MULTIMODAL MACHINE LEARNING

AISHWARYA KAMATH
AGENDA

- **Background**
  - a. Tasks & datasets
  - b. Influential early approaches
  - c. Large scale pre-training and shortcomings

- **MDETR**
  - a. Modulated Detection
  - b. Architecture
  - c. Loss functions
  - d. Results
Common tasks and datasets for vision+text understanding
Task 1: Expression Generation
Generate referring expression for this target person.

Algorithm: The girl playing wii

Task 2: Expression Comprehension
Which object is “Girl on the left” indicating?

walking people  wipers on trains  zebra lying on savanna

Image and Trace:

In the front portion of the picture we can see a dried grass area with dried twigs. There is a woman standing wearing light blue jeans and a red sleeve long sleeve length shirt. This woman is holding a black jacket in her hand. On the other hand she is holding a balloon which is peach in colour. On the top of the picture we see a clear blue sky with clouds. The hair colour of the woman is brownish.

Image

(CxC score) Ranked Captions
(4.48) Home plate at a professional baseball game, batter not quite ready.
(4.46) three players on the base ball diamond, all headed for a base.
(4.15) Baseball team mates and another player on the diamond.
(4.98) A batter, catcher and umpire in a baseball game.
(4.95) A batter, catcher and umpire in a baseball game.

(4.92) A dog wearing a striped elf hat sits in the snow.
(5.0) A dog is wearing an elf hat in the snow.
(5.0) A dog wearing an elf hat sits in the snow.
(4.25) Brown and white dog in Christmas hat standing in the snow.
(4.98) A dog that is wearing a christmas hat on its head.

Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metallic thing that is left of the big sphere?
Q: There is a sphere with the same size as the metal cube. Is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or red things?

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

How many slices of pizza are there?
Is this a vegetarian pizza?

A1. Are the napkin and the cup the same color? yes
A2. Are the napkin and the cup the same color? yes
A3. Is the small table both oval and wooden? yes
A4. Are there any fruit to the left of the tray the cup is on top of? yes
A5. Are there any cups to the left of the tray on top of the table? no
B1. What is the brown animal sitting inside of? box
B2. What is the large container made of? cardboard
B3. What animal is in the box? bear
B4. Is there a bag to the right of the green door? no
B5. Is there a box inside the plastic bag? no
Some influential early approaches
Show, Attend and Tell

Neural Image Caption Generation with Visual Attention

Main Idea:
Use a RNN with attention to the visual features to generate captions.
Neural Image Caption Generation with Visual Attention

Bottom-Up and Top-Down (BUTD) Attention for Image Captioning and Visual Question Answering

Won VQA Challenge 2017

Main idea:

**Attention over objects** instead of grid features

★ Serves as the image feature extractor for most vision+language models in years following.
Bottom-Up and Top-Down Attention

Instead of performing attention over a regular grid, attend to object regions

Bottom-Up and Top-Down Attention

Train on Visual Genome with:
- 1600 filtered object classes
- 400 filtered attribute classes

Main Idea:
Use different attention modules for object identity, location and relation to others.
**MAttNet**

Inferring and Executing Programs for Visual Reasoning

Neural module networks for compositional learning

Main Idea: Model predicts explicit program that represents the reasoning process and uses this in the execution engine to produce an answer.
Seq2Seq program generator + Neural Module Network executor

Feature Modulation

FiLM, MoVie (used in winning VQA Challenge 2020)

Main Idea:

Use features from the text to **manipulate** the visual stream (using affine transformations).
Feature Modulation - FiLM

Feature Modulation - MoVie

Transformers for vision+text understanding
Two main types

1. Cross encoder models
2. Dual encoder models

Main idea: Extract features from images and text, feed it through transformer layers.

Pre-training on massive datasets using cross-modal alignment tasks.
LXMERT/ViLBERT
Dual encoder + Cross attention

Main Idea:
Use separate vision encoder and text encoder to encode vision and text followed by cross attention between the two.
LXMERT pre-training tasks
Main Idea:
Use a single cross-encoder to encode text and vision.
Main Idea:
Use a cross-encoder to encode text and vision, while using object tags as anchors.
OSCAR

Performance bottlenecked by object detection
Should we go brute-force?

- Recent paper pre-train the detector on all available detection datasets
- Impressive performance on all downstream tasks
- **5.6 Million** Images
- Still bounded by 1848 object categories and 524 attribute categories,

Main idea: