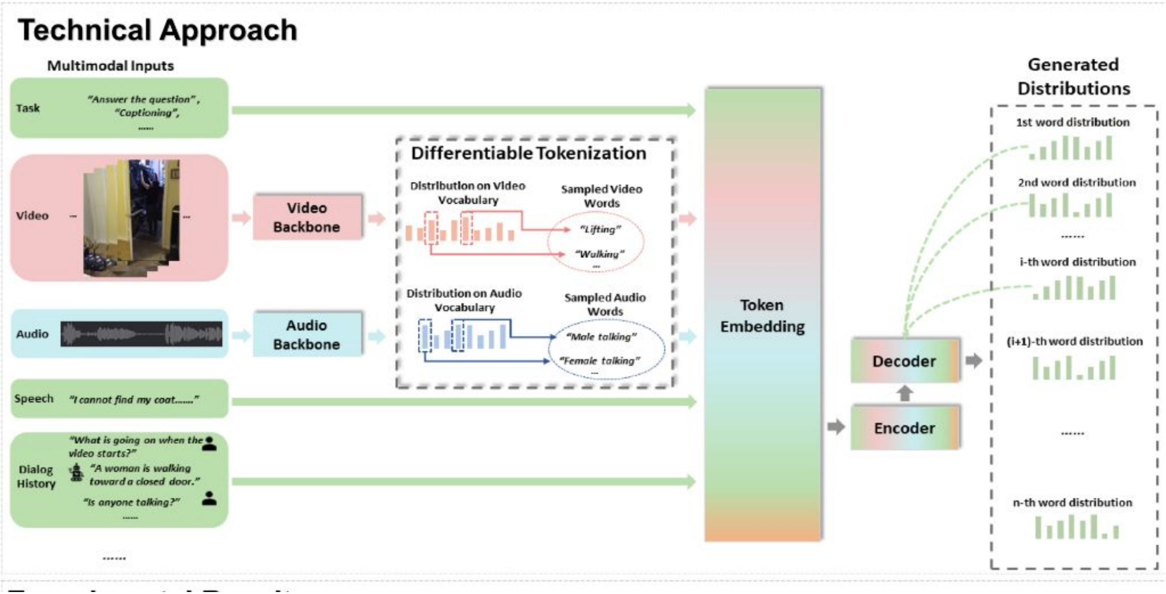
Models	Use Caption?	CIDERr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR
MA-VDS [17]	No	0.727	0.078	0.109	0.161	0.256	0.277	0.113
Simple [41]	No	0.905	0.095	0.130	0.183	0.279	0.303	0.122
VX2TEXT (Ours)	No	**1.357**	**0.127**	**0.166**	**0.222**	**0.317**	**0.356**	**0.152**
MTN [25]	Yes	1.249	0.128	0.173	0.241	0.357	0.355	0.162
MTN-TMT [30]	Yes	1.357	0.142	-	-	-	0.371	0.171
VX2TEXT (Ours)	Yes	**1.605**	**0.154**	**0.197**	**0.260**	**0.361**	**0.393**	**0.178**

- Video Captioning: TVC

Models	# Samples for Multimodal Pretext	Test			
		CIDERr	BLEU-4	ROUGE-L	METEOR
HERO [29]	7.6M	0.500	0.124	0.342	0.176
MMT [28]	0	0.454	0.109	0.328	0.169
HERO [29]	0	0.437	0.109	0.326	0.165
VX2TEXT (Ours)	0	**0.483**	**0.119**	**0.331**	**0.174**

- Interpretation of Video Tokens on TVC
- Meaningful semantics of video tokens after the end-to-end training process

