Vision-Language Pre-training
Part II
Linjie Li
Researcher
Agenda

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• Compressing VLP models
• Robustness/causality/fairness of VLP models
• Multilingual VLP
Agenda

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• Compressing VLP models
• Robustness/causality/fairness of VLP models
• Multilingual VLP
Training Strategies in VLP

Paired Image-text Data

Model

Pre-training Task I
Pre-training Task II
Pre-training Task III

Downstream Adaptation

Model I
Model II
Model III
Model IV
Model V
Model VI
Model VII
Model VIII
Model IX
Image description generation in particular has been studied in a few recent papers [9, 11, 15, 30]. In summary, our contributions are:

which is associated with 5 human generated descriptions. The ImageClef
been created by hand in the past. The UIUC Pascal Sentence data set
contains 10k images with associated human descriptions. However neither of these collections is
large enough to facilitate reasonable image based matching necessary for our goals, as demonstrated
by our experiments on captioning with varying collection size (Sec 3). In addition this is the first –
transfer captions from our data set to a query image. Often a variety of features related to document content [23], surface [25], events [19] or feature com-
binations [28] are used in the selection process to produce sentences that reflect the most significant

A description generation method that utilizes both global representations and direct esti-
mates of image content (objects, actions, stuff, attributes, and scenes) to produce relevant
transfer captions from our data set to a query image.

Figure 1: A large novel data set containing images from the web with associated captions written by
people, filtered so that the descriptions are likely to refer to visual content. Some small collections of captioned images have
widely exist on the web?

Q: Can we leverage un-paired data that widely exist on the web?
Unsupervised VLP

• Pre-training without paired image-text data

The [MASK] is traditionally [MASK] with the same number of lit [MASK] as the age of the [MASK], or a number candle representing their age. The celebrated individual…
Unsupervised VLP

- Pre-training without paired image-text data

Supervised with paired image-text
- Model: Image+Text
- VQA: 70.87
- NLVR: 73.69
- Flickr30K: 79.80
- RefCOCO+: 72.54

Unsupervised with unpaired image-text
- Model: Image only
- VQA: 70.74 (-0.13)
- NLVR: 71.38 (-2.31)
- Flickr30K: 76.05 (-3.75)
- RefCOCO+: 71.91 (-0.63)

Unsupervised VLP (Li et al. 2021)
Training Strategies in VLP

Paired Image-text Data

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing.

Q: Can we design better pre-training tasks?

- Pre-training Task I: MLM
- Pre-training Task II: MRM
- Pre-training Task III: ITM

Downstream Adaptation

Model I
Model II
Model III
Model IV
Model V
Model VI
Model VII
Model VIII
Model IX
Contrastive Vision-Language Pre-training

UNITER (Chen et al. 2020)

- Dot Product
- Cross Attention
- [CLS] Features
- Word Features
- Region Features
Contrastive Vision-Language Pre-training

UNITER (Chen et al. 2020)

LighteningDOT (Sun et al. 2021)
Contrastive Vision-Language Pre-training

UNITER (Chen et al. 2020)

LighteningDOT (Sun et al. 2021)

Contrastive learning

Enabling real-time image-text retrieval

> 1000X speed up
Contrastive Vision-Language Pre-training

- ViLBERT [Lu, 2019] 70.92
- VisualBERT [Li, 2019] 71.00
- UNITER [Chen, 2019] 72.46
- LXMERT [Tan, 2019] 72.54
- CVLP 73.05

Supervise Learning Task
Regression Task
Self-supervise Learning Task
Contrastive Learning Task

Object Predictions
Object Attributes Predictions
RoI-Feature Regression
Masked Word Token Predictions
Image-Sentence Match Predictions
RoI-Feature Contrastive Learning

Vision-and-Language Pre-trained Model on VQA Task (Test-std)

CVLP (Shi et al. 2020)
Contrastive Vision-Language Pre-training

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLBERT [Lu, 2019]</td>
<td>70.92</td>
</tr>
<tr>
<td>VisualBERT [Li, 2019]</td>
<td>71.00</td>
</tr>
<tr>
<td>UNITER [Chen, 2019]</td>
<td>72.46</td>
</tr>
<tr>
<td>LXMERT [Tan, 2019]</td>
<td>72.54</td>
</tr>
<tr>
<td>CVLP</td>
<td>73.05</td>
</tr>
</tbody>
</table>

Vision-and-Language Pre-trained Model on VQA Task (Test-std)

CVLP (Shi et al. 2020)
This paper presents two extractive approaches for image description generation. The first uses global image representations to select relevant captions (Sec 3). The second incorporates features derived from noisy estimates of image content (Sec 5). Of course, the first requirement for any extractive model is the availability of a sufficiently large relevant document set constructed from images with associated human-generated descriptions. A large novel data set containing images from the web with associated captions written by people, filtered so that the descriptions are likely to refer to visual content.

In summary, our contributions are:

1. A large novel data set containing images from the web with associated captions written by people, filtered so that the descriptions are likely to refer to visual content.
2. Extractive methods for image description generation.
3. The UIUC Pascal Sentence data set, which contains 10k images with associated human descriptions. However, neither of these collections is large enough to serve as our document for web-scale captioning.

Studying the association between words with pictures has been explored in a variety of tasks, including:

- Labeling faces in news photographs with associated captions [2].
- Finding a correspondence between keywords and image regions [1, 6], or for moving beyond objects to mid-level recognition [8, 16, 17, 12].
- Between keywords and image regions [1, 6], or for moving beyond objects to mid-level recognition [8, 16, 17, 12].
- Often a variety of features related to document content [23], surface [25], events [19] or feature combinations [28] are used in the selection process to produce sentences that reflect the most significant concepts in the document.

Image description generation in particular has been studied in a few recent papers [9, 11, 15, 30]. Typically, models are trained on small collections of paired images with associated captions written by humans. Often these models are trained on data sets with a specific task in mind (e.g., multi-choice questions about images [14] or image captioning [27]). Even when the task of the model is similar to the task in which it will be used, the model is trained on a different data set.

