



AAAI-23

BridgeTower: Building Bridges Between Encoders in Vision-Language Representation Learning

Xiao Xu^{1,2}, Chenfei Wu², Shachar Rosenman³, Vasudev Lal³, Wanxiang Che¹, Nan Duan²

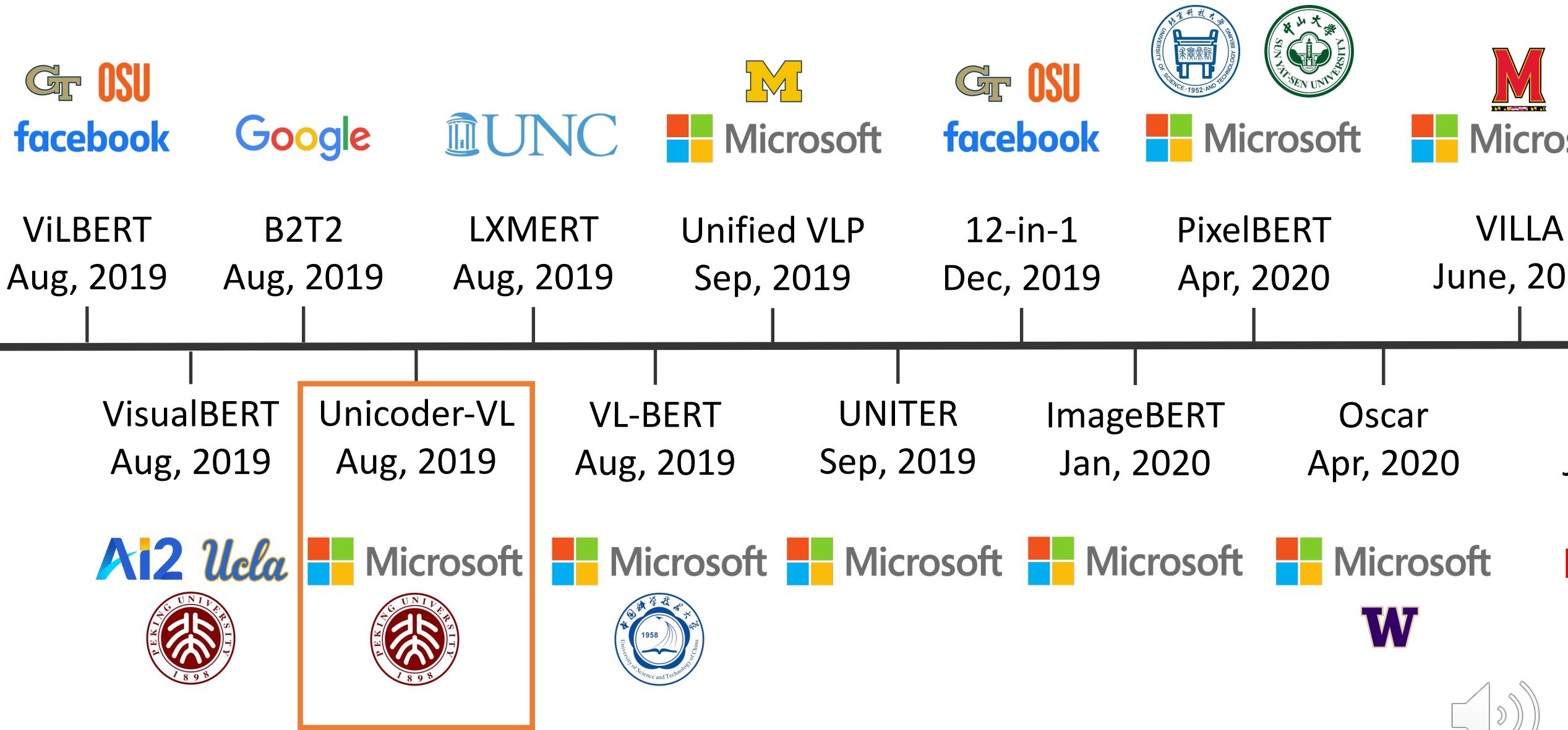
¹Harbin Institute of Technology, ²Microsoft Research Asia, ³Intel Labs

Presenter: Xiao Xu

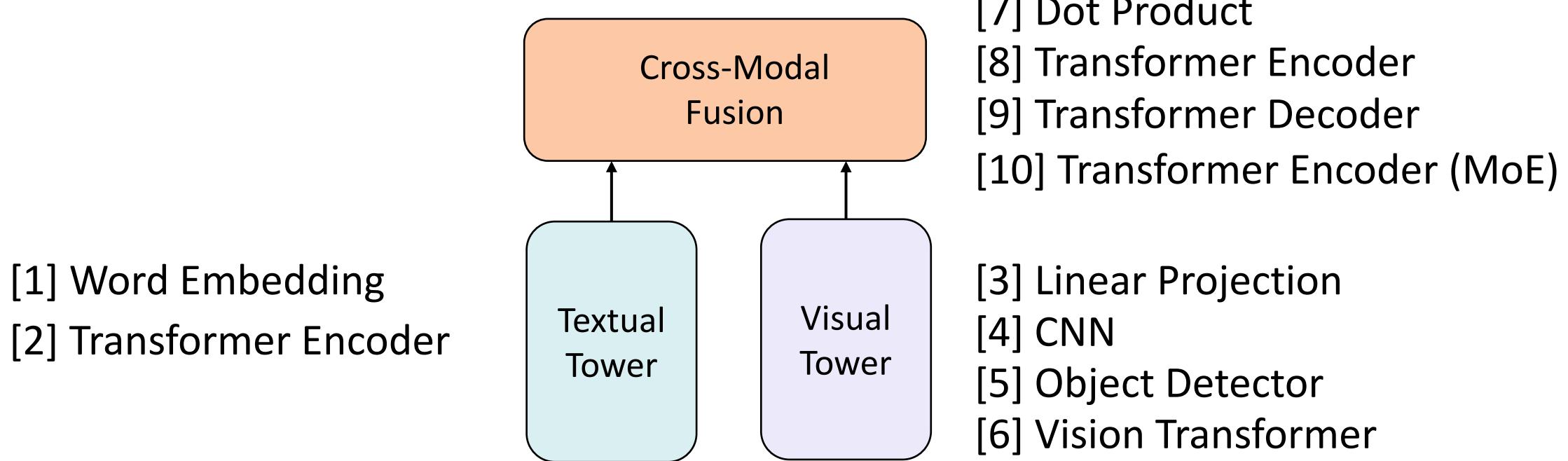


1 Work done during the internship of Microsoft Research Asia.

Vision-Language Pre-training Background

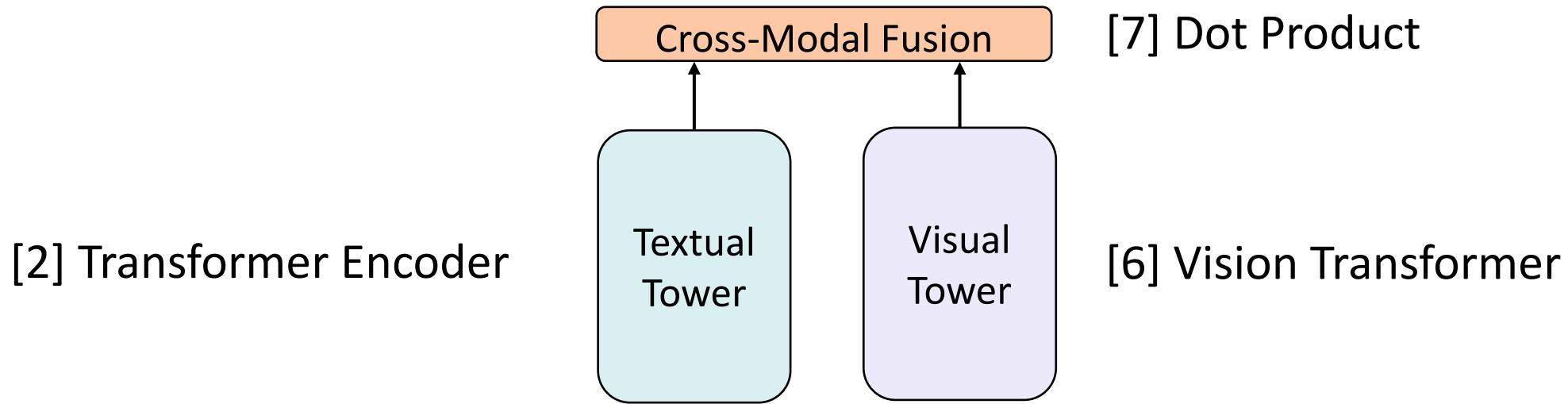


Traditional Two-Tower Architecture

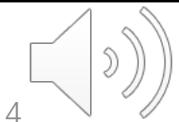


	CLIP	SOHO	Unicoder-VL	ViLT	VL-T5	METER	UniT	OFA	VLMo	BLIP
Textual-Tower	[2]	[1]	[1]	[1]	[1]	[2]	[2]	[1]	[1]	[2]
Visual-Tower	[6]	[4]	[5]	[3]	[5]	[6]	[6]	[4]	[3]	[6]
Cross-Fusion	[7]	[8]	[8]	[8]	[8] + [9]	[8]	[9]	[8] + [9]	[10]	[8] + [9]

