Friedrich-Alexander-Universität
Technische Fakultät

FAU

Advanced Deep Learning

Multimodal Learning

K. Breininger, V. Christlein

Artificial Intelligence in Medical Imaging + Pattern Recognition Lab,

Friedrich-Alexander-Universität Erlangen-Nürnberg SoSe 2023

Slides in part courtesy of Jindong Li (FAU)

1. Introduction to Multimodal Learning
2. Multimodal Learning Architectures
3. Improving Multimodal Learning
4. Outlook

Goals for Today

You will be able to ...

- explain the term “multimodal” in the context of deep learning.
- discuss applications of multimodal learning.
- explain the main differences between architectures for multimodal learning and argue benefits and disadvantages.
- discuss the connection between multimodal learning and self-supervised learning.

What we will not talk about (in detail):

- **Multi-task** learning
- Multimodal learning in the context of medical data (CT, MRI, ...)

1. Introduction to Multimodal Learning

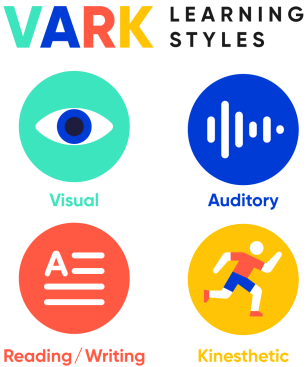
2. Multimodal Learning Architectures

3. Improving Multimodal Learning

4. Outlook

What is Multimodal Learning?

- **Modality**: Modes of digital data
- **Multimodal learning**: model combination of different modalities



Source: https://www.learnupon.com/blog/multimodal-learning/

Unimodal vs. Multimodal Learning

Unimodal Learning

- Unimodal data
- (Often) less complex architecture
- High performance possible with large datasets

Multimodal Learning

- Multimodal data
- (Often) more complex architecture
- Potentially higher performance through different data sources

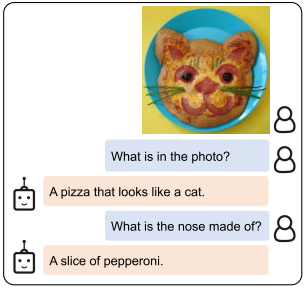
Multimodal Learning Tasks

Image Text Retrieval

- Retrieve samples in one modality relevant to samples in other modality
- Subtasks
  - image-to-text retrieval
  - text-to-image retrieval

Visual Question Answering

- Predict answer given question about an image

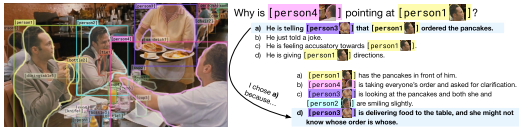


Source: Li, Li, Savarese, et al. 2023 [1]

Multimodal Learning Tasks (cont.)

Visual Reasoning

- Understand relation and interaction between objects



Source: Zellers, Bisk, Farhadi, et al. 2019 [2]

Visual Entailment

- Determine if a hypothesis can be concluded from an image

- *Two woman are holding packages.*
- *The sisters are hugging goodbye while holding to go packages after just eating lunch.*
- *The men are fighting outside a deli.*

+

- *Entailment*
- *Neutral*
- *Contradiction*

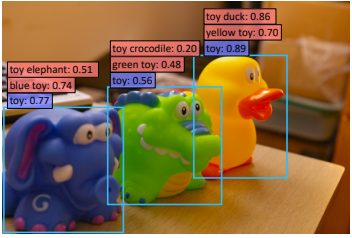
=

Premise *Hypothesis* *Answer*

Source: Xie, Lai, Doran, et al. 2019 [3]

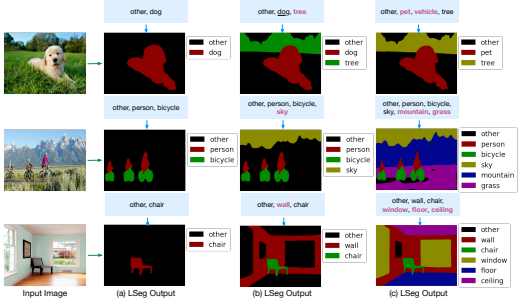
Multimodal Learning Tasks (cont.)

Language-guided Detection



Source: Gu, Lin, Kuo, et al. 2022 [4]

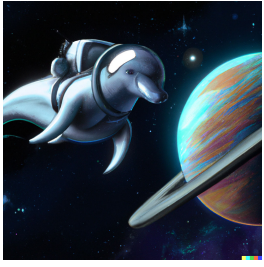
Language-guided Segmentation



Source: Li, Weinberger, Belongie, et al. 2022 [5]

Language-guided Image Generation

e. g., Dall·e 2[1]



A dolphin in an astronaut suit on Saturn

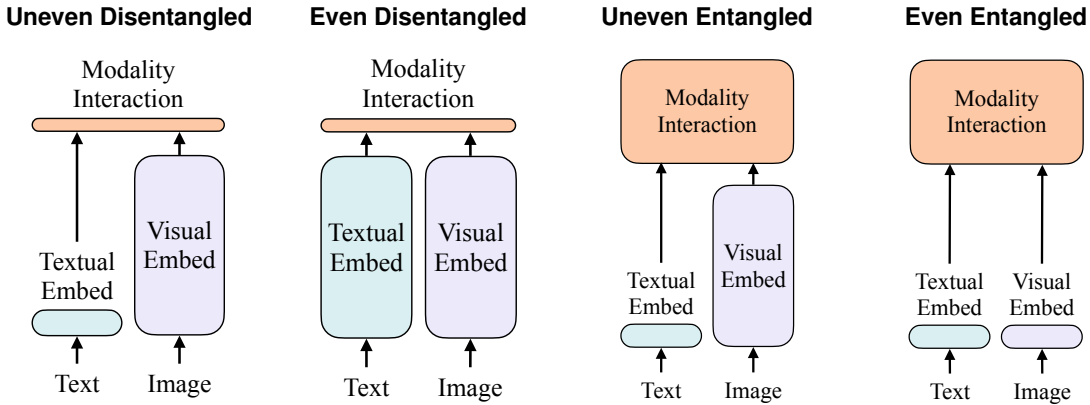
A teddy bear on a skateboard on Times Square

Source: Ramesh, Dhariwal, Nichol, et al. 2022 [6]

[1] Ramesh, Dhariwal, Nichol, et al.: Hierarchical Text-Conditional Image Generation with CLIP Latents (2022) [6]

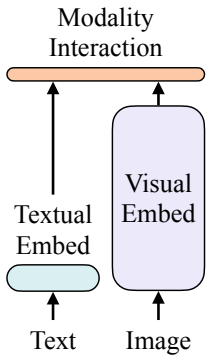
1. Introduction to Multimodal Learning
2. Multimodal Learning Architectures
3. Improving Multimodal Learning
4. Outlook

Concepts



Source: Kim, Son, and Kim 2021 [7]

Uneven Disentangled



- **Heavy** visual embedder
- **Light** textual embedder
- **Weak** modality interaction

VSE++[2]

- Visual Embedder: VGG-19 or ResNet-152
- Textual Embedder: RNN-based language model
- Modality interaction: inner product

Source: Kim, Son, and Kim 2021 [7]

[2] Faghri, Fleet, Kiros, et al.: "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives" (2018) [8]

Uneven Disentangled

VSE++

Caption Retrieval

Model	R@1	R@5	R@10
sm-LSTM [9]	53.2	83.1	91.5
VSE++	**64.6**	**90.0**	**95.7**

Image Retrieval

Model	R@1	R@5	R@10
sm-LSTM	40.7	75.8	87.4
VSE++	**52.0**	**84.3**	**92.0**

Summary

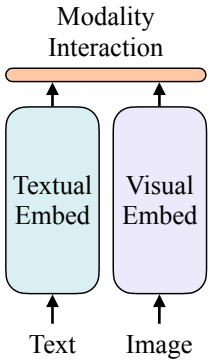
- Performs better than other methods
- Can't solve difficult tasks (VQA, VR, ...)

Even Disentangled

- **Heavy** visual embedder
- **Heavy** textual embedder
- **Weak** modality interaction

CLIP (Contrastive Language-Image Pretraining)[3]

- Visual Embedder: Vision Transformer
- Textual Embedder: Transformer
- Modality Interaction: Scaled pairwise cosine similarities

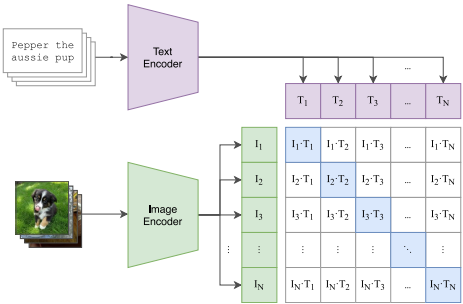


Source: Kim, Son, and Kim 2021 [7]

[3] Radford, Kim, Hallacy, et al.: Learning Transferable Visual Models From Natural Language Supervision (2021) [10]

Even Disentangled

CLIP



Training

1. Get sequences from both encoders
2. Apply symmetric contrastive loss with $\langle \mathbf{v}, \mathbf{u} \rangle = \frac{\mathbf{v}^T \mathbf{u}}{\|\mathbf{v}\| \|\mathbf{u}\|}$:

$$\ell_i^{(T \to I)} = -\log \frac{\exp\left(\langle \boldsymbol{T}_i, \boldsymbol{I}_i \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle \boldsymbol{T}_i, \boldsymbol{I}_k \rangle / \tau\right)}$$

$$\ell_i^{(I \to T)} = -\log \frac{\exp\left(\langle \boldsymbol{I}_i, \boldsymbol{T}_i \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle \boldsymbol{I}_i, \boldsymbol{T}_k \rangle / \tau\right)}$$

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \ell_i^{(T \to I)} + \ell_i^{(I \to T)}$$

Source: Radford, Kim, Hallacy, et al. 2021 [10]

Even Disentangled

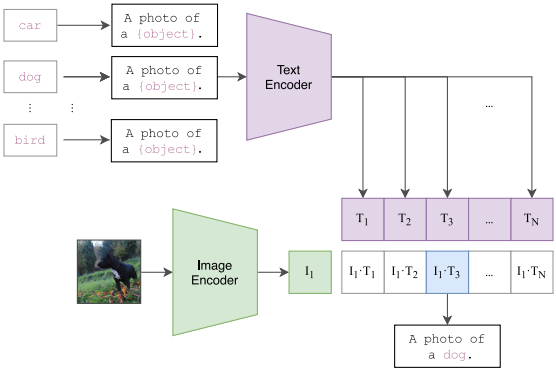
CLIP

Dataset

- 400 mio (image,text) pairs from the Internet
- 500 k queries (words occ. at least 100 times in English Wikipedia)
- About 20 k (image,text) pairs per query

Even Disentangled

CLIP



Zero-shot

- Encode class names as potential text pairing
- Use for image-based zero-shot prediction

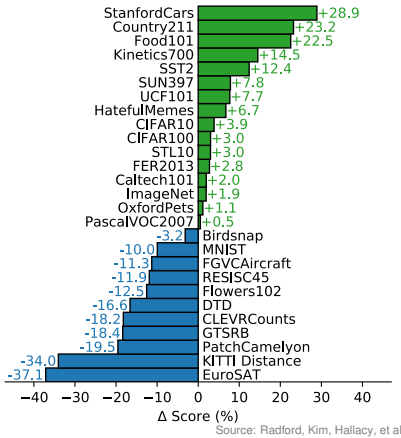
Source: Radford, Kim, Hallacy, et al. 2021 [10] (adapted)

Even Disentangled

CLIP

Zero-shot Performance

- 27 datasets
- Prompt engineering (e. g., add “a type of food” for Food101) pet.
- Ensembling of multiple text embeddings
- Zero-shot CLIP vs. linear probe ResNet-50 ImageNet features
- Zero-shot CLIP better at **16** datasets
- Bad at highly specialized and abstract tasks