One of the challenges of image captioning is the generation of short, but informative descriptions. Often a model is trained on data sets consisting of a few hundred or thousand images with associated human-generated captions. Generating a caption for a query picture that summarizes the image content is one of the major tasks in the field of computer vision. In our photo captioning problem, we would like to generate a caption for a query picture that summarizes the image content. We do this by considering a large relevant document set constructed from noisy estimates of image content.

Figure 1: Paired Image-text Data

Training Strategies in VLP

Paired Image-text Data

Model

Pre-training Task I: MLM
Pre-training Task II: MRM
Pre-training Task III: ITM

Q: Can we have one model for all?

Model I
Model II
Model III
Model IV
Model V
Model VI
Model VII
Model VIII
Model IX

Downstream Adaptation
Multi-task Training

All these task require visually-grounded language understanding skills.

12-in-1 (Lu et al. 2020)
Multi-task Training

<table>
<thead>
<tr>
<th>Model</th>
<th># models</th>
<th># parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independently train for each task</td>
<td>12</td>
<td>12X270M = 3B</td>
</tr>
</tbody>
</table>

12-in-1 (Lu et al. 2020)
# Multi-task training

<table>
<thead>
<tr>
<th>Model</th>
<th># models</th>
<th># parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independently train for each task</td>
<td>12</td>
<td>12X270M = 3B</td>
</tr>
<tr>
<td>Train all tasks together (12-in-1)</td>
<td>1</td>
<td>270M</td>
</tr>
</tbody>
</table>
Multi-task training

All these task share similar visually-grounded language understanding skills.
Training Strategies in VLP

Paired Image-text Data

Model

Pre-training Task I: MLM
Pre-training Task II: MRM
Pre-training Task III: ITM

Downstream Adaptation

Model I
Model II
Model III
Model IV
Model V
Model VI
Model VII
Model VIII
Model IX

Q: Can we tackle all VL tasks with a single objective?
Unifying VL Tasks via Text Generation

**UNITER**
(Chen et al. 2020)

Credit to: Jaemin Cho
Unifying VL Tasks via Text Generation

UNITER
(Chen et al. 2020)

VL-T5/-BART
(Cho et al. 2021)

Credit to: Jaemin Cho
Unifying VL Tasks via Text Generation

**UNITER**
(Chen et al. 2020)

**VL-T5-/BART**
(Cho et al. 2021)

Credit to: Jaemin Cho
Unifying VL Tasks via Text Generation

(a) Our vision-and-language framework

VL-T5/-BART (Cho et al. 2021)
Agenda

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• Compressing VLP models
• Robustness/causality/fairness of VLP models
• Multilingual VLP
Great success of VLP models

Can we apply VLP to other VL tasks?
Great success of VLP models

Can we apply VLP to other VL tasks?  YES!
Diverse Applications of VLP

Visual Dialog (Murahari et al. 2021)

VisDial-BERT (Wang et al. 2020)

VD-BERT (Wang et al. 2020)
Diverse Applications of VLP

Vision Language Navigation
(Anderson et al. 2017)

PREVALENT (Hao et al. 2020)
Diverse Applications of VLP

Novel Object Captioning
(Agrawal et al. 2019)

VIVO (Hu et al. 2020)
Diverse Applications of VLP

Masked language modeling (MLM)
- logo
- clock
- coors
- \( f^w_{\text{mask}} \)

Relative position prediction (RPP)
- relative position: "left"
- \( f_{\text{obj}} \)
- \( f_{\text{ocr}} \)

Image-text matching (ITM)
- Matched: 0/1
- \( f^P \)
- \( \text{SAFEWAY} \)

Text-VQA (Singh et al. 2019)
- Question: what number is on the bike on the right? ---- A: the number is 317

Text-Captioning (Sidorov et al. 2020)
Diverse Applications of VLP

Q: Which American president is associated with the stuffed animal seen here?
A: Teddy Roosevelt

OK-VQA (Sidorov et al. 2019)

Reasoning over Vision and Language:
Exploring the Benefits of Supplemental Knowledge
(Shevchenko et al. 2021)
Diverse Applications of VLP

Application: Fashion Product Searching System

Kaleido-BERT (Zhuge et al. 2021)

Pre-training Kaleido-BERT (Sec. 3.6)
- Task #3: AKPM (with five sub-tasks)
- Task #2: TIM
- Task #1: AMLM

Cross-Modality Transformer (Sec. 3.5)

AGM (Sec. 3.4)
- Winter [MASK] fur collar coats [MASK] girl

Aligned Information

AAG (Sec. 3.3)

KPG (Sec. 3.2)
- Fashion Caption: Winter faux fur collar coats for girl
TAP: Text-Aware Pre-training

• Towards VLP model that can read

Question: what **number** is on the bike on the right? ---- A: the number is **317**

A group of motorcyclists with **number** **317, 44, 30, 338, 598** racing outdoor.

Text-VQA (Singh et al. 2019)  Text-Captioning (Sidorov et al. 2020)

Credit to: Zhengyuan Yang
TAP: Text-Aware Pre-training

Credit to: Zhengyuan Yang
TAP: Text-Aware Pre-training

Credit to: Zhengyuan Yang
TAP: Text-Aware Pre-training

Credit to: Zhengyuan Yang
# TAP: Text-Aware Pre-training

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TextVQA</th>
<th>ST-VQA</th>
<th>TextCaps</th>
<th>CC-OCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training images</td>
<td><strong>22K</strong></td>
<td><strong>19K</strong></td>
<td><strong>22K</strong></td>
<td><strong>1.37M</strong></td>
</tr>
<tr>
<td>Text</td>
<td>35K OCR-QA pairs</td>
<td>26K OCR-QA pairs</td>
<td>110K OCR-Caption</td>
<td>One caption per image</td>
</tr>
<tr>
<td>Image source</td>
<td>Open Image</td>
<td>ICDAR 2013/15, ImageNet, VizWiz, IIIT Scene Text Retrieval, Visual Genome, COCO-Text</td>
<td>TextVQA</td>
<td>Conceptual captions</td>
</tr>
</tbody>
</table>

TAP (Yang et al. 2021)

Credit to: Zhengyuan Yang
TAP: Text-Aware Pre-training

Credit to: Zhengyuan Yang

TAP (Yang et al. 2021)
Agenda

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• Compressing VLP models
• Robustness/causality/fairness of VLP models
• Multilingual VLP
Great success of VLP models

Can VLP help unimodal tasks?