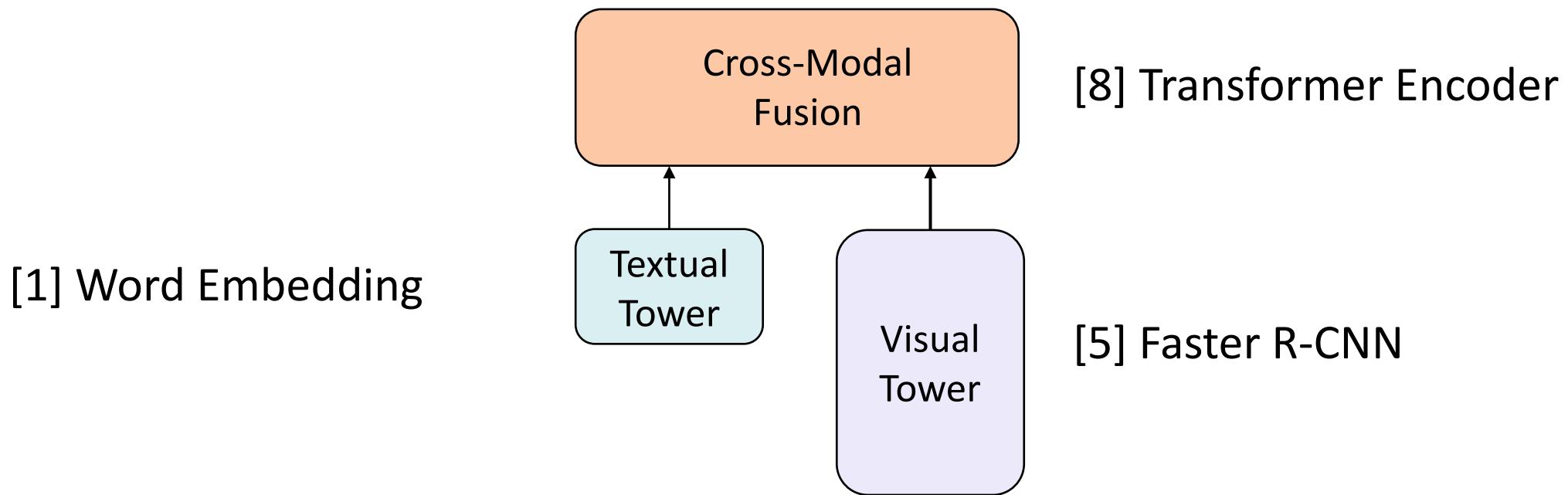
Traditional Two-Tower Architecture



	CLIP	SOHO	Unicoder-VL	ViLT	VL-T5	METER	UniT	OFA	VLMo	BLIP
Textual-Tower	[2]	[1]	[1]	[1]	[1]	[2]	[2]	[1]	[1]	[2]
Visual-Tower	[6]	[4]	[5]	[3]	[5]	[6]	[6]	[4]	[3]	[6]
Cross-Fusion	[7]	[8]	[8]	[8]	[8] + [9]	[8]	[9]	[8] + [9]	[10]	[8] + [9]



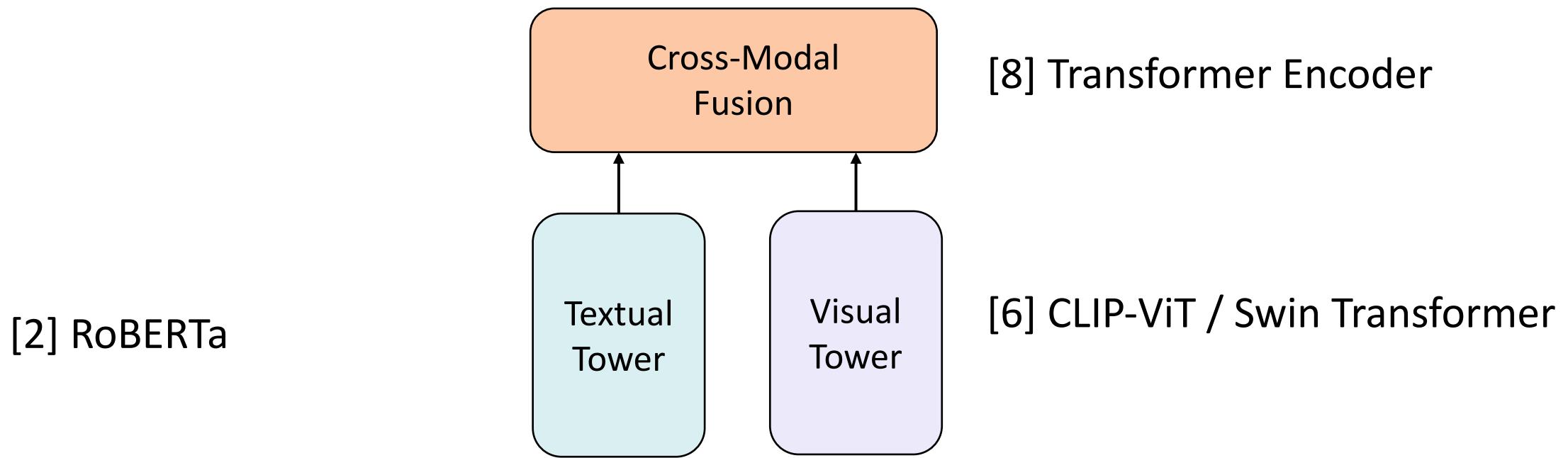
Traditional Two-Tower Architecture



	CLIP	SOHO	Unicoder-VL	ViLT	VL-T5	METER	UniT	OFA	VLMo	BLIP
Textual-Tower	[2]	[1]	[1]	[1]	[1]	[2]	[2]	[1]	[1]	[2]
Visual-Tower	[6]	[4]	[5]	[3]	[5]	[6]	[6]	[4]	[3]	[6]
Cross-Fusion	[7]	[8]	[8]	[8]	[8] + [9]	[8]	[9]	[8] + [9]	[10]	[8] + [9]



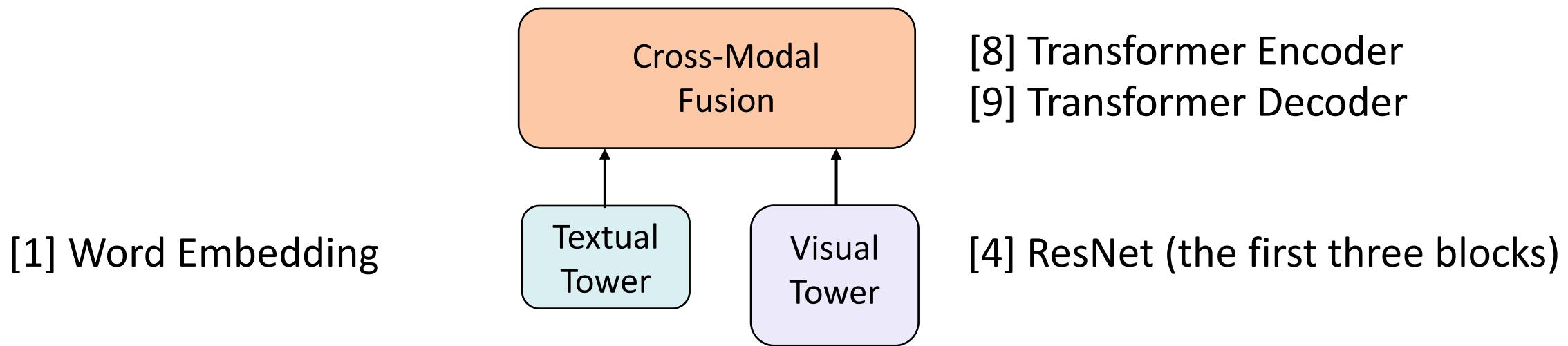
Traditional Two-Tower Architecture



	CLIP	SOHO	Unicoder-VL	ViLT	VL-T5	METER	UniT	OFA	VLMo	BLIP
Textual-Tower	[2]	[1]	[1]	[1]	[1]	[2]	[2]	[1]	[1]	[2]
Visual-Tower	[6]	[4]	[5]	[3]	[5]	[6]	[6]	[4]	[3]	[6]
Cross-Fusion	[7]	[8]	[8]	[8]	[8] + [9]	[8]	[9]	[8] + [9]	[10]	[8] + [9]

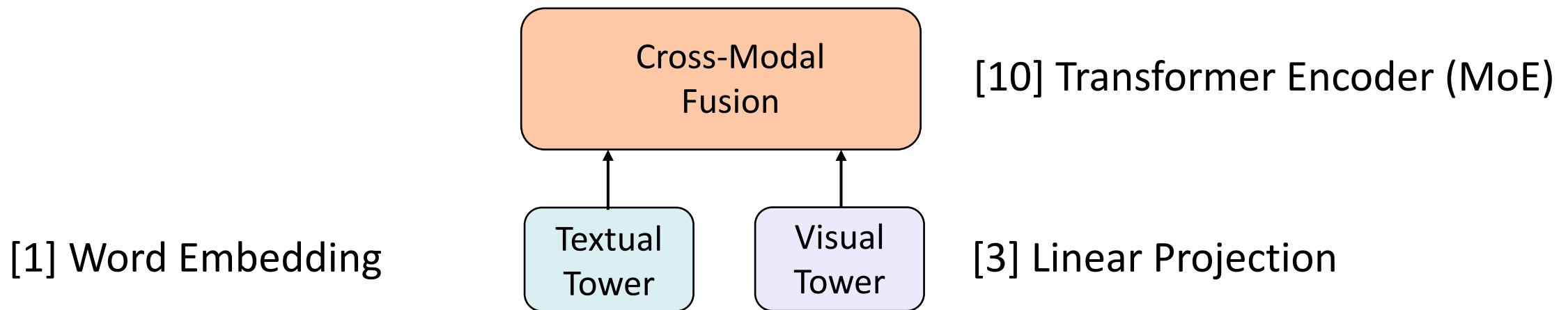


Traditional Two-Tower Architecture

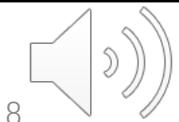


	CLIP	SOHO	Unicoder-VL	ViLT	VL-T5	METER	UniT	OFA	VLMo	BLIP
Textual-Tower	[2]	[1]	[1]	[1]	[1]	[2]	[2]	[1]	[1]	[2]
Visual-Tower	[6]	[4]	[5]	[3]	[5]	[6]	[6]	[4]	[3]	[6]
Cross-Fusion	[7]	[8]	[8]	[8]	[8] + [9]	[8]	[9]	[8] + [9]	[10]	[8] + [9]

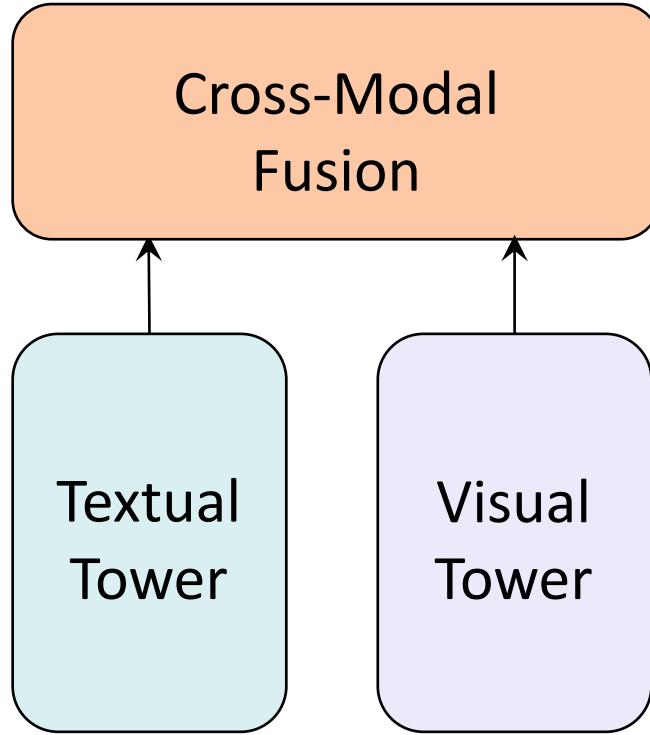
Traditional Two-Tower Architecture



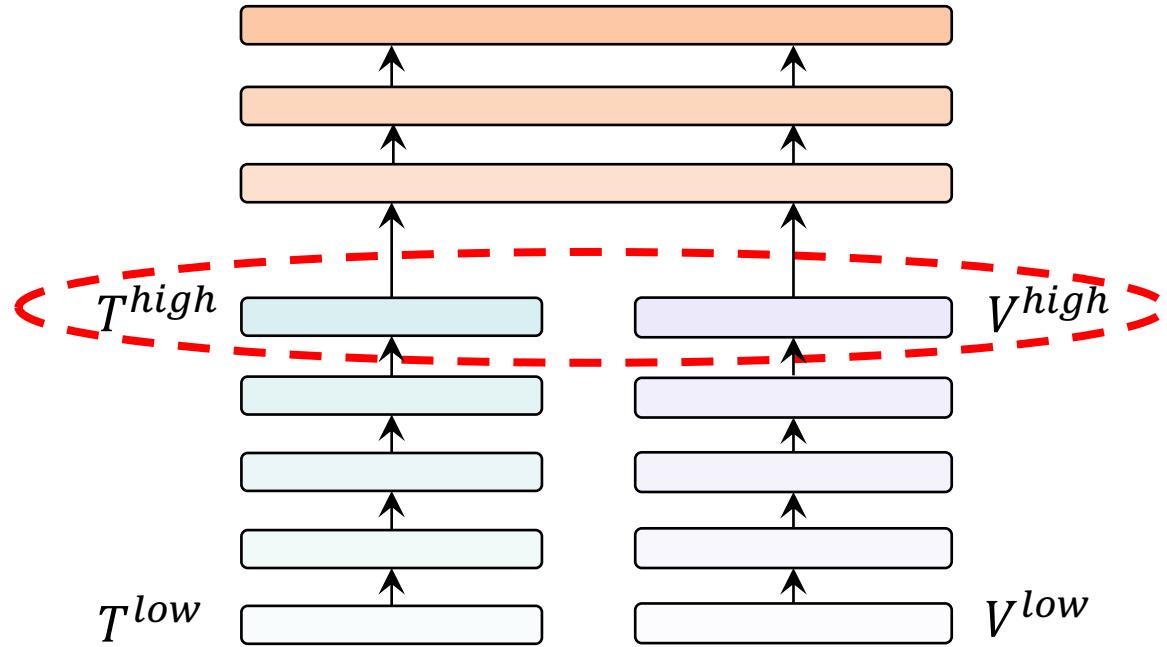
	CLIP	SOHO	Unicoder-VL	ViLT	VL-T5	METER	UniT	OFA	VLMo	BLIP
Textual-Tower	[2]	[1]	[1]	[1]	[1]	[2]	[2]	[1]	[1]	[2]
Visual-Tower	[6]	[4]	[5]	[3]	[5]	[6]	[6]	[4]	[3]	[6]
Cross-Fusion	[7]	[8]	[8]	[8]	[8] + [9]	[8]	[9]	[8] + [9]	[10]	[8] + [9]



Motivation



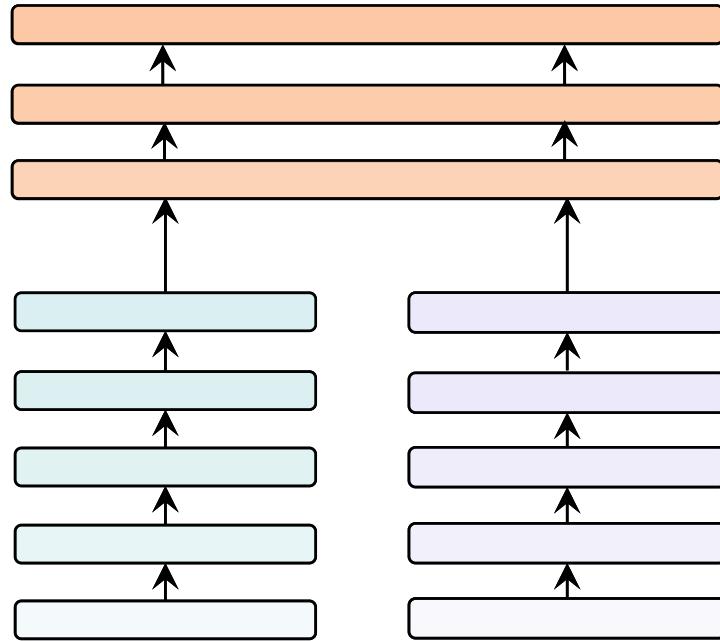
Two-Tower architecture
only use the **last-layer** uni-modal features.



Numerous works proved: **different layers** encode different types of **semantic** information.

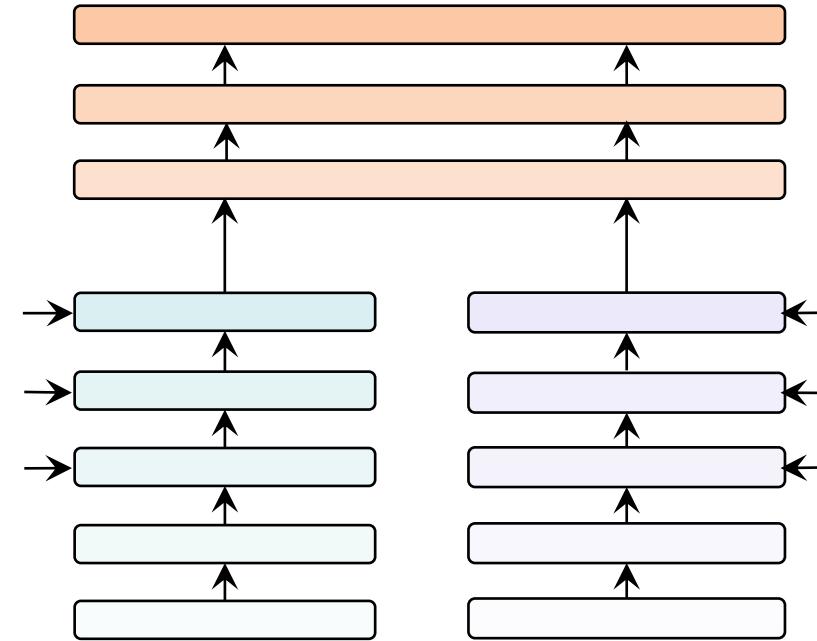
*Question: can we build a **bridge** between **different layers** of uni-modal towers and the cross-modal fusion module?*

Two-Tower vs BridgeTower



Two-Tower

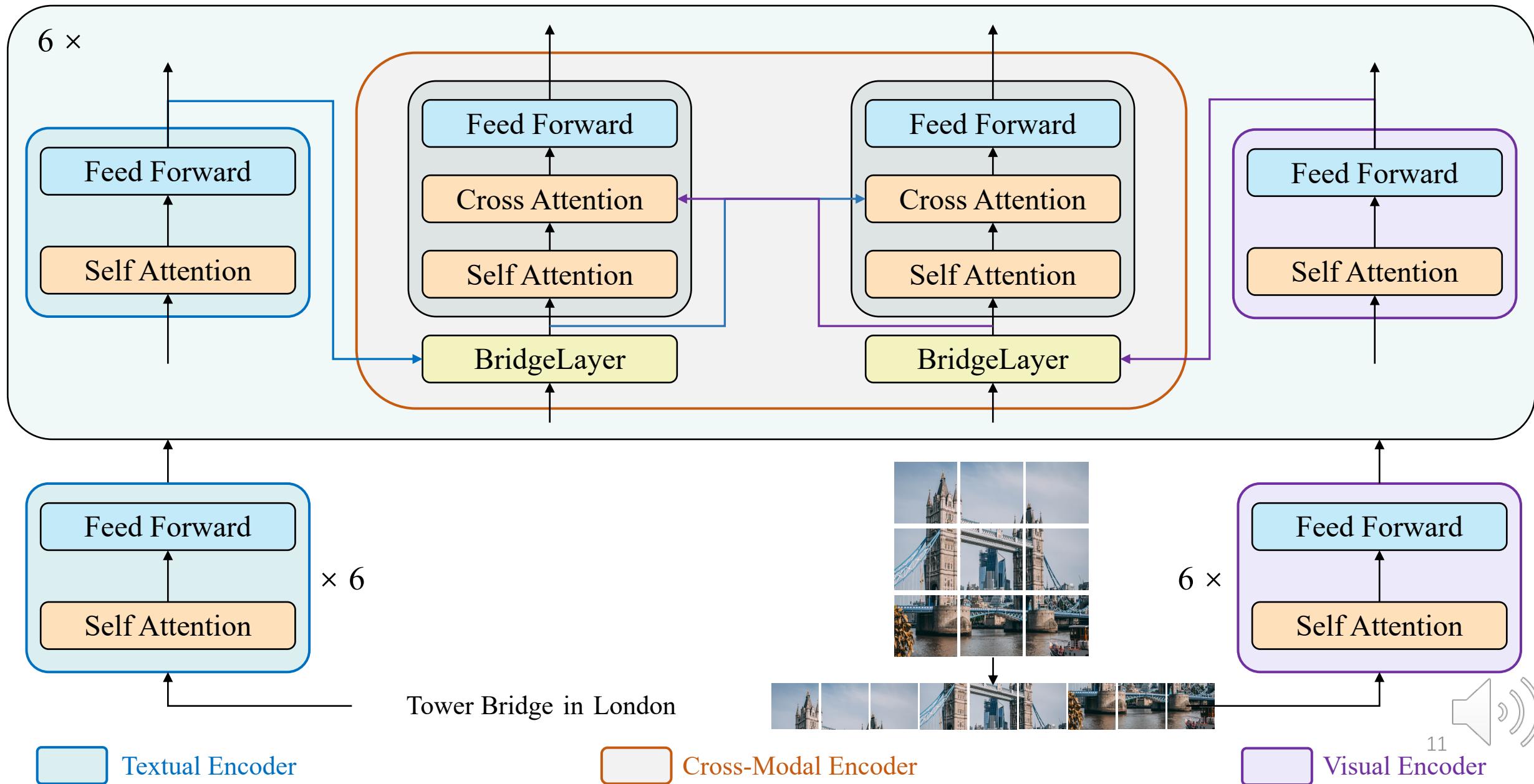
*only fuse the
last layer features*



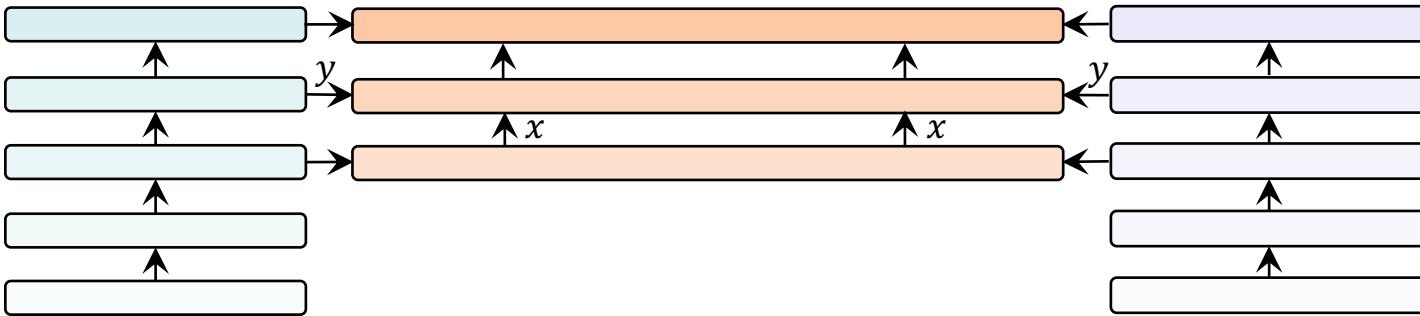
BridgeTower

*gradually fuse
multiple top layer features*

Our BridgeTower Architecture



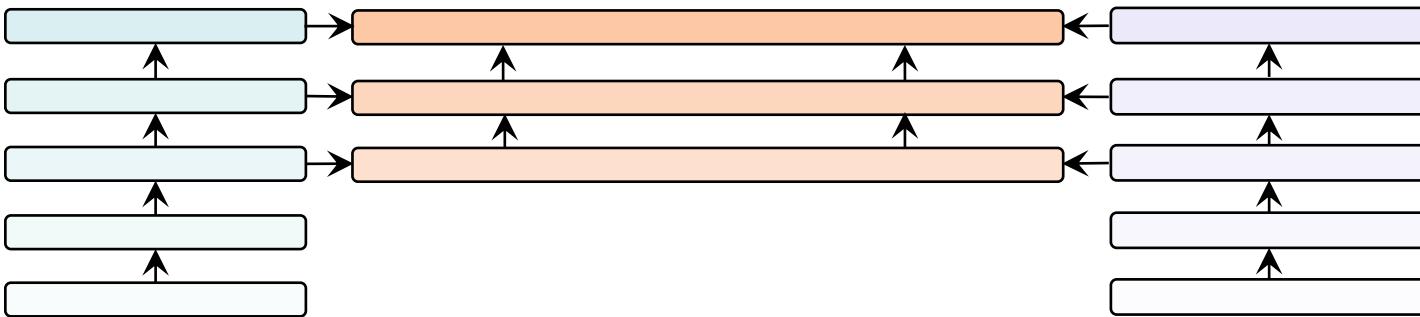
Ablation Study



Design I: Definition of Bridges



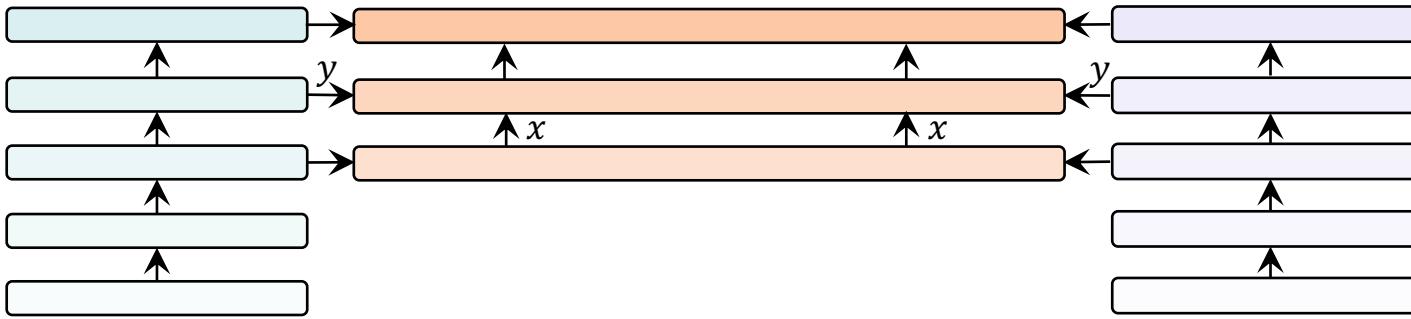
Design II: Number of Layers



Design III: Number of Bridges



Design I: Definition of Bridges

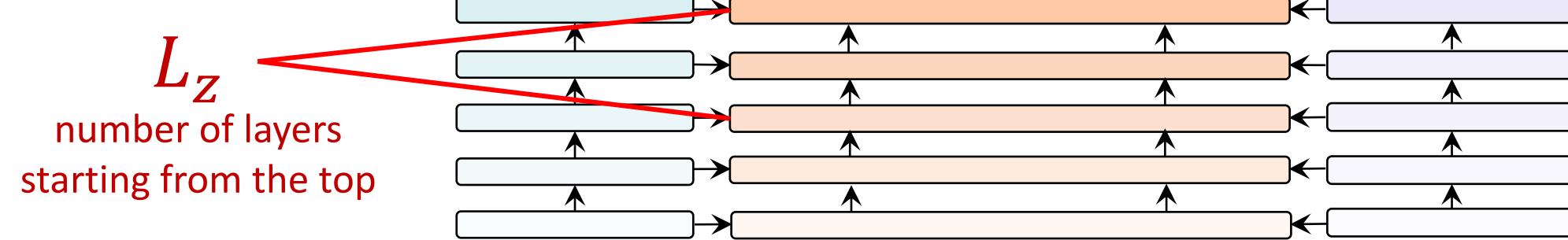


BridgeLayer(x, y)	# Params	Test-Dev	RSUM
(a) $x + y$	18.4K	75.18	533.8
(b) $x \odot y$	18.4K	73.41	530.4
(c) $\alpha x + (1 - \alpha) y, \alpha \in \mathbb{R}^{D_Z}$	26.0K	75.09	532.9
(d) $\alpha x + (1 - \alpha) y, \alpha = \sigma(\mathbf{W}[x; y])$	11.8M	75.13	533.1
(e) $\mathbf{W}[x; y]$	11.8M	74.55	532.2
(f) $\mathbf{W}_2(\text{GeLU}(\mathbf{W}_1[x; y]))$	35.4M	74.26	530.2
(g) MCA(x, y)	23.6M	73.67	514.3
(h) FFN(MCA(x, y))	70.8M	73.54	511.1
(i) $x + y + \mathbf{W}_*[x; y]$	11.8M	75.10	533.1

x: the output cross-modal representation of the previous layer

y: the corresponding uni-modal representation

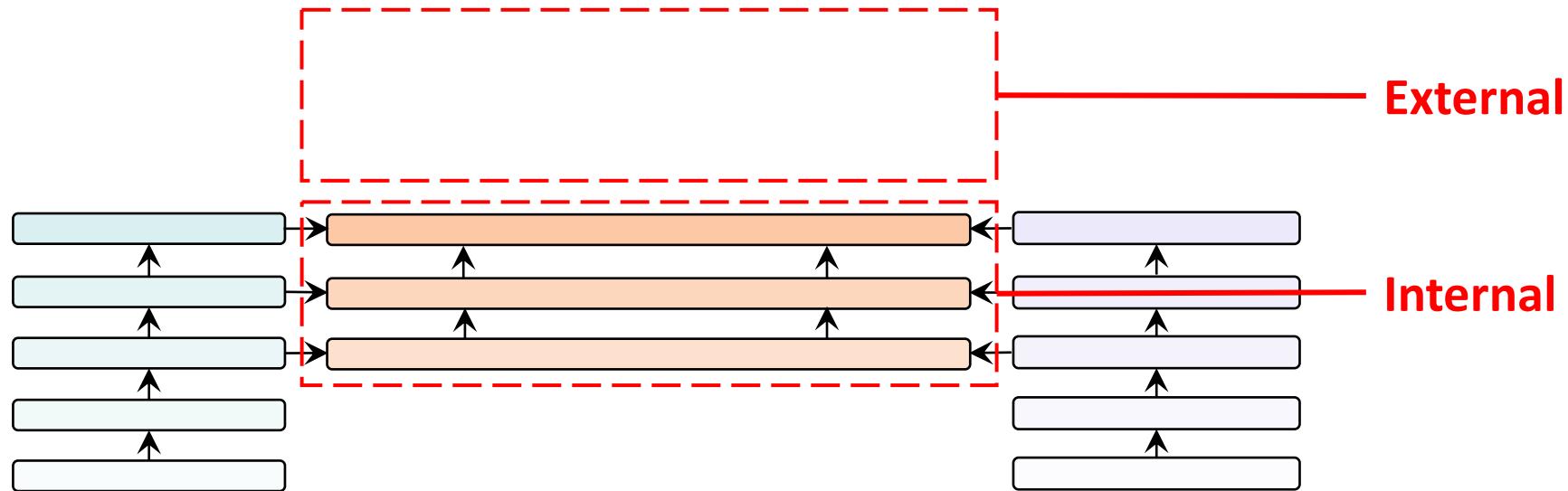
Design II: Number of Layers



L_Z	# Params	VQAv2 Test-Dev		Flickr30K RSUM	
		METER	Ours	METER	Ours
2	37.8M	72.84	74.12 (\uparrow 1.28)	526.0	527.1 (\uparrow 1.1)
3	56.8M	73.47	74.36 (\uparrow 0.89)	526.5	528.6 (\uparrow 2.1)
4	75.6M	73.71	75.00 (\uparrow 1.29)	527.9	529.7 (\uparrow 1.8)
5	94.6M	73.80	74.98 (\uparrow 1.18)	528.8	531.8 (\uparrow 3.0)
6	113.4M	74.04	75.18 (\uparrow 1.14)	530.7	533.8 (\uparrow 3.1)
8	151.2M	73.97	75.07 (\uparrow 1.10)	530.0	531.6 (\uparrow 1.6)
10	189.0M	73.45	75.06 (\uparrow 1.61)	529.6	531.7 (\uparrow 2.1)
12	226.8M	71.88	74.94 (\uparrow 3.06)	528.7	531.4 (\uparrow 2.7)

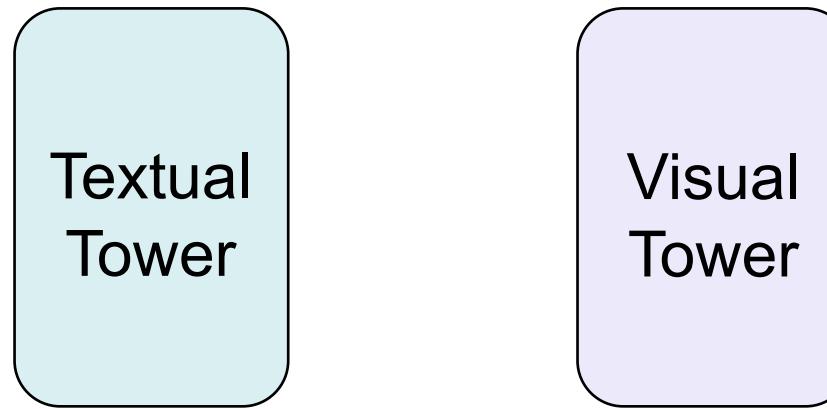


Design III: Number of Bridges



	# Internal	# External	VQAv2 Test-Dev	Flickr30K RSUM
BridgeTower	6	0	75.18	533.8
	4	2	75.06 (\downarrow 0.12)	533.1 (\downarrow 0.7)
	3	3	74.97 (\downarrow 0.21)	532.8 (\downarrow 1.0)
	2	4	74.71 (\downarrow 0.47)	532.3 (\downarrow 1.5)
Two-Tower(METER)	0	6	74.04 (\downarrow 1.14)	530.7 (\downarrow 3.1)

Apply Different Uni-modal Backbones



Visual Backbone	Textual Backbone	VQAv2 Test-Dev		Flickr30K RSUM	
		METER	Ours	METER	Ours
DeiT B-224/16	RoBERTa	69.98	70.83 (\uparrow 0.85)	448.0	455.7 (\uparrow 7.7)
ViT B-224/16	RoBERTa	70.26	72.24 (\uparrow 1.98)	472.7	476.9 (\uparrow 4.2)
ViT B-384/16	RoBERTa	70.52	72.38 (\uparrow 1.86)	472.8	477.1 (\uparrow 4.3)
CLIP-VIT-B/32	RoBERTa	72.19	72.91 (\uparrow 0.72)	508.8	512.0 (\uparrow 3.2)
CLIP-VIT-B/16	BERT	74.09	74.89 (\uparrow 0.80)	522.1	526.5 (\uparrow 4.4)
CLIP-VIT-B/16	RoBERTa	74.04	75.18 (\uparrow 1.14)	530.7	533.8 (\uparrow 3.1)

Pre-training Settings

- **Pre-training Objectives**

- Masked Language Modeling – MLM
- Image-Text Matching – ITM

- **Pre-training Datasets**

- 4M Images, ~9M Image-Text Pairs

	COCO	VG	CC	SBU
# Images	113K	108K	2.9M	860K
# Captions	567K	4.8M	2.9M	860K

Hyperparameters	BRIDGE{TOWER}_BASE	BRIDGE{TOWER}_LARGE
Number of Layers	6	6
Hidden size	768	1,024
FFN inner hidden size	3,072	4,096
Number of Attention heads	12	16
Dropout Ratio	0.1	0.1
Attention dropout	0.1	0.1
Total Steps	100k	100k
Batch Size	4,096	4,096
Textual Encoder	RoBERTa _{BASE}	RoBERTa _{LARGE}
Visual Encoder	CLIP-ViT-B	CLIP-ViT-L
Patch Size	16	14
Image Resolution	288	294



Results on VQAv2 Dataset

Model	# Pre-train Images	Visual Backbone	Test-Dev Overall	Test-Standard			
				Yes/No	Number	Other	Overall
<i>Base-Size Models</i>							
ViLT _{BASE} (Kim, Son, and Kim 2021)	4M	ViT-B-384/32	71.26	-	-	-	-
UNITER _{BASE} (Chen et al. 2020)*	4M	Faster R-CNN	72.70	-	-	-	72.91
VILLA _{BASE} (Gan et al. 2020)*	4M	Faster R-CNN	73.59	-	-	-	73.67
UNIMO _{BASE} (Li et al. 2021b)	4M	Faster R-CNN	73.79	-	-	-	74.02
ALBEF _{BASE} (Li et al. 2021a)*	4M	DeiT-B-224/16	74.54	-	-	-	74.70
ALBEF _{BASE} (Li et al. 2021a)*	14M	DeiT-B-224/16	75.84	-	-	-	76.04
VinVL _{BASE} (Zhang et al. 2021)	5.7M	ResNeXt-152	75.95	-	-	-	76.12
VLMO _{BASE} (Wang et al. 2021a)	4M	BEiT-B-224/16	76.64	-	-	-	76.89
BLIP _{BASE} (Li et al. 2022b)*	14M	DeiT-B-224/16	77.54	-	-	-	77.62
METER _{BASE} (Dou et al. 2022)	4M	CLIP-ViT-B-224/16	77.68	92.49	58.07	69.20	77.64
mPLUG (Li et al. 2022a)*	4M	CLIP-ViT-B-224/16	77.94	-	-	-	77.96
OFA _{BASE} (Wang et al. 2022b) **	54M	ResNet-101	77.98	-	-	-	78.07
SimVLM _{BASE} (Wang et al. 2021c)*	1.8B	ResNet-101	77.87	-	-	-	78.14
BLIP _{BASE} (Li et al. 2022b)*	129M	DeiT-B-224/16	78.24	-	-	-	78.17
BRIDGETOWER _{BASE} (Ours)	4M	CLIP-ViT-B-224/16	78.66	92.92	60.69	70.51	78.73
BRIDGETOWER _{BASE} (Ours)*	4M	CLIP-ViT-B-224/16	79.10	93.06	62.19	70.69	79.04
<i>Large-Size Models</i>							
UNITER _{LARGE} (Chen et al. 2020)*	4M	Faster R-CNN	73.82	-	-	-	74.02
VILLA _{LARGE} (Gan et al. 2020)*	4M	Faster R-CNN	74.69	-	-	-	74.87
UNIMO _{LARGE} (Li et al. 2021b)	4M	Faster R-CNN	75.06	-	-	-	75.27
VinVL _{LARGE} (Zhang et al. 2021)	5.7M	ResNeXt-152	76.52	92.04	61.50	66.68	76.63
SimVLM _{LARGE} (Wang et al. 2021c)	1.8B	ResNet-152	79.32	-	-	-	79.56
VLMO _{LARGE} (Wang et al. 2021a)	4M	BEiT-L-224/16	79.94	-	-	-	79.98
OFA _{LARGE} (Wang et al. 2022b) **	54M	ResNet-152	80.43	93.32	67.31	72.71	80.67
BRIDGETOWER _{LARGE} (Ours)	4M	CLIP-ViT-L-224/14	81.25	94.69	64.58	73.16	81.15
BRIDGETOWER _{LARGE} (Ours)*	4M	CLIP-ViT-L-224/14	81.52	94.80	66.01	73.45	81.49



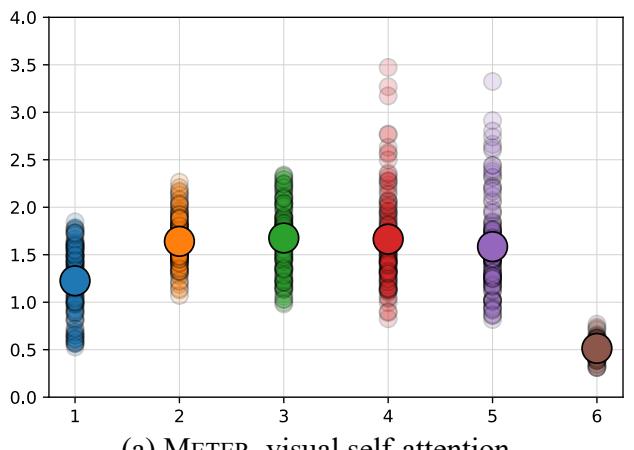
Results on VQAv2 Dataset

Model	# Pre-train Images	Visual Backbone	Test-Dev Overall	Test-Standard			Overall
				Yes/No	Number	Other	
<i>Large-Size Models</i>							
UNITER _{LARGE} (Chen et al. 2020) *	4M	Faster R-CNN	73.82	-	-	-	74.02
VILLA _{LARGE} (Gan et al. 2020) *	4M	Faster R-CNN	74.69	-	-	-	74.87
UNIMO _{LARGE} (Li et al. 2021b)	4M	Faster R-CNN	75.06	-	-	-	75.27
VinVL _{LARGE} (Zhang et al. 2021)	5.7M	ResNeXt-152	76.52	92.04	61.50	66.68	76.63
SimVLM _{LARGE} (Wang et al. 2021c)	1.8B	ResNet-152	79.32	-	-	-	79.56
VLMO _{LARGE} (Wang et al. 2021a)	4M	BEiT-L-224/16	79.94	-	-	-	79.98
OFA _{LARGE} (Wang et al. 2022b) **	54M	ResNet-152	80.43	93.32	67.31	72.71	80.67
BRIDGETOWER_{LARGE} (Ours)	4M	CLIP-ViT-L-224/14	81.25	94.69	64.58	73.16	81.15
BRIDGETOWER_{LARGE} (Ours) *	4M	CLIP-ViT-L-224/14	81.52	94.80	66.01	73.45	81.49
<i>Huge or even Larger Size Models</i>							
SimVLM _{HUGE} (Wang et al. 2021c)	1.8B	ResNet-101	80.03	93.29	66.54	72.23	80.34
METER _{HUGE} (Dou et al. 2022)	14M	Florence-CoSwin-H	80.33	94.25	64.37	72.30	80.54
mPLUG (Li et al. 2022a) *	14M	CLIP-ViT-L-224/14	81.27	-	-	-	81.26
GIT2 (Wang et al. 2022a) *	10.5B	DaViT(4.8B)	81.74	92.90	67.06	75.77	81.92
OFA _{HUGE} (Wang et al. 2022b) **	54M	ResNet-152	82.0	94.66	71.44	73.35	81.98
Flamingo (Alayrac et al. 2022) *	2.3B	NFNet-F6	82.0	-	-	-	82.1
CoCa (Yu et al. 2022) *	4.8B	ViT-G-288/18	82.3	94.55	70.25	74.46	82.33
BEiT-3 (Wang et al. 2022c)	28M	BEiT-3	84.19	96.43	73.63	75.92	84.18
PaLI (Chen et al. 2022)	1.6B	ViT-E-224	84.3	96.13	69.07	77.58	84.34