MINDs Lab



Input

- Task
- Video
- Audio
- Speech
- Dialog

Applications

- Video Question Answering
- Audio-Visual Scene Aware **Dialog**
- Captioning

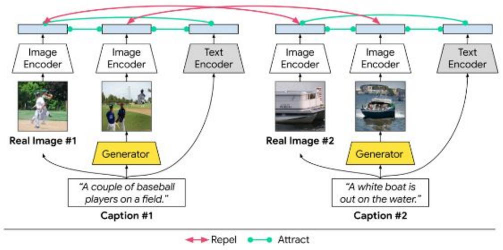
Google Research

Cross-Modal Contrastive Learning for Text-to-Image Generation

Han Zhang*, Jing Yu Koh*, Jason Baldridge, Honglak Lee, Yinfei Yang

CVPR VIRTUAL JUNE 19-25

XMC-GAN Overview



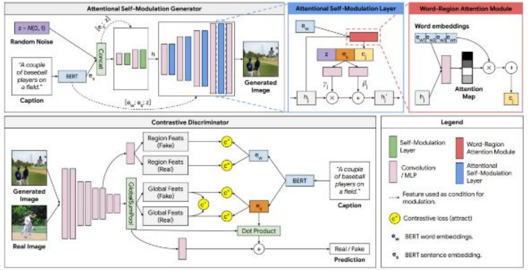
A simple **one-stage** GAN *without* object-level annotation can outperform prior object-driven and multi-stage approaches.

The proposed cross-modal losses maximize the **mutual information** between image-text pairs through **contrastive losses**.

Contrastive losses are used to train both the *discriminator (D)* and *Generator (G)*

- Train *D* to learn more robust and discriminative feature, **less prone** to mode collapse.
- Train *G* to **enforce the consistency** between generated images and conditional text descriptions.

Model / Ablation



Maximize the mutual information between the corresponding pairs:

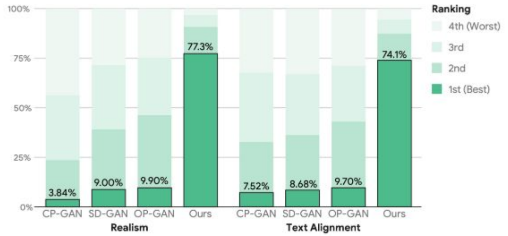
(1) image and sentence (**S**)
(2) generated / real image with the same description (**I**)
(3) image regions and words (**W**)

S	W	I	IS ↑	FID ↓	R-prec ↑	SOA-C ↑	SOA-I ↑
Real Images [17]			34.88	6.09	69.36	76.17	80.12
			15.89	39.28	21.41	8.99	25.72
✓			23.50	19.25	53.57	24.57	45.41
	✓		20.72	24.38	44.42	20.50	39.12
		D	18.90	29.71	31.16	12.73	30.89
		VGG	21.54	39.58	35.89	17.41	35.08
		D + VGG	23.61	21.14	47.04	23.87	44.41
✓	✓		26.02	14.25	64.94	30.49	51.60
✓	✓	D	28.06	12.96	65.36	34.21	54.23
✓	✓	VGG	30.55	**11.12**	**70.98**	39.36	59.10
✓	✓	D + VGG	**30.66**	11.93	69.86	**39.85**	**59.78**

All three cross-modal contrastive pairs are important.