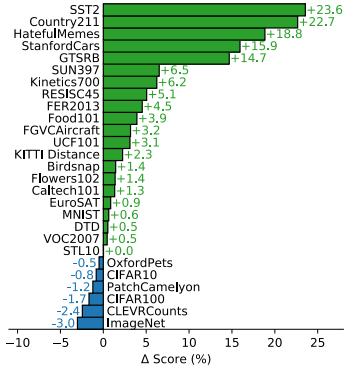
Source: Radford, Kim, Hallacy, et al. 2021 [10]

Even Disentangled

CLIP

Representation Learning Performance

- Comparison between linear probes of CLIP and EfficientNet [12]
- Linear probe of CLIP better on 21 datasets
- Overly narrow supervision on ImageNet



Source: Radford, Kim, Hallacy, et al. 2021 [10]

Even Disentangled

CLIP

Summary

- Excellent general performance
- Especially for zero-shot scenarios
- Unable to solve difficult tasks

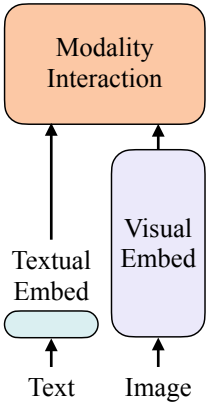
→ The standard module to combine image + text modalities for many generative models (Dall·E, StableDiffusion, etc.)

Uneven Entangled

- **Heavy** visual embedder
- **Light** textual embedder
- **Strong** modality interaction

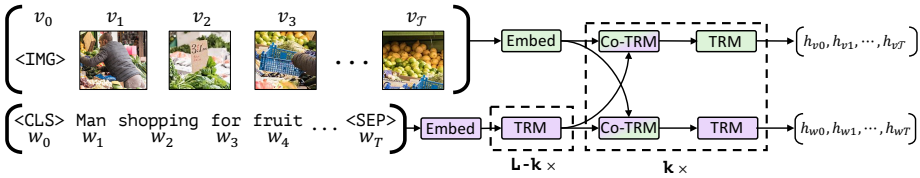
ViLBERT[4]

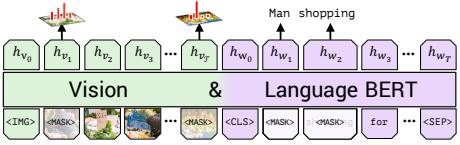
- Visual Embedder: Pre-trained Faster R-CNN (ResNet-101 backbone)
- Textual Embedder: Transformer
- Modality Interaction: Transformer

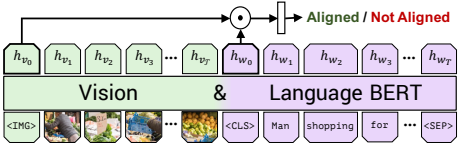


Uneven Entangled

VilBERT







- Mask text: as in BERT
- Image mask: KL Div. between class distribution of patch and of pretrained detection model

Source: Kim, Son, and Kim 2021 [7]

Uneven Entangled

VilBERT

Method	VQA	VR	Image Retrieval			Zero-Shot Image Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
SOTAs	70.22	43.1	48.60	77.70	85.20	-	-	-
ViLBERT	**70.55**	**54.04**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

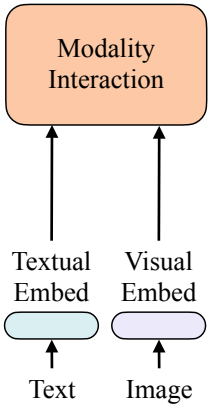
+ Good performance
– Object detection overhead

Even Entangled

- **Light** visual embedder
- **Light** textual embedder
- **Strong** modality interaction

ViLT[5]

- Visual embedder: Linear projection of flattened patches
- Textual embedder: Tokenization
- Modality interaction: Transformer



Source: Kim, Son, and Kim 2021 [7]

[5] Kim, Son, and Kim: "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision" (2021) [7]

Even Entangled

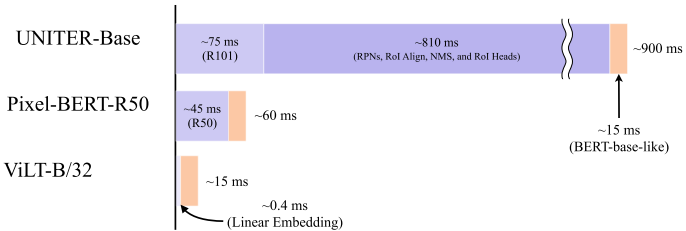
ViLT

Why is the region feature popular?