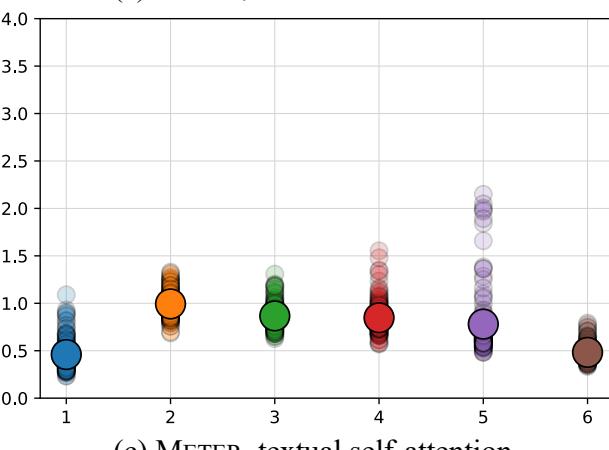
Results on SNLI-VE and Flickr30K Dataset

Model	# Pre-train Images	SNLI-VE		Flickr30K (1K test set)						RSUM
		dev	test	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	
<i>Pre-trained on More Data</i>										
ALIGN _{BASE} (Jia et al. 2021)	1.8B	-	-	84.9	97.4	98.6	95.3	99.8	100.0	576.0
ALBEF _{BASE} (Li et al. 2021a)	14M	80.80	80.91	85.6	97.5	98.9	95.9	99.8	100.0	577.7
<i>Pre-trained on CC, SBU, MSCOCO and VG datasets</i>										
ViLT _{BASE} (Kim, Son, and Kim 2021)	4M	-	-	64.4	88.7	93.8	83.5	96.7	98.6	525.7
UNITER _{LARGE} (Chen et al. 2020)	4M	79.30	79.38	75.6	94.1	96.8	87.3	98.0	99.2	550.9
VILLA _{LARGE} (Gan et al. 2020)	4M	80.18	80.02	76.3	94.2	96.8	87.9	97.5	98.8	551.5
UNIMO _{LARGE} (Li et al. 2021b)	4M	81.11	80.63	78.0	94.2	97.1	89.4	98.9	99.8	557.5
ALBEF _{BASE} (Li et al. 2021a)	4M	80.14	80.30	82.8	96.7	98.4	94.3	99.4	99.8	571.4
METER-CLIP-ViT _{BASE} (Dou et al. 2022)	4M	80.86	81.19	82.2	96.3	98.4	94.3	99.6	99.9	570.7
BRIDGETOWER_{BASE} (Ours)	4M	81.11	81.19	85.8	97.6	98.9	94.7	99.6	100.0	576.6

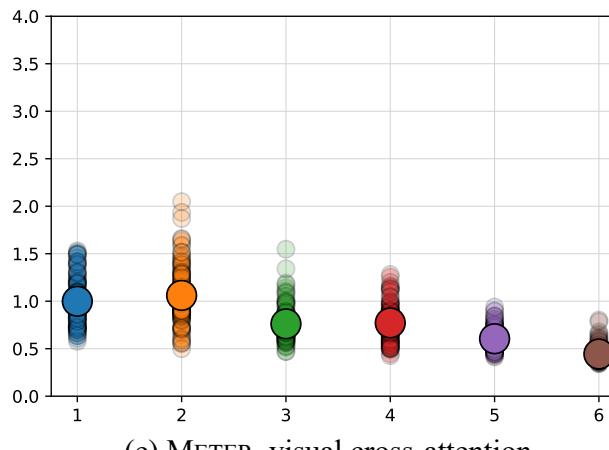
KL Divergence Visualization



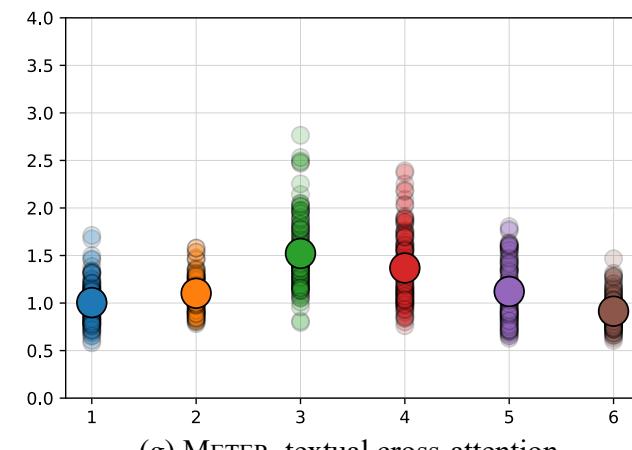
(a) METER, visual self-attention



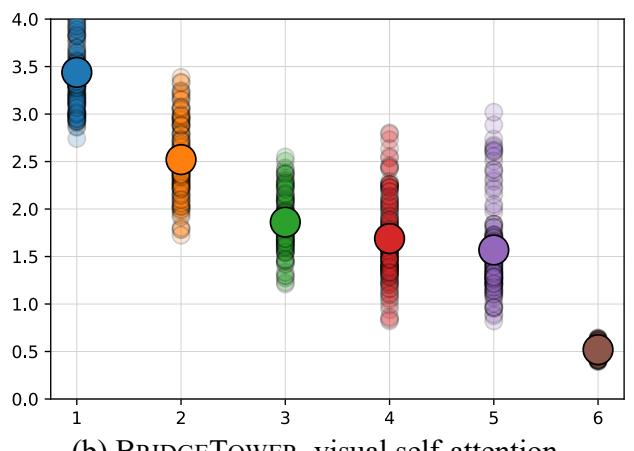
(c) METER, textual self-attention



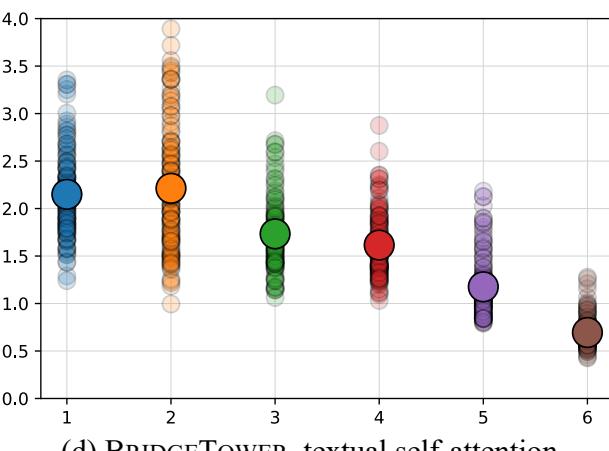
(e) METER, visual cross-attention



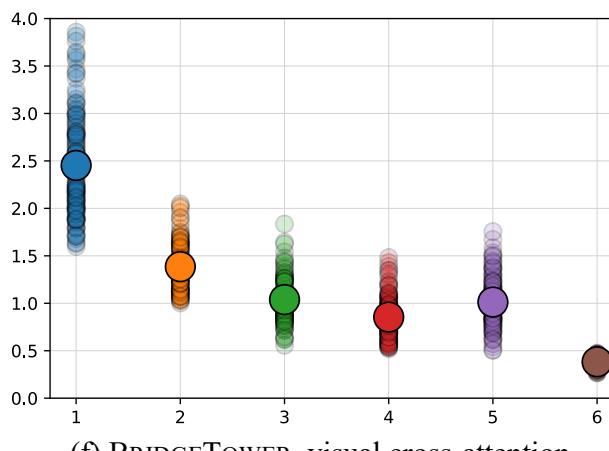
(g) METER, textual cross-attention



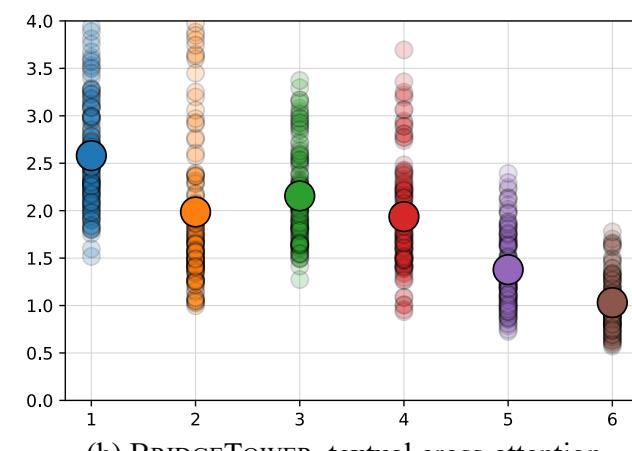
(b) BRIDGETOWER, visual self-attention



(d) BRIDGETOWER, textual self-attention



(f) BRIDGETOWER, visual cross-attention



(h) BRIDGETOWER, textual cross-attention

Higher/lower KL divergence means that different attention heads pay attention to different/similar tokens.