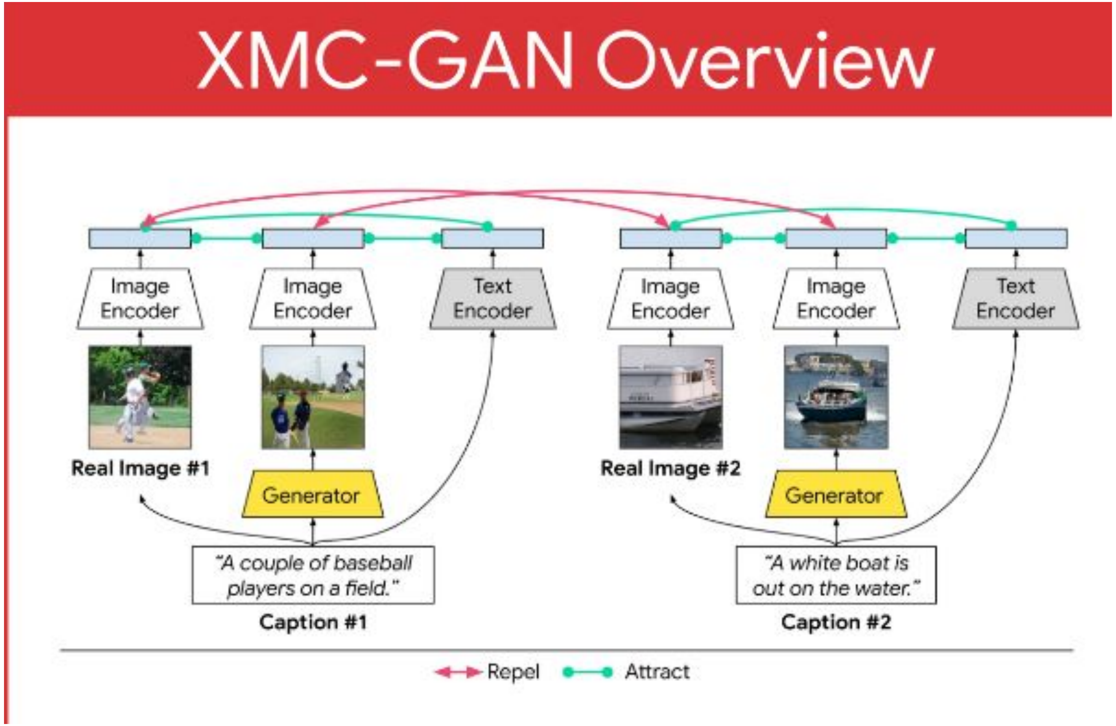
Evaluation

XMC-GAN generated images depict clearer scenes and objects as compared to previous SOTA approaches.



In large scale human evaluations (1000 independent annotators), XMC-GAN generated images are significantly preferred.

MINDs Lab



- Cross modality contrastive learning on Text-to-Image GAN
- Attract (pull)
  (real image, fake image, text)
- Repel (push)
  (real image vs real image)
  (fake image vs fake image)

Audio-Visual Instance Discrimination with Cross-Modal Agreement

Pedro Morgado — UC San Diego
Nuno Vasconcelos — UC San Diego
Ishan Misra — Facebook AI Research

CVPR VIRTUAL JUNE 14-19

1 Summary

Audio-Visual Instance Discrimination (AVID)

- Self-supervised framework to learn audio and video representations.
- AVID seeks to identify audio-video pairs originating from the same instance from a large set of options.
- Cross-modal discrimination, as opposed to within-modal discrimination, is crucial for learning audio and video representations that transfer well to action recognition and environmental sound classification tasks.

Cons of AVID

- Positive sets limited to audio-video pairs from the same instances.
- Negative sets contain instances from semantically related instances.
- Within modal similarities are left unconstrained.

Positive Expansion by Cross-Modal Agreement (CMA)

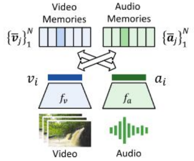
- CMA identifies which instances are similar in both audio and visual space to form more accurate and diverse positive sets.
- Within-modal discrimination of positive sets calibrates within-modal similarities and improve performance on downstream tasks.

2 Audio-Visual Instance Discrimination

Overview

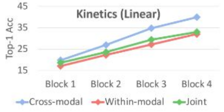
- 0.5 sec video and 2s audio signals extracted from each instance.
- Neural network encoders extract video and audio features, independently.
- Slow moving representations maintained in memory banks.
- Cross-Modal Contrastive NCE Loss [4,5]

$$L_{AVID} = L_{NCE}(v_i \to \bar{a}_i) + L_{NCE}(a_i \to \bar{v}_i)$$



Is cross-modal supervision critical for learning good representations?

- We compared cross-modal to within-modal and joint supervision.
- Cross-modal supervision outperforms others by significant margins.

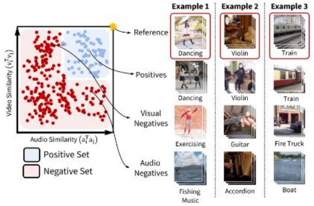


3 Cross-Modal Agreement

Goal: Expand positive set beyond the instance itself and calibrate within modal similarities.

Procedure:

1. Start from an AVID pre-trained model
2. Compute pairwise agreement scores: $\rho_{ij} = \min(v_i^T v_j, a_i^T a_j)$
3. Define positive set as $P_i = TopK_j(\rho_{ij})$ and negative set as $N_i = D \setminus P_i$
4. Optimize $L_{CMA} = \sum_{j \in P_i} L_{NCE}(v_i \to \bar{v}_j) + L_{NCE}(a_i \to \bar{a}_j)$



CMA finds instances with both visual and audio similarity.