- Discrete, semantic
- Down-stream tasks are related to objects
- Cached as sequences in advance

What do we get?

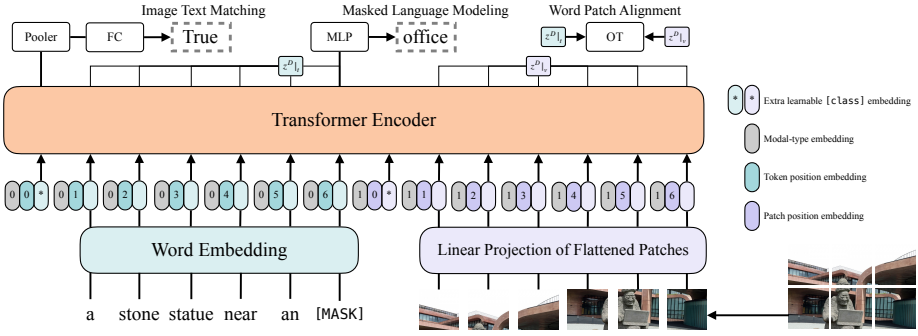
- Limited categories of objects
- Slow on inference
- Sub-optimal (fixed pre-trained model)



Source: Kim, Son, and Kim 2021 [7]

Even Entangled

ViLT



- Image Text Matching
  - Randomly replace input image w. another
  - Add FC to classify pairs
- Masked Language Modeling w. whole word masking
- Word Patch Alignment

Source: Kim, Son, and Kim 2021 [7]

Performance

Model	Time	Visual Question Answering	Visual Reasoning
VinVL-Base [15]	650ms	**75.95**	**83.08**
ViLT-B/32	**15ms**	71.26	76.13

Model	Time	Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER-Base [16]	900ms	**85.9**	**97.1**	**98.8**	**72.5**	**92.4**	**96.1**
ViLT-B/32	**15ms**	83.5	96.7	98.6	64.4	88.7	93.8

Even Entangled

ViLT

Summary

+ Extremely high inference speed
– Sub-optimal performance
- Training computational demanding
  - 64 * NVIDIA V100 GPUs
  - Batch size: 4096
  - 200K steps
- Word patch alignment is slow

1. Introduction to Multimodal Learning
2. Multimodal Learning Architectures
3. Improving Multimodal Learning
4. Outlook

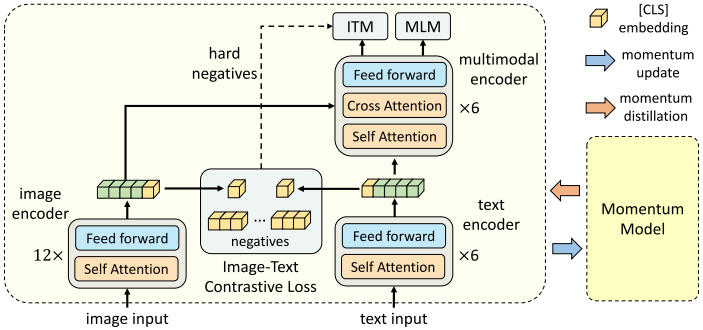
What did we learn?

The things we should use

- Vision Transformer
- Strong modality interaction
- Image augmentation
- Image text contrastive loss
- Masked language modeling loss
- Image text matching loss

Align before Fuse

Vision and Language Representation Learning with Momentum Distillation [17]



Source: Li, Selvaraju, Gotmare, et al. 2021 [17]

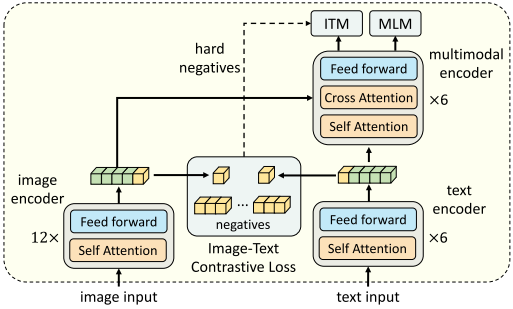
Align before Fuse

Architecture

- Visual encoder: 12-layer vision transformer ViT-B/16
- Textual encoder: 6 layers of transformer
- Modality interaction: 6 layers of transformer
- Modality interaction through **cross attention** layers