Conclusion & Future

- Conclusion:
 - We introduced **BridgeTower**, a **simple but effective** architecture for VL pre-training.
 - We studied different design choices for **bridges**.
 - We show that BridgeTower achieves **SOTA** results on multiple downstream tasks.
- Future:
 - More Pre-training Objectives (currently we only use **two**)
 - Larger-Scale Pre-training (currently only **4M** data)
 - More Modalities (currently only **two** modalities)

Integrated into Hugging Face – Transformers



The sidebar shows the following navigation structure:

- MAIN EN 78,811
- Image Processor
- MODELS
- TEXT MODELS
- VISION MODELS
- AUDIO MODELS
- MULTIMODAL MODELS
 - AltCLIP
 - BLIP
 - BridgeTower**
 - Chinese-CLIP
 - CLIP
 - CLIPSeg
 - Data2Vec
 - Donut
 - FLAVA
 - GIT
 - GroupViT
 - LayoutLMV2
 - LayoutLMV3
 - LayoutxLM
 - LXMERT
 - OneFormer
 - OWL-ViT
 - Perceiver
 - Speech Encoder Decoder Models
 - TrOCR
 -
 - OWL-ViT
 - Perceiver
 - Speech Encoder Decoder Models
 - TrOCR
 -

BridgeTower

Overview

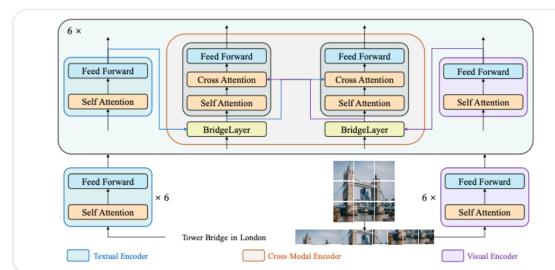
The BridgeTower model was proposed in

[BridgeTower: Building Bridges Between Encoders in Vision-Language Representative Learning](#) by Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, Nan Duan. The goal of this model is to build a bridge between each uni-modal encoder and the cross-modal encoder to enable comprehensive and detailed interaction at each layer of the cross-modal encoder thus achieving remarkable performance on various downstream tasks with almost negligible additional performance and computational costs.

This paper has been accepted to the [AAAI'23](#) conference.

The abstract from the paper is the following:

Vision-Language (VL) models with the TWO-TOWER architecture have dominated visual-language representation learning in recent years. Current VL models either use lightweight uni-modal encoders and learn to extract, align and fuse both modalities simultaneously in a deep cross-modal encoder, or feed the last-layer uni-modal representations from the deep pre-trained uni-modal encoders into the top cross-modal encoder. Both approaches potentially restrict vision-language representation learning and limit model performance. In this paper, we propose BRIDGETOWER, which introduces multiple bridge layers that build a connection between the top layers of uni-modal encoders and each layer of the crossmodal encoder. This enables effective bottom-up cross-modal alignment and fusion between visual and textual representations of different semantic levels of pre-trained uni-modal encoders in the cross-modal encoder. Pre-trained with only 4M images, BRIDGETOWER achieves state-of-the-art performance on various downstream vision-language tasks. In particular, on the VQAv2 test-std set, BRIDGETOWER achieves an accuracy of 78.73%, outperforming the previous state-of-the-art model METER by 1.09% with the same pre-training data and almost negligible additional parameters and computational costs. Notably, when further scaling the model, BRIDGETOWER achieves an accuracy of 81.15%, surpassing models that are pre-trained on orders-of-magnitude larger datasets.



BridgeTower

Overview

Usage

BridgeTowerConfig

BridgeTowerTextConfig

BridgeTowerVisionConfig

BridgeTowerImageProcessor

BridgeTowerProcessor

BridgeTowerModel

BridgeTowerForMaskedLM

BridgeTowerForImageAndText

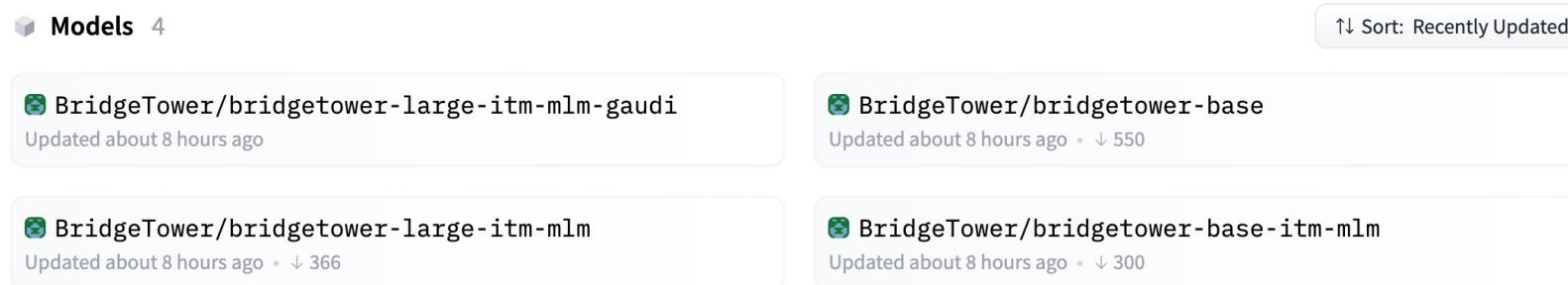
Retrieval

- Source Code: <https://github.com/huggingface/transformers/tree/main/src/transformers/models/bridgetower>
- Documentation: https://huggingface.co/docs/transformers/main/en/model_doc/bridgetower



Integrated into Hugging Face – Transformers

- Pre-trained models released on Hugging Face – Model Hub
 - <https://huggingface.co/BridgeTower>



- Model Variants
 - Number of parameters:

	Textual Encoder	Visual Encoder	Cross-Modal Encoder	Total
BridgeTower _{Base}	124M	86M	113M	323M
BridgeTower _{Large}	355M	304M	200M	859M



Usage – Image-Text Matching

```
from transformers import BridgeTowerProcessor, BridgeTowerForImageAndTextRetrieval
import requests
from PIL import Image

url = "http://images.cocodataset.org/val2017/000000039769.jpg"
image = Image.open(requests.get(url, stream=True).raw)
texts = ["An image of two cats chilling on a couch", "A football player scoring a goal"]

processor = BridgeTowerProcessor.from_pretrained("BridgeTower/bridgetower-base-itm-mlm")
model = BridgeTowerForImageAndTextRetrieval.from_pretrained("BridgeTower/bridgetower-base-itm-mlm")

# forward pass
scores = dict()
for text in texts:
    # prepare inputs
    encoding = processor(image, text, return_tensors="pt")
    outputs = model(**encoding)
    scores[text] = outputs.logits[0,1].item()

# {'An image of two cats chilling on a couch': 4.8437371253967285,
#  'A football player scoring a goal': -6.897047996520996}
```



Usage – Masked Language Modeling

```
from transformers import BridgeTowerProcessor, BridgeTowerForMaskedLM
from PIL import Image
import requests

url = "http://images.cocodataset.org/val2017/000000360943.jpg"
image = Image.open(requests.get(url, stream=True).raw).convert("RGB")
text = "a <mask> looking out of the window"

processor = BridgeTowerProcessor.from_pretrained("BridgeTower/bridgetower-base-itm-mlm")
model = BridgeTowerForMaskedLM.from_pretrained("BridgeTower/bridgetower-base-itm-mlm")

# prepare inputs
encoding = processor(image, text, return_tensors="pt")

# forward pass
outputs = model(**encoding)

results = processor.decode(outputs.logits.argmax(dim=-1).squeeze(0).tolist())

print(results)
# a cat looking out of the window.
```



Next Steps

- ❑ Pre-training Scripts for BridgeTower
- ❑ Fine-tuning Scripts and Notebooks for More Downstream Tasks



Thanks & QA

Xiao Xu^{1,2}, Chenfei Wu², Shachar Rosenman³, Vasudev Lal³, Wanxiang Che¹, Nan Duan²

¹Harbin Institute of Technology, ²Microsoft Research Asia, ³Intel Labs

Presenter: Xiao Xu