Visually similar instances with low audio similarity can still be sampled as hard visual negatives (and vice-versa).

4 Experiments and results

Downstream tasks

- Action recognition on UCF and HMDB datasets.
- Environmental Sound Classification on ESC and DCASE datasets.

Pre-training DB – Kinetics (240K videos)

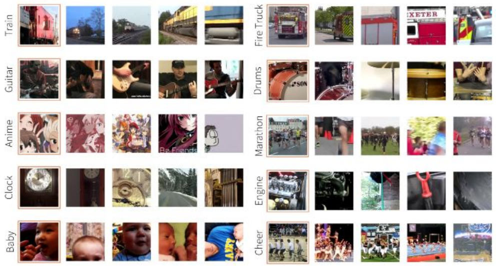
	UCF	HMDB	ESC	DCASE
L3 [1]	74.4	47.8	-	-
AVTS [2]	85.8	56.9	76.7	91
XDC [3]	86.8	52.6	78.5	-
AVID	86.9	59.9	77.6	93.0
AVID-CMA	87.5	60.8	79.1	93.0

Pre-training DB – Audioset (2M videos)

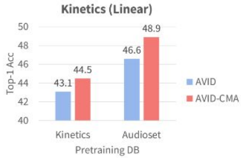
	UCF	HMDB	ESC	DCASE
L3 [1]	82.3	51.6	79.3	93.0
AVTS [2]	89.0	61.6	76.7	91.0
XDC [3]	93.0	63.7	85.8	-
AVID	91.0	64.1	89.2	96.0
AVID-CMA	91.5	64.7	89.1	96.0

Visual nearest neighbors obtained from a model trained by AVID-CMA

Semantically similar videos are grouped together despite their visual diversity.



CMA enhances visual representations for action recognition, as shown by the gains on the downstream linear classification task on Kinetics.



5 References

[1] Arandjelovic, Zisserman. "Look, listen and learn." CVPR, 2017.

[2] Korbar, Tran, Torresani. "Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization." NeurIPS, 2018.

[3] Alwassel, Mahajan, Korbar, Torresani, Ghanem, Tran. "Self-supervised learning by cross-modal audio-video clustering." NeurIPS, 2020.

[4] Michael Gutmann and Aapo Hyva ̈rinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models." ICAIS, 2010

[5] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv, 2018.

[S1] Morgado, Misra, Vasconcelos, "Robust Cross-Modal Instance Discrimination", CVPR, 2021.

6 Code

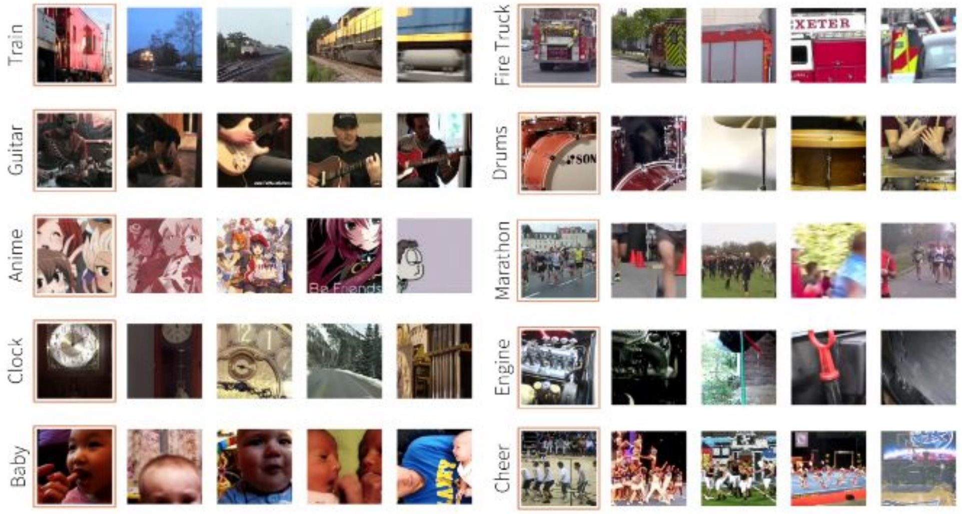
https://github.com/facebook research/AVID-CMA

SCAN ME

MINDs Lab

Visual nearest neighbors obtained from a model trained by AVID-CMA

Semantically similar videos are grouped together despite their visual diversity.