Three core objectives:

- Image-text contrastive loss (ITC)
- Masked language modeling (MLM)
- Image-text matching (ITM)

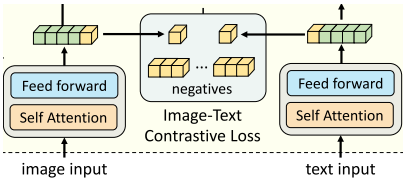


Align before Fuse

Objectives I

Image-text contrastive (ITC) loss

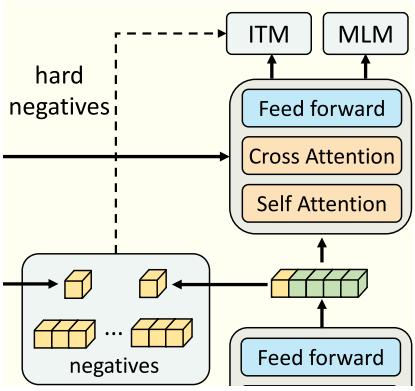
- Core idea:
  Align unimodal embeddings before fusing them
- Take **class tokens** from outputs of both encoders
- **Project** and **normalize** class tokens
- Queue of recent inputs (cf. MoCo [23]) for negative samples
- Cosine similarity between positive and negative pairs
- Calculate ITC loss (cross entropy)



Objectives II

Masked Language Modeling

- Core idea: **Self-supervised pretraining**
- Input tokens replaced by `[MASK]` with 15% probability
- Joint image + text representation used for BERT-like masked prediction

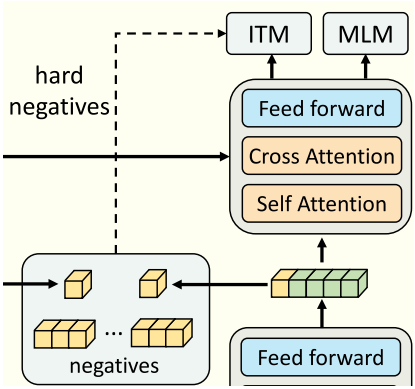


Align before Fuse

Objectives III

Image-text matching (ITM)

- Core idea:
  Enforce alignment throughout encoder
- Classification of multimodal embedding as matched vs. non-matched
- **Hard negative** sampling:
  Sample the **negative pairs** according to **similarity** from all **ITC negative samples**



Source: Li, Selvaraju, Gotmare, et al. 2021 [17]

Align before Fuse

Momentum Distillation

Web data is not perfect

- Text contains unrelated words
- Images contain undescribed entities
- ITC / MLM penalize regardless of correctness

"polar bear in the [MASK]"

GT: wild
Top-5 pseudo-targets:
1. zoo
2. pool
3. water
4. pond
5. wild

Momentum Distillation for ITC and MLM

- **Exponential-moving-average** versions (MoD) of base model
- MoD model generates **pseudo targets**
- Predictions matched against **GT** (1-α) and **momentum model prediction** (α)

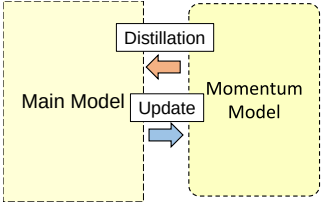


Image-Text Retrieval

Model	# pre-train images	Text Retrieval			Zero-Shot Text Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
ALIGN [18]	1.2 B	95.3	99.8	100.0	88.6	98.7	99.7
ALBEF	14 M	**95.9**	**99.8**	**100.0**	**94.1**	**99.5**	**99.7**

Model	Image Retrieval			Zero-Shot Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
ALIGN	84.9	97.4	98.6	75.7	93.8	96.8
ALBEF	**85.6**	**97.5**	**98.9**	**82.8**	**96.3**	**98.1**

Visual Question Answering, Visual Reasoning, and Visual Entailment

Model	Visual Question Answering	Visual Reasoning	Visual Entailment
VILLA [19]	73.67	79.30	79.03
ALBEF	**76.04**	**83.14**	**80.91**

- Inference speed **more than 10** $\times$ **faster** than VILLA

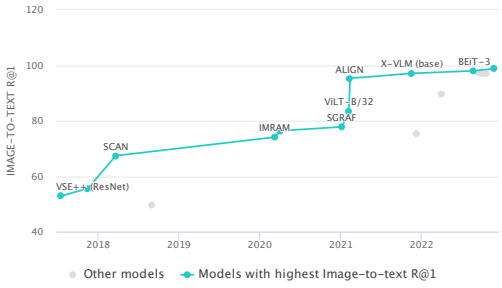
Align before Fuse

Summary

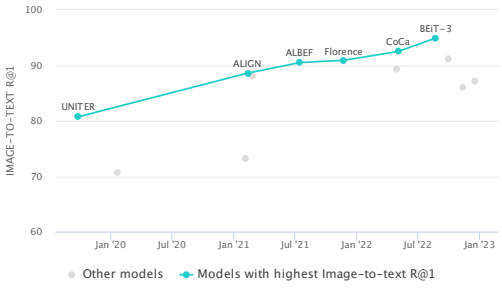
- Achieve state-of-the-art performance
- Fast inference speed
- Low computational demand for training
  - 8 * NVIDIA A100 GPUs
  - Batch size: 512
  - 30 epochs

Outlook

- Significant improvement in recent years



Flickr30K Image Retrieval



Flickr30K Zero-Shot Image Retrieval

Source: https://paperswithcode.com/sota/cross-modal-retrieval-on-flickr30k
Source: https://paperswithcode.com/sota/zero-shot-cross-modal-retrieval-on-flickr30k

Outlook

How Multimodal Learning benefits from Vision Transformer?

- Scalability
- Attention mechanism
- Global receptive field
- Improved transfer learning / pretrained models

Outlook

There is more in Multimodal Learning

- Audio-visual multimodal learning
  - VoViT[6]
- Audio-language multimodal learning
  - Audio-to-language: Whisper[7]
  - Language-to-audio: MelNet[8]
- ...

[6] Montesinos, Kadandale, and Haro: "VoViT: Low Latency Graph-Based Audio-Visual Voice Separation Transformer" (2022) [20]

[7] Radford, Kim, Xu, et al.: Robust Speech Recognition via Large-Scale Weak Supervision (2022) [21]

[8] Vasquez and Lewis: "MelNet: A Generative Model for Audio in the Frequency Domain" (2019) [22]

NEXT TIME

ADVANCED

ON DEEP LEARNING

Denoising Diffusion Models

Text prompt

medium-full off-center shot,
35 mm Kodachrome film still,
capturing a Japanese woman peacing out and waving down a taxi,
wearing a gingham print dress made of silk,
blue/white palette,
accessorized by sleek pearl earrings,
another moody late-night in Tokyo
–ar 1:1

Generate

Source: https://twitter.com/nickfloats/status/1645522764084772875

Comprehensive Questions

- What are the core differences between multimodal models?
- How does zero-shot learning/transfer learning work in the context of multimodal models like CLIP?
- What are the main evaluation metrics and how are they computed?
- What are dataset requirements for multi-modal models?
- What is the role of the momentum distillation in ALBEF?
- Why is ALBEF more efficient compared to other techniques?

Further Reading

- Overview “book” from a seminar course at LMU (Munich):
  Cem Akkus, Luyang Chu, Vladana Djakovic, et al. Multimodal Deep Learning. 2023. arXiv: 2301.04856 [cs.CL]
- Course on Multimodal Machine Learning at Carnegie Mellon University (with videos):
  https://cmu-multicomp-lab.github.io/mmml-course/fall2022/
- Corresponding review article:
  Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency.
  Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions
  2023. arXiv: 2209.03430 [cs.LG]
- Hands on tutorial on multimodal learning from KDD ’22:
  https://github.com/dsaidgovsg/multimodal-learning-hands-on-tutorial

References

References I

[1] Junnan Li, Dongxu Li, Silvio Savarese, et al.
BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Languag
arXiv:2301.12597 [cs]. Jan. 2023.

[2] Rowan Zellers, Yonatan Bisk, Ali Farhadi, et al. “From Recognition to Cognition: Visual Commonsense Reasoning”. en. In:
2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, June 2019, pp. 6713–6724.

[3] Ning Xie, Farley Lai, Derek Doran, et al.
Visual Entailment: A Novel Task for Fine-Grained Image Understanding. arXiv:1901.06706 [cs]. Jan. 2019.

[4] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, et al. “Open-vocabulary Object Detection via Vision and Language Knowledge Distillation”. In: International Conference on Learning Representations. 2022.

References II

[5] Boyi Li, Kilian Q Weinberger, Serge Belongie, et al. “Language-driven Semantic Segmentation”. In: International Conference on Learning Representations. 2022.

[6] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs]. Apr. 2022.

[7] Wonjae Kim, Bokyung Son, and Ildoo Kim. “ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision”. en. In: Proceedings of the 38th International Conference on Machine Learning. ISSN: 2640-3498. PMLR, July 2021, pp. 5583–5594.

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, et al. “VSE++: Improving Visual-Semantic Embeddings with Hard Negatives”. In: British Machine Vision Conference (BMVC). July 2018.

References III

[9] Yan Huang, Wei Wang, and Liang Wang. “Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM”. en. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, July 2017, pp. 7254–7262.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning Transferable Visual Models From Natural Language Supervision. 18–24 Jul 2021.

[11] Yonglong Tian, Dilip Krishnan, and Phillip Isola. “Contrastive Multiview Coding”. en. In: Computer Vision – ECCV 2020. Vol. 12356. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 776–794.

[12] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, et al. “Self-Training With Noisy Student Improves ImageNet Classification”. en. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, June 2020, pp. 10684–10695.

References IV

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: Advances in Neural Information Processing Systems. Vol. 32. Curran Associates, Inc., 2019.

[14] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, et al. “RandAugment: Practical Automated Data Augmentation with a Reduced Search Space”. In: Advances in Neural Information Processing Systems. Vol. 33. Curran Associates, Inc., 2020, pp. 18613–18624.

[15] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, et al. “VinVL: Revisiting Visual Representations in Vision-Language Models”. en. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, June 2021, pp. 5575–5584.

References V

[16] Yen-Chun Chen, Linjie Li, Licheng Yu, et al. “UNITER: UNiversal Image-TExt Representation Learning”. en. In: Computer Vision – ECCV 2020. Vol. 12375. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 104–120.

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, et al. “Align before Fuse: Vision and Language Representation Learning with Momentum Distillation”. In: Advances in Neural Information Processing Systems. Vol. 34. Curran Associates, Inc., 2021, pp. 9694–9705.

[18] Chao Jia, Yinfei Yang, Ye Xia, et al. “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision”. en. In: Proceedings of the 38th International Conference on Machine Learning. ISSN: 2640-3498. PMLR, July 2021, pp. 4904–4916.

References VI

[19] Zhe Gan, Yen-Chun Chen, Linjie Li, et al. “Large-Scale Adversarial Training for Vision-and-Language Representation Learning”. In: Advances in Neural Information Processing Systems. Vol. 33. Curran Associates, Inc., 2020, pp. 6616–6628.

[20] Juan F. Montesinos, Venkatesh S. Kadandale, and Gloria Haro. “VoViT: Low Latency Graph-Based Audio-Visual Voice Separation Transformer”. en. In: Computer Vision – ECCV 2022. Vol. 13697. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 310–326.

[21] Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [cs, eess]. Dec. 2022.

[22] Sean Vasquez and Mike Lewis. “MelNet: A Generative Model for Audio in the Frequency Domain”. en. In: (Dec. 2019).

References VII

[23] Kaiming He, Haoqi Fan, Yuxin Wu, et al. “Momentum Contrast for Unsupervised Visual Representation Learning”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2020.

[24] Cem Akkus, Luyang Chu, Vladana Djakovic, et al. Multimodal Deep Learning. 2023. arXiv: 2301.04856 [cs.CL].

[25] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Question 2023. arXiv: 2209.03430 [cs.LG].