V+L Tasks
- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning
VL for V/L

- **VL for V**: A scalable way to learn visual representations
  - SOTA computer vision models rely on carefully annotated labels/bounding boxes for learning
  - Self-supervised learning is scalable, but supervision signal is weak
  - Image-text pairs widely exist on the web
VL for V/L

- **VL for V**: A scalable way to learn visual representations
  - Early attempts on using human annotated image-text pairs: COCO, VG

**ICMLM** (Sariyildiz et al. 2020)

**VirTex** (Desai and Johnson, 2020)
**VL for V/L**

- **VL for V**: A scalable way to learn visual representations
  - Scaling up to billions of web-crawled image alt-text data!

**CLIP** (Radford et al. 2021)

~400M image alt-text pairs!

**ALIGN** (Jia et al., 2020)

Over one billion image alt-text pairs!!!
**VL for V/L**

- **VL for V**: A scalable way to learn visual representations
  - Scaling up to billions of web-crawled image alt-text data!
  - Achieving strong performance while closes the “robustness gap” by up to 75%

<table>
<thead>
<tr>
<th>DATASET</th>
<th>IMAGENET RESNET101</th>
<th>CLIP VIT-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>76.2%</td>
<td>76.2%</td>
</tr>
<tr>
<td>ImageNet V2</td>
<td>64.3%</td>
<td>70.1%</td>
</tr>
<tr>
<td>ImageNet Rendition</td>
<td>37.7%</td>
<td>88.9%</td>
</tr>
<tr>
<td>ObjectNet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageNet Sketch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageNet Adversarial</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CLIP (Radford et al. 2021)
VL for V/L

- **VL for L**: Great potential to enhance language representations
  - A picture is worth a thousand words

**Vokenization** (Tan and Bansal, 2020)
Vokenization

- Goal: improving language understanding with contextualized, visual-grounded supervision

Vokenization (Tan and Bansal, 2020)
Vokenization

- Goal: Improving Language Understanding with Contextualized, Visual-Grounded Supervision

How to generate vokens?

Voken Classification Task

How to generate vokens?

Voken Classification Task

Vokenization (Tan and Bansal, 2020)
Vokenization

- Vokenization process: assign each token with a relevant image

Vokenization (Tan and Bansal, 2020)
Vokenization

- Improve language understanding with related visual information

+2.7% average improvement

Learns token-image alignment

<table>
<thead>
<tr>
<th>Method</th>
<th>SST-2</th>
<th>QNLI</th>
<th>QQP</th>
<th>MNL1</th>
<th>SQuAD v1.1</th>
<th>SQuAD v2.0</th>
<th>SWAG</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT_{6L/512H}</td>
<td>88.0</td>
<td>85.2</td>
<td>87.1</td>
<td>77.9</td>
<td>71.3/80.2</td>
<td>57.2/60.8</td>
<td>56.2</td>
<td>75.6</td>
</tr>
<tr>
<td>BERT_{6L/512H} + Voken-cls</td>
<td>89.7</td>
<td>85.0</td>
<td>87.3</td>
<td>78.6</td>
<td>71.5/80.2</td>
<td>61.3/64.6</td>
<td>58.2</td>
<td>76.8</td>
</tr>
<tr>
<td>BERT_{12L/768H}</td>
<td>89.3</td>
<td>87.9</td>
<td>83.2</td>
<td>79.4</td>
<td>77.0/85.3</td>
<td>67.7/71.1</td>
<td>65.7</td>
<td>79.4</td>
</tr>
<tr>
<td>BERT_{12L/768H} + Voken-cls</td>
<td>92.2</td>
<td>88.6</td>
<td>88.6</td>
<td>82.6</td>
<td>78.8/86.7</td>
<td>68.1/71.2</td>
<td>70.6</td>
<td>82.1</td>
</tr>
</tbody>
</table>

Example 1: Humans learn language by listening, speaking, writing, reading

Vokenization (Tan and Bansal, 2020)
Video-Language for Language

Video-aided Unsupervised Grammar Induction

Songyang Zhang\textsuperscript{1}, Linfeng Song\textsuperscript{2}, Lifeng Jin\textsuperscript{2}, Kun Xu\textsuperscript{2}, Dong Yu\textsuperscript{2} and Jiebo Luo\textsuperscript{1}
\textsuperscript{1}University of Rochester, Rochester, NY, USA
szhang83@ur.rochester.edu, jluo@cs.rochester.edu
\textsuperscript{2}Tencent AI Lab, Bellevue, WA, USA
{lfsong,lifengjin,kxkunxu,dyu}@tencent.com

NAACL 2021 Best Long Paper

Pre-training

Sentence: A squirrel jumps on stump.

Downstream

Parse tree for the sentence “The man falls to the floor.”
Agenda

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• **Compressing VLP models**
• Robustness/causality/fairness of VLP models
• Multilingual VLP
Great success of VLP models

How about efficiency? Can we compress a large VLP model while preserving its performance and transferability?
Model Compression Technique

• Low-rank approximation and sparsity
• Neural Architecture Search
• Knowledge distillation
• Pruning
• Quantization
Compressing VLP Models

- Low-rank approximation and sparsity
- Neural Architecture Search
- Knowledge distillation
- Pruning
- Quantization
Compressing VLP Models via Distillation

- Large VLP Model
  - Region feature extractor
    - R101 (BUTD, 2017)
    - X152 (VinVL, 2021)
  - Transformer
    - BERT\text{BASE} (12/768/3072)
    - BERT\text{LARGE}

- Compact VLP Model
  - Region feature extractor
    - Two-stage Efficient Extractor (TEE)
  - Transformer
    - MiniLM (12/384/1536)

Credit to: Jianfeng wang
Compressing VLP Models via Distillation

• Large VLP Model (Teacher)
  • Region feature extractor
    • R101 (BUTD, 2017)
    • X152 (VinVL, 2021)
  • Transformer
    • BERT\textsubscript{BASE} (12/768/3072)
    • BERT\textsubscript{LARGE}

• Compact VLP Model (Student)
  • Region feature extractor
    • Two-stage Efficient Extractor (TEE)
  • Transformer
    • MiniLM (12/384/1536)

Credit to: Jianfeng wang
Compressing VLP Models via Distillation

DistillVLM (Fang et al. 2021)
Compressing VLP Models via Distillation

DistillVLM (Fang et al. 2021)
Compressing VLP Models via Distillation

Adapted Teacher VLM

Teacher VLM → Strong Detector → Visual Token Extraction → Object Proposal → Light Detector → Student VLM

Transformer Block 1 → Transformer Block 2 → ... → Transformer Block L

Self-Attention Heads: 

\[ \mathcal{L}_{ATT} \]

FC Layer: 

\[ \mathcal{L}_{MLM} + \mathcal{L}_{ITM} + \mathcal{L}_{LOGITS} \]

DistillVLM (Fang et al. 2021)
Compressing VLP Models via Distillation

DistillVLM (Fang et al. 2021)

Distillation on (1) logits, (2) hidden states and (3) attention matrix
## Compressing VLP Models via Distillation

DistillVLM (Fang et al. 2021)

<table>
<thead>
<tr>
<th>Method</th>
<th># Param</th>
<th># I-T Pairs</th>
<th>VisualFeat.</th>
<th>P. D.</th>
<th>F. D.</th>
<th>COCO Captioning</th>
<th>VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B@4   M   C   S</td>
<td>test-std test-dev</td>
</tr>
<tr>
<td>UVLP [74]</td>
<td>111.7M</td>
<td>3M</td>
<td>ResNeXt101</td>
<td>✗</td>
<td>✗</td>
<td>36.5  28.4 116.9 21.2</td>
<td>70.7 –</td>
</tr>
<tr>
<td>OSCAR$_B$ [37]</td>
<td>111.7M</td>
<td>7M</td>
<td>R101-F</td>
<td>✗</td>
<td>✗</td>
<td>36.5  30.3 123.7 23.1</td>
<td>73.4 73.2</td>
</tr>
<tr>
<td>MiniVLM [65]</td>
<td>34.5M</td>
<td>7M</td>
<td>TEE</td>
<td>✗</td>
<td>✗</td>
<td>34.3  28.1 116.7 21.3</td>
<td>– –</td>
</tr>
<tr>
<td>MiniVLM [65]</td>
<td>34.5M</td>
<td>14M</td>
<td>TEE</td>
<td>✗</td>
<td>✗</td>
<td>35.6  28.6 119.8 21.6</td>
<td>69.4 69.1</td>
</tr>
<tr>
<td>DistillVLM</td>
<td>34.5M</td>
<td>7M</td>
<td>TEE</td>
<td>✗</td>
<td>✗</td>
<td>34.0  28.0 115.7 21.1</td>
<td>69.0 68.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✗</td>
<td>✓</td>
<td>34.5  28.2 117.1 21.5</td>
<td>69.2 69.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✗</td>
<td>35.2  28.6 120.1 21.9</td>
<td>69.7 69.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>35.6  28.7 120.8 22.1</td>
<td>69.8 69.6</td>
</tr>
</tbody>
</table>

DistillVLM improves over MiniVLM, but still more compact and faster than large VLP models!
Compressing VLP Models via Pruning

- A popular direction: *lottery ticket hypothesis*

What to prune?  How to prune?  How often?  When to prune?

Unstructured  Magnitude  Iterative  'Before'

Image credit to: https://roberttlange.github.io/posts/2020/06/lottery-ticket-hypothesis/
Lottery Ticket Hypothesis

• *ICLR 2019 Best Paper by MIT*: The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

• **LTH**: A randomly initialized, dense neural network contains a subnetwork that is initialized such that — when trained in isolation — it can match the test accuracy of the original network after training for at most the same number of iterations. - Frankle & Carbin (2019, p.2)

• An emerging sub-field in deep learning regarding sparse neural networks

https://arxiv.org/cs

The Lottery Ticket Hypothesis: Finding Sparse, Trainable ...

by J Frankle · 2018 · Cited by 314 — Based on these results, we articulate the "lottery ticket hypothesis:" dense, randomly-initialized, feed-forward networks contain subnetworks ("winning tickets") ...

Playing Lottery Tickets with Vision and Language

- **Existence**: Can we draw VLP winning tickets successfully for VL downstream tasks?

- **Transferability**: Can we find tickets that transfer universally to all downstream VL tasks?

- **Compatibility**: Can we find tickets compatible with adversarial training to enhance the performance?
Playing Lottery Tickets with Vision and Language

Algorithm 1 Iterative Magnitude Pruning for V+L Tickets.

**Input** Initial mask $m = 1^d$; Pre-trained parameters $\theta_0$ and task-specific parameters $\phi_i$; rewinding step $i$ (could be 0), sparsity level $s$, total training step $t$.

Train the pre-trained V+L model $f(x; m \odot \theta_0, \phi_0)$ to step $i$: $f(x; m \odot \theta_i, \phi_i)$.

**repeat**

Train $f(x; m \odot \theta_i, \phi_i)$ to step $t$: $f(x; m \odot \theta_t, \phi_t)$.

Prune 10% of non-zero weights of $m \odot \theta_t$ based on the magnitudes and update $m$ accordingly.

**until** the sparsity of $m$ reaches $s$

**Return** $f(x; m \odot \theta_s, \cdot) = 0$

A **winning ticket** is a sub-network that matches the performance of the original full dense network.

Playing lottery tickets with VL (Gan et al. 2021)
Playing Lottery Tickets with Vision and Language

- **Existence**: Can we draw VLP winning tickets successfully for VL downstream tasks?

- **VLM can play lottery tickets too**: We confirm that “relaxed” winning tickets that match 99% of the full accuracy can be found at 50%-70% sparsity across all the tasks.

- **Transferability**: Can we find tickets that transfer universally to all downstream VL tasks?

- **One ticket to win them all**: Matching subnetworks found via IMP on pre-training tasks transfer universally. Unexpectedly, matching subnetworks found via IMP on each downstream task also transfer to other tasks reasonably well.

- **Compatibility**: Can we find tickets compatible with adversarial training to enhance the performance?

- **Enhancing tickets with adversarial training**: Though the found winning tickets are sparse neural networks, adversarial training can be still helpful to enhance the performance across all the tasks considered.
Agenda

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• Compressing VLP models
• Robustness/fairness of VLP models
• Multilingual VLP
Great success of VLP models

How robust are these pre-trained V+L Models?

V+L Tasks
- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning
**Similar Data Distribution**

![Graph showing similar data distribution between training and test sets]

**Little-to-None Linguistic Variations**

- **Original**: Q: What is in the basket? A: Remote
- **Rephrasing**: Q: What can be seen inside the basket? A: Remote
- **Logical Transformation**: Q: Is remote in the basket? A: Yes

**Without Visual Content Manipulations**

- COCO: Common Objects in Context
- Movie Clips

**Standard V+L Tasks**

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning
Robust VQA Benchmarks

• Compilation of 9 diverse VQA datasets covering 4 types of robustness
• Note that robustness here *is not* adversarial robustness

**Linguistic Variation (Lingual)**
- VQA-Rephrasings

**Logical Reasoning (Reason)**
- VQA-LOL Compose
- VQA-LOL Supplement
- VQA-Introspect
- GQA

**Visual Content Manipulation (Visual)**
- IV-VQA
- CV-VQA

**Answer Distribution Shift (Answer)**
- VQA-CP v2
- GQA-OOD

MANGO (Li et al. 2020)
# Robust VQA Benchmarks

<table>
<thead>
<tr>
<th>Type</th>
<th>Benchmark</th>
<th>Metric</th>
<th>Q Type</th>
<th>Train Source</th>
<th>#IQ</th>
<th>len(Q)</th>
<th>Val #IQ</th>
<th>len(Q)</th>
<th>Test #IQ</th>
<th>len(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GQA [26]</td>
<td>Acc.</td>
<td>All</td>
<td>-</td>
<td>943K</td>
<td>8.76</td>
<td>132K</td>
<td>8.77</td>
<td>13K</td>
<td>8.51</td>
</tr>
<tr>
<td>Visual</td>
<td>IV-VQA [2]</td>
<td>#flips</td>
<td>All</td>
<td>VQA v2 train</td>
<td>444K</td>
<td>6.20</td>
<td>120K</td>
<td>5.85</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GQA-OOD [32]</td>
<td>Acc.</td>
<td>All</td>
<td>GQA train</td>
<td>943K</td>
<td>8.76</td>
<td>51K</td>
<td>8.09</td>
<td>3K</td>
<td>7.70</td>
</tr>
</tbody>
</table>

**Table 1:** Detailed descriptions of each downstream benchmark, including robustness type, evaluation metric, question type, training data source and statistics on train, val, test data in terms of number of Image-Question pairs (#IQ) and average question length (len(Q)). We use the training data provided with the benchmark unless specified otherwise. Results on val split are reported when test split is not available. Acc. is short for Accuracy. M✓S✓ is a consistency measure between main questions and sub-questions in VQA-Introspect. #flips is the number of predictions mismatched before and after visual content manipulation.
MANGO Framework

- Adversarial Noise Generator

MANGO (Li et al. 2020)
MANGO Framework

- Adversarial Noise Generator

Minimize for V+L models

Maximize for Adv. Noise Generator
MANGO Framework

• Adversarial Noise Generator

$$\min_{\theta} \max_{\phi_v} \mathbb{E}_{(v,w,y) \sim D} \mathbb{E}_{\alpha \sim \mathcal{N}(0,1)} \left[ \mathcal{L}_{std}(\theta, \phi_v) + \beta \mathcal{R}_{at}(\theta, \phi_v) \right]$$

$$\mathcal{L}_{std}(\theta, \phi_v) = \mathcal{L}_{BCE}(f_\theta(v, w), y)$$

$$\mathcal{R}_{at}(\theta, \phi_v) = \mathcal{L}_{BCE}(f_\theta(v + g_\phi_v(\alpha), w), y) + \mathcal{L}_{kl}(f_\theta(v + g_\phi_v(\alpha), w), f_\theta(v, w))$$

VQA task loss on clean inputs

VQA task loss on perturbed inputs

KL Divergence between clean inputs and perturbed inputs
MANGO Framework

• Adversarial Noise Generator

\[
\min_{\theta} \max_{\phi_v} \mathbb{E}_{v,w,y \sim D} \mathbb{E}_{\alpha \sim \mathcal{N}(0,1)} [\mathcal{L}_{std}(\theta, \phi_v) + \beta \mathcal{R}_{at}(\theta, \phi_v)]
\]

\[
\mathcal{L}_{std}(\theta, \phi_v) = \mathcal{L}_{\text{BCE}}(f_\theta(v, w), y)
\]

\[
\mathcal{R}_{at}(\theta, \phi_v) = \mathcal{L}_{\text{BCE}}(f_\theta(v + g_{\phi_v}(\alpha), w), y)
+ \mathcal{L}_{kl}(f_\theta(v + g_{\phi_v}(\alpha), w), f_\theta(v, w))
\]