- Audio-Visual Instance Discrimination: matching video-audio pair
- The same abstract, the same representation.

MINDs Lab

M^{3}P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training

Minheng Ni†, Haoyang Huang†, Lin Su†, Edward Cui, Taroon Bhart, … Dongdong Zhang and Nan Duan

Harbin Institute of Technology Microsoft Corporation

CVPR VIRTUAL JUNE 19-25

Introduction

Recently, we witness the rise of a new paradigm of natural language processing (NLP), where general knowledge is learned from raw texts by self-supervised pre-training and then applied to downstream tasks by task-specific fine-tuning.

The multilingual pre-trained language models cannot handle vision data directly, whereas many pre-trained multimodal models are trained on English corpora thus cannot perform very well on non-English languages.

Moreover, relying on high-quality machine translation engines to generate such data from English multimodal corpora is both time-consuming and computationally expensive.

To address these challenges, this paper presents M3P, a Multitask Multilingual Multimodal Pre-trained model, which aims to learn universal representations that can map objects occurred in different modalities or texts expressed in different languages into a common semantic space.

M^{3}P: Multitask Multilingual Multimodal Pre-training

We use the self-attentive transformer architecture of BERT, and design two pre-training objectives with three types of data streams. Multitask training is employed into the pre-training stage to optimize all pre-training objectives simultaneously for better performance.

Data Stream

- **Multilingual Monomodal Stream** To apply multilingual pre-training, we use raw multilingual text as Multilingual Monomodal Stream.
- **Monolingual Multimodal Stream** To apply multimodal pre-training, we use raw image-text pair as Monolingual Multimodal Stream.
- **Multimodal Code-switched Stream** We generate Multimodal Code-switched Stream from Monolingual Multimodal Stream by code-switch.

Multilingual Training

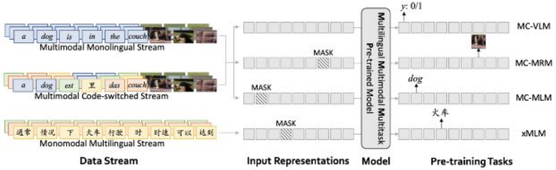
Multilingual Training aims to learn grammar or syntax from well formed multilingual sentences.

- **xMLM** This task performs masked language modeling based on the multilingual corpus.

Multimodal Code-switched Training

Multimodal Code-switched Training aims to learn different languages from the shared vision modal and the alignment between vision and non-English texts.

- **MC-VLM** This task aims to learn alignment between multilingual texts and images with mixed data stream.
- **MC-MRM** This task aims to learn vision representations with multilingual text as the context in mixed data stream.
- **MC-MLM** This task aims to learn the representation of different languages based on the shared vision modal.



Experiments

Model	Multi30K				MSCOCO		
	en	de	fr	cs	en	ja	zh
Monolingual supervised results							
EmbN [31]	72.0	60.3	54.8	46.3	76.8	73.2	73.5
PAR. EmbN [11]	69.0	62.6	60.6	54.1	78.3	76.0	74.8
S-LIWE [32]	76.3	72.1	63.4	59.4	80.9	73.6	70.0
MULE [15]	70.3	64.1	62.3	57.7	79.0	75.9	75.6
SMALR [1]	74.5	69.8	65.9	64.8	81.5	77.5	76.7
Monolingual results with multimodal pre-training							
Unicoder-VL (w/o fine-tune) [19]	72.0	-	-	-	63.7	-	-
Unicoder-VL (w/ fine-tune on en) [19]	**88.1**	-	-	-	**89.2**	-	-
Multilingual results with multimodal pre-training							
M^{3}P (w/o fine-tune)	57.9	36.8	27.1	20.4	63.1	33.3	32.3
M^{3}P (w/ fine-tune on en)	87.4	58.5	46.0	36.8	88.6	53.8	56.0
M^{3}P (w/ fine-tune on each)	87.4	82.1	67.3	65.0	88.6	80.1	75.8
M^{3}P (w/ fine-tune on all)	87.7	**82.7**	**73.9**	**72.2**	88.7	**87.9**	**86.2**

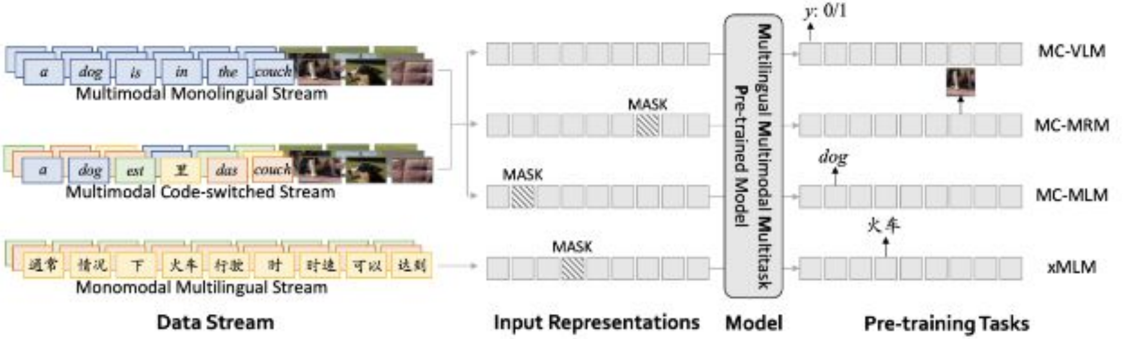
Our M^{3}P model obtains the state-of-the-art results in all non-English languages, which shows its exciting multilingual multimodal transfer capability.

Conclusion

We present M3P, the first known effort on combining multilingual pre-training and multimodal pre-training into a unified framework.

We proposed Multimodal Code-switched Training to further alleviate the issue of lacking enough labeled data for non-English multimodal tasks and avoid the tendency to model the relationship between vision and English text.

- **MC-VLM** This task aims to learn alignment between multilingual texts and images with mixed data stream.
- **MC-MRM** This task aims to learn vision representations with multilingual text as the context in mixed data stream.
- **MC-MLM** This task aims to learn the representation of different languages based on the shared vision modal.



Inputs
- Multimodal monolingual
- Multimodal code-switched
- Monomodal multilingual

Pretasks
- VLM: matching language-image
- MRM: MLM vision representation version
- MLM: masked language modeling

MINDs Lab

Multimodal Contrastive Training for Visual Representation Learning

Xin Yuan, University of Chicago
Zhe Lin, Adobe Research
Jason Kuen, Adobe Research
Jianming Zhang, Adobe Research
Yilin Wang, Adobe Research
Michael Maire, University of Chicago
Ajinkya Kale, Adobe
Baldo Faieta, Adobe

Overview

Goal:

- Improve the quality of pre-trained visual representations
- Learn more generic visual features for various tasks

Approach:

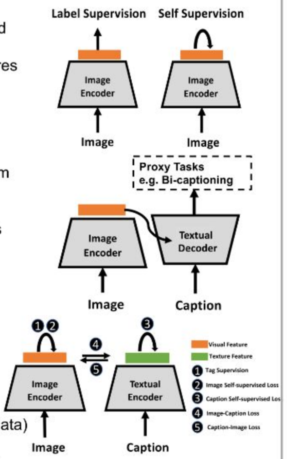
- Exploit intrinsic data properties within each modality
- Extract semantic information from cross-modal correlation
- Combine intra- and inter-modal similarity preservation objectives

Consequences:

- Unifies multi-modal training in a flexible framework
- Visual representations can be transferred and achieve excellent performance

Experimental Results:

- ResNet50 on ImageNet:
  - Pre-train on COCO (10x less data)
  - 55.3% Top-1 Acc
- Generalize across various tasks
- Effective on large-scale Stock images dataset



Training with Multi-modal Contrastive Learning

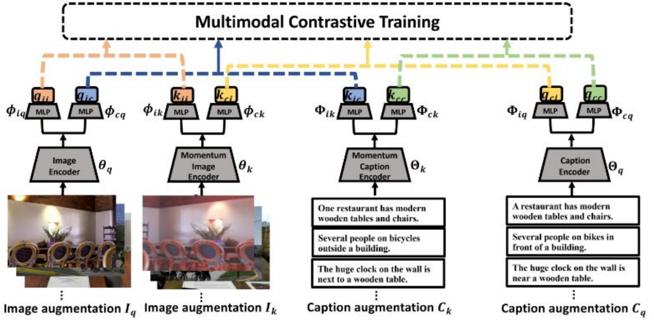
Our framework is composed of two contrastive training schemes:

Intra-modal (orange and green paths)

- Train encoders for each individual modality in a self-supervised manner
- Additional textual encoder captures semantics from augmented sentence
- Involve the tag information to improve the visual representations

Inter-modal (yellow and blue paths)

- Embed the visual and textual features into common space
- Visual-semantic contrastive loss to force similar samples to be closer



Contrastive Objectives

Visual Contrastive Learning

- Image encoder $f_{iq}(\cdot;\theta,\phi_{iq})$,
- Momentum encoder $f_{ik}(\cdot;\theta,\phi_{ik})$
- ***Query*** and ***key*** features embedding: $q_{ii}^{j} = f_{iq}(I_j^{+};\theta_q,\phi_{iq});\ k_{ii}^{j} = f_{ik}(I_j^{*};\theta_k,\phi_{ik})$
- $J_{ii} = -log\frac{\exp(q_{ii}\cdot k_{ii}^{+}/\tau)}{\sum_{j=0}^{K}\exp(q_{ii}\cdot k_{ii}^{j}/\tau)}$
- More closely semantic-aligned visual features with tag supervision
- $P = \{k_{ii}^{p} | \forall p: t_p \cdot t_j > \varepsilon\}$
- $J_{tag} = -\frac{1}{|P|}\sum_{p\in P} log\frac{\exp(q_{ii}\cdot k_{ii}^{p}/\tau)}{\sum_{j=0}^{K}\exp(q_{ii}\cdot k_{ii}^{j}/\tau)}$

Textual Contrastive Learning

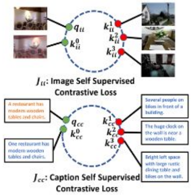
- $q_{cc}^{j} = f_{cq}(c_j^{+};\Theta_q,\Phi_{cq}); k_{cc}^{j} = f_{ck}(c_j^{*};\Theta_k,\Phi_{ck})$
- $J_{ii} = -log\frac{\exp(q_{ii}\cdot k_{ii}^{+}/\tau)}{\sum_{j=0}^{K}\exp(q_{ii}\cdot k_{ii}^{j}/\tau)}$

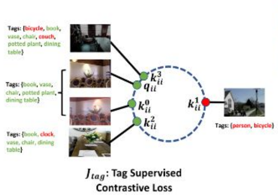
Image-to-Caption Contrastive Learning

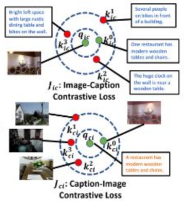
- $q_{ic}^{j} = f_{iq}(I_j^{+};\theta_q,\phi_{cq}),\ k_{ic}^{j} = f_{ck}(c_j^{*};\Phi_k,\Phi_{ik})$
- $J_{ic} = \sum_{j=1}^{K}[\alpha - q_{ic}\cdot k_{ic}^{+} + q_{ic}\cdot k_{ic}^{j}]_{+}$

Caption-to-Image Contrastive Learning

- $q_{ci}^{j} = f_{cq}(c_j^{+};\Theta_q,\Phi_{iq}),\ k_{ci}^{j} = f_{ik}(I_j^{*};\theta_k,\phi_{ck})$
- $J_{ci} = \sum_{j=1}^{K}[\alpha - q_{ci}\cdot k_{ci}^{+} + q_{ci}\cdot k_{ci}^{j}]_{+}$







Experimental Results

Linear Cls. on ImageNet

- Outperforms VirTex and ICMLM by ***2.1% and 3.0%***
- Further improve by ***0.4%*** by leveraging tags
- Better performance than Sup. Pre-training on IN-100

Model	Pretrain Dataset	Supervision	Top-1 (%)
IN-Sup	IN-1K	Label	76.1
IN-Sup	IN-100	Label	53.3
MoCo-v2[1]	COCO	NA	49.3
VirTex[2]	COCO	Caption	52.8
$ICMLM_{tfm}$[3]	COCO	Caption	51.9
Ours	**COCO**	**Caption**	**54.9**
Ours(with tag)	**COCO**	**Caption+Tag**	**55.3**