Perturbations generated via a small neural network
MANGO Framework

• Random Masking

Motivation: significant mismatch in the distribution of question lengths and image regions between training and test splits of robustness benchmarks
## Experimental Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA</td>
<td>N/A</td>
<td></td>
<td>56.59</td>
<td>48.99</td>
<td>50.54</td>
<td>50.05</td>
<td>63.17</td>
<td>7.53</td>
<td>78.44</td>
<td>69.52</td>
<td>52.70</td>
<td>74.69</td>
</tr>
<tr>
<td>UNITER _B</td>
<td>40.98</td>
<td>64.56</td>
<td>54.54</td>
<td>50.00</td>
<td>56.80</td>
<td>59.99</td>
<td>60.65</td>
<td>8.47</td>
<td>40.67</td>
<td>69.33</td>
<td>53.43</td>
<td>72.70</td>
</tr>
<tr>
<td>MANGO _B</td>
<td>42.80</td>
<td>65.80</td>
<td>56.22</td>
<td>56.49</td>
<td>58.33</td>
<td>60.65</td>
<td>60.26</td>
<td>7.43</td>
<td>38.25</td>
<td>48.38</td>
<td>55.79</td>
<td>3.45</td>
</tr>
<tr>
<td>VILLA _B</td>
<td>42.37</td>
<td>65.91</td>
<td>55.44</td>
<td>57.58</td>
<td>58.94</td>
<td>60.73</td>
<td>60.26</td>
<td>6.69</td>
<td>35.52</td>
<td>52.76</td>
<td>56.40</td>
<td>74.26</td>
</tr>
<tr>
<td>MANGO _VB</td>
<td>43.08</td>
<td>65.91</td>
<td>55.44</td>
<td>57.58</td>
<td>58.94</td>
<td>60.73</td>
<td>60.26</td>
<td>6.69</td>
<td>35.52</td>
<td>52.76</td>
<td>56.40</td>
<td>74.26</td>
</tr>
<tr>
<td>UNITER _L</td>
<td>43.37</td>
<td>67.64</td>
<td>58.60</td>
<td>60.50</td>
<td>62.14</td>
<td>61.10</td>
<td>45.27</td>
<td>68.33</td>
<td>59.45</td>
<td>68.16</td>
<td>58.66</td>
<td>74.26</td>
</tr>
<tr>
<td>MANGO _L</td>
<td>45.27</td>
<td>68.33</td>
<td>59.45</td>
<td>60.50</td>
<td>62.14</td>
<td>61.10</td>
<td>58.83</td>
<td>62.60</td>
<td>58.83</td>
<td>62.00</td>
<td>58.29</td>
<td>74.26</td>
</tr>
<tr>
<td>VILLA _L</td>
<td>44.33</td>
<td>68.16</td>
<td>58.66</td>
<td>58.29</td>
<td>62.00</td>
<td>61.38</td>
<td>59.45</td>
<td>62.60</td>
<td>58.83</td>
<td>62.00</td>
<td>58.29</td>
<td>74.26</td>
</tr>
<tr>
<td>MANGO _VL</td>
<td>45.31</td>
<td>68.27</td>
<td>58.66</td>
<td>58.29</td>
<td>62.00</td>
<td>61.38</td>
<td>59.45</td>
<td>62.60</td>
<td>58.83</td>
<td>62.00</td>
<td>58.29</td>
<td>74.26</td>
</tr>
</tbody>
</table>

- Comparison with SOTA, MANGO pushes state-of-the-art performance by a large margin on 7 out of 9 benchmarks.
- On VQA-CP v2 and GQA, the SOTA methods exploit additional task-specific information (for example, scene graphs).

**MANGO** (Li et al. 2020)
What about Adversarial Robustness?

• MANGO => Adversarial VQA

Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models

Linjie Li¹, Jie Lei², Zhe Gan¹, Jingjing Liu¹
¹Microsoft  ²UNC Chapel Hill
{lindesy.li, zhe.gan, jingj}@microsoft.com
jielei@cs.unc.edu
Adversarial VQA

Sears (Ribeiro et al. 2018)
Q: What surrounds the sign?
Pred: Grass

Q: What about the sign?
Pred: Nothing

TextFooler (Jin et al. 2020)
Q: What kind of cars are featured in the picture?
Pred: Trucks

Q: What kind of automobiles are featured in the picture?
Pred: Motocycles

Semem+PSO (Zang et al. 2020)
Q: How many windows are on the right side of the train?
Pred: 3

Q: How many skylights are on the right side of the train?
Pred: 0

Automatically generated adversarial questions are often incorrect.

AVQA (Li et al. 2021)
Adversarial VQA

Human Performance: 80.78
Adversarial VQA

- A *dynamically evolving adversarial* VQA benchmark with *human-and-model-in-the-loop*
# Adversarial VQA

SOTA VQA models fail within 2 tries on average

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image Source</th>
<th>#Img</th>
<th>IsCollected</th>
<th>#IQ Total/Verified</th>
<th>Model error rate Mean/Median</th>
<th>#Tries</th>
<th>Time (sec.) per verified ex.</th>
<th>Data Split Train/Dev/Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>CC</td>
<td>13.7K</td>
<td>✓</td>
<td>93.1K/45.6K</td>
<td>48.9%/35.2%</td>
<td>1.6/1</td>
<td>71.0</td>
<td>53.6K/3.3K/10.0K</td>
</tr>
<tr>
<td>R2</td>
<td>CC</td>
<td>13.1K</td>
<td>✓</td>
<td>70.4K/37.8K</td>
<td>56.1%/49.0%</td>
<td>1.5/1</td>
<td>54.2</td>
<td>42.8K/2.7K/8.3K</td>
</tr>
<tr>
<td>R3</td>
<td>Various</td>
<td>11.1K</td>
<td>✓</td>
<td>60.4K/30.0K</td>
<td>49.8%/36.3%</td>
<td>1.6/1</td>
<td>57.3</td>
<td>35.3K/2.2K/6.6K</td>
</tr>
<tr>
<td>AVQA</td>
<td>Various</td>
<td>37.9K</td>
<td>✓</td>
<td>223.9K/113.4K</td>
<td>50.7%/40.0%</td>
<td>1.6/1</td>
<td>61.6</td>
<td>131.7K/8.2K/24.9K</td>
</tr>
</tbody>
</table>
## Adversarial VQA