Object Detection on VOC

- Fine-tune ResNet-50-C4 backbones on VOC trainval 07+12 split
- Significantly outperforms self-supervised method which uses COCO

Model	Pretrain Dataset	AP_{50}	AP	AP_{75}
IN-Sup	IN-1K	81.6	54.3	59.7
MoCo-v2[1]	IN-1K	82.4	57.0	63.6
MoCo-v2[1]	COCO	75.4	48.4	52.1
VirTex[2]	COCO	81.4	55.6	61.5
Ours	**COCO**	**82.1**	**56.1**	**62.4**

Cross-modal Search on COCO 1K test-set

- Consistently performs better than all competing methods
- … generate 2048-dglobal pooled features, …mapped to 1024-dby fully connected layers

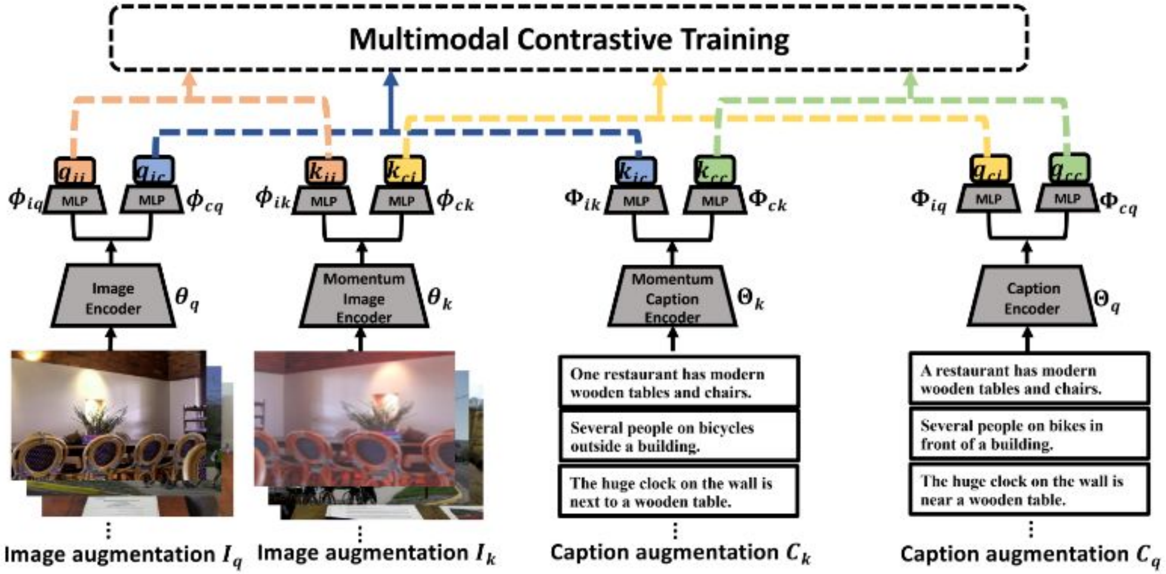
Method	Image-to-Text			Text-to-Image		
	R@1	R@10	Med r	R@1	R@10	Med r
IN-Sup	57.9	92.7	1.0	42.8	87.0	2.0
MoCo-v2[1]	51.6	90.0	1.0	39.0	84.8	2.0
VirTex[2]	58.1	93.0	1.0	44.0	88.5	2.0
Ours	**58.4**	**93.4**	**1.0**	**45.1**	**90.0**	**2.0**

Ablation Study on Separate MLP Design

- Separate design consistently yields better visual features
- Final design (128-d for intra-modal;1024-d for inter-modal) performs best (***54.9%***)

Reference

[1] Xinlei Chen, Haoqi Fan, Ross B. Girshick, Kaiming He. Improved baselines with momentum contrastive learning, arxiv 2020.
[2] Karan Desai, Justin Johnson. Virtex: Learning visual representations from textual annotations, CVPR 2021.
[3] Mert Bulent Sariyildiz, Julien Perez, Diane Larlus. Learning visual representations with caption annotations, ECCV 2020.



- One-tool:
  the same meaning,
  the same representation.

Conclusion

- We explored multi-modal joint representation learning.
- The simple technique has potentials for various applications.