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Data</th>
<th>R1 dev/test</th>
<th>R2 dev/test</th>
<th>R3 dev/test</th>
<th>AVQA dev/test</th>
<th>VQA v2 test-dev</th>
<th>Δ(v2, AVQA) test-dev, test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUTD</td>
<td>VQA v2 +VGQA</td>
<td>20.80/19.28</td>
<td>18.77/18.85</td>
<td>21.17/21.31</td>
<td>20.18/19.60</td>
<td>67.60</td>
<td>48.00</td>
</tr>
<tr>
<td></td>
<td>ALL</td>
<td>24.96/22.11</td>
<td>22.62/22.78</td>
<td>24.03/23.70</td>
<td>23.94/22.71</td>
<td>67.52</td>
<td><strong>44.81</strong></td>
</tr>
<tr>
<td>UNITER-B</td>
<td>VQA v2 +VGQA</td>
<td>20.60/17.91</td>
<td>17.86/18.55</td>
<td>19.45/20.20</td>
<td>19.39/18.66</td>
<td>72.70</td>
<td>54.04</td>
</tr>
<tr>
<td></td>
<td>+R1</td>
<td>26.03/22.94</td>
<td>17.30/17.36</td>
<td>19.41/20.23</td>
<td>21.48/20.37</td>
<td>72.98</td>
<td>52.61</td>
</tr>
<tr>
<td></td>
<td>+R1+R2</td>
<td>26.60/24.76</td>
<td>23.21/23.86</td>
<td>18.60/19.09</td>
<td>23.58/23.14</td>
<td>72.75</td>
<td>49.61</td>
</tr>
<tr>
<td></td>
<td>ALL</td>
<td>26.85/24.93</td>
<td>23.38/23.92</td>
<td>23.76/23.02</td>
<td>24.94/24.14</td>
<td>72.66</td>
<td>48.52</td>
</tr>
<tr>
<td>UNITER-L</td>
<td>VQA v2 +VGQA</td>
<td>25.04/23.72</td>
<td>17.82/17.49</td>
<td>18.86/19.34</td>
<td>21.11/20.54</td>
<td>73.82</td>
<td>53.28</td>
</tr>
<tr>
<td></td>
<td>+R1</td>
<td>29.31/26.63</td>
<td>19.34/18.66</td>
<td>19.53/18.30</td>
<td>23.60/21.93</td>
<td>73.89</td>
<td>51.96</td>
</tr>
<tr>
<td></td>
<td>+R1+R2</td>
<td>30.13/28.15</td>
<td>23.11/23.54</td>
<td>16.09/16.47</td>
<td>24.46/23.84</td>
<td>73.77</td>
<td>49.93</td>
</tr>
<tr>
<td></td>
<td>ALL</td>
<td><strong>30.80/28.45</strong></td>
<td>22.95/23.11</td>
<td>23.75/21.88</td>
<td><strong>26.11/25.07</strong></td>
<td><strong>74.15</strong></td>
<td><strong>49.08</strong></td>
</tr>
<tr>
<td>LXMERT</td>
<td>VQA v2 +VGQA</td>
<td>19.76/18.15</td>
<td>18.98/18.79</td>
<td>20.75/20.26</td>
<td>19.72/18.86</td>
<td>72.31</td>
<td>53.45</td>
</tr>
<tr>
<td></td>
<td>+R1</td>
<td>23.89/22.65</td>
<td>19.01/17.91</td>
<td>21.47/20.85</td>
<td>21.64/20.58</td>
<td>72.51</td>
<td>51.93</td>
</tr>
<tr>
<td></td>
<td>+R1+R2</td>
<td>26.76/24.86</td>
<td>23.28/24.11</td>
<td>19.16/18.93</td>
<td>23.80/23.23</td>
<td>72.61</td>
<td>49.38</td>
</tr>
<tr>
<td></td>
<td>ALL</td>
<td>26.35/24.55</td>
<td><strong>23.84/24.02</strong></td>
<td>24.00/22.90</td>
<td>24.94/23.98</td>
<td>72.42</td>
<td>48.44</td>
</tr>
</tbody>
</table>
AVQA includes diverse question types.
# Adversarial VQA

<table>
<thead>
<tr>
<th>Round</th>
<th>Count</th>
<th>OCR</th>
<th>Reasoning</th>
<th>Visual Concept Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Position</td>
<td>Low-level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Relation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commonsense</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>23.3%</td>
<td>10.7%</td>
<td>14.7%</td>
<td>8.3%</td>
</tr>
<tr>
<td>R2</td>
<td>30.0%</td>
<td>22.7%</td>
<td>12.0%</td>
<td>27.7%</td>
</tr>
<tr>
<td>R3</td>
<td>35.3%</td>
<td>13.0%</td>
<td>13.0%</td>
<td>28.3%</td>
</tr>
<tr>
<td>Ave.</td>
<td>29.6%</td>
<td>15.4%</td>
<td>13.2%</td>
<td>21.4%</td>
</tr>
</tbody>
</table>

**Q1:** What creature is in the box?  
**A1:** Bird, Conf: 53.2%

**Q2:** What is the box made of?  
**A1:** Wood, Conf: 99.8%
**A2:** Cardboard (0.9), Paper (0.3)

**Q1:** How many football players can you see off the ground?  
**A1:** 0, Conf: 92.2%
**A1:** 1 (1.0)

**Q1:** Is the person next to the dog standing or squatting down?  
**A1:** walking, Conf: 15.4%
**A1:** squatting down (1.0)
What about Fairness?

Worst of Both Worlds:
Biases Compound in Pre-trained Vision-and-Language Models

Tejas Srinivasan  
University of Southern California  
tejas.srinivasan@usc.edu

Yonatan Bisk  
Carnegie Mellon University  
ybisk@cs.cmu.edu
Agenda

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• Model Compression
• Robustness/fairness of VLP models
• Multilingual VLP
Multilingual VLP

- Visual Question Answering
  What is the animal in the picture? dog

- Multi-modal Verification
  The woman is sitting next to a cat. false

- Caption-based Image Retrieval
  A woman sitting on a bench next to a dog.
Multilingual VLP

M3P (Ni et al. 2021)

UC2 (Zhou et al. 2021)
M3P: Multitask Multilingual Multimodal Pre-training

M3P (Ni et al. 2021)
M3P: Multitask Multilingual Multimodal Pre-training

M3P (Ni et al. 2021)
M3P: Multitask Multilingual Multimodal Pre-training

Conceptual Caption (CC)

Multimodal Monolingual Stream

Multimodal Code-switched Stream

Monomodal Multilingual Stream

Data Stream

Input Representations

Model

Pre-training Tasks

Multilingual Multimodal Multitask Pre-trained Model

y: 0/1

dog

MC-VLM

MC-MRM

MC-MLM

xMLM

M3P (Ni et al. 2021)
M3P: Multitask Multilingual Multimodal Pre-training

M3P (Ni et al. 2021)
UC2: Universal Cross-lingual Cross-modal Pre-training

UC2 (Zhou et al. 2021)
UC2: Masked Region-to-Token Modeling

English: a woman sitting on a bench next to a dog
Chinese: 一个女人坐在一条大狗旁边的长椅上
Japanese: 犬の隣のベンチに座っている女性
French: une femme assise sur un banc à côté d'un gros chien

Credit to: Mingyang Zhou

UC2 (Zhou et al. 2021)
UC2: Visual Translation Language Modeling

Credit to: Mingyang Zhou

UC2 (Zhou et al. 2021)
## UC2: Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>EN</th>
<th>DE</th>
<th>FR</th>
<th>CS</th>
<th>EN</th>
<th>ZH</th>
<th>JA</th>
<th>Meta-Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALR</td>
<td>74.5</td>
<td>69.8</td>
<td>65.9</td>
<td>64.8</td>
<td>81.5</td>
<td>77.5</td>
<td>76.7</td>
<td>73.0</td>
</tr>
<tr>
<td>M³P</td>
<td>87.7</td>
<td>82.7</td>
<td>73.9</td>
<td>72.2</td>
<td>88.7</td>
<td>86.2</td>
<td>87.9</td>
<td>82.8</td>
</tr>
<tr>
<td>UNITER</td>
<td>87.7</td>
<td>81.2</td>
<td>81.9</td>
<td>80.2</td>
<td>88.4</td>
<td>87.3</td>
<td>82.2</td>
<td>84.1</td>
</tr>
<tr>
<td>UC²</td>
<td><strong>88.2</strong></td>
<td><strong>84.5</strong></td>
<td><strong>83.9</strong></td>
<td><strong>81.2</strong></td>
<td><strong>88.1</strong></td>
<td><strong>89.8</strong></td>
<td><strong>87.5</strong></td>
<td><strong>86.2</strong></td>
</tr>
</tbody>
</table>

English: a woman sitting on a bench next to a dog

Chinese: 一个女人坐在一条大狗旁边的长椅上

Japanese: 犬の隣のベンチに座っている女性

French: une femme assise sur un banc à côté d'un gros chien

Credit to: Mingyang Zhou

UC² (Zhou et al. 2021)
Conclusion

• Advanced training strategies in VLP
• Diverse applications of VLP
• VL for V/L
• Compressing VLP models
• Robustness/fairness of VLP models
• Multilingual VLP
Challenges and Future Directions

• **Fairness:**
  - We observe that there are severe biases in VLP models. How can we improve the fairness of VLP models?

• **Adversarial robustness:**
  - We observe that VLP models can be easily attacked. How can we enhance the adversarial robustness of VLP models?

• **Training efficiency:**
  - How can we obtain training efficiency rather than inference/parameter efficiency? This could be especially useful for pre-training.
Thank you!
CVPR 2021 Tutorial
Multi-task training

<table>
<thead>
<tr>
<th>Model</th>
<th># models</th>
<th># parameters</th>
<th>Average Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independently train for each task</td>
<td>12</td>
<td>12X270M = 3B</td>
<td>67.24</td>
</tr>
<tr>
<td>Train all tasks together (12-in-1)</td>
<td>1</td>
<td>270M</td>
<td>69.08</td>
</tr>
<tr>
<td>Further fine tune on each task</td>
<td>12</td>
<td>12X270M = 3B</td>
<td><strong>70.24 (+3)</strong></td>
</tr>
</tbody>
</table>

12-in-1 (Lu et al. 2020)
UC2: Data Augmentation via Machine Translation

UC2 (Zhou et al. 2021)
**Training Strategies**

- **E2E Pre-training**
  - ViLT (Kim et al. 2021)
  - SOHO (Huang et al. 2021)
- **VLP for V/L**
  - ViL-BERT (Cho et al. 2021)
  - LightningDOT (Sun et al. 2021)
  - Unsupervised VLP (Li et al. 2021)
- **Model Compression**
  - Pixel-BERT (Huang et al. 2020)
  - OS2 (Shi et al. 2020)
  - VILLA (Gan et al. 2020)
  - VL-BART (Cho et al. 2021)

**Diverse Applications**

- **Diverse Applications**
  - Kaleido-BERT (Zhuge et al. 2021)
  - VisDial-BERT (Murahari et al. 2021)
  - VLP w/Knowledge (Shevchenko et al. 2021)
  - VD-BERT (Wang et al. 2020)
  - PREVALENT (Sariyildiz et al. 2020)
  - VIVO (Hao et al. 2020)
  - TAP (Yang et al. 2020)
  - MINILM (Wang et al. 2020)
  - CLIP (Radford et al. 2021)
  - VirTex (Desai et al. 2021)
  - DistillVLM (Fang et al. 2021)

**Enhanced Visual Representations**

- **Enhanced Visual Representations**
  - OSCAR (Li et al. 2020)
  - VinVL (Zhang et al. 2021)
  - Behind the Scene (Cao et al. 2020)

**Probing Analysis**

- **Probing Analysis**
  - What does BERT with vision look at? (Li et al. 2020)

**Pioneering work in VLP**

- **Pioneering work in VLP**
  - LXMERT (Tan and Bansal 2019)
  - UniCoder-VL (Li et al. 2020)
  - UniPER (Chen et al. 2020)
  - VLP (Zhou et al. 2020)
  - VLP-BERT (Su et al. 2020)
  - ViL-BERT (Lu et al. 2019)
  - UNITER (Chen et al. 2020)
  - VisuBERT (Li et al. 2019)

**Robustness/Causality/Fairness**

- **Robustness/Causality/Fairness**
  - MANGO (Li et al. 2020)
  - CATT (Yang et al. 2021)
  - AVQA (Li et al. 2021)

- **Worst of Both Worlds**
  - VILLA (Gan et al. 2020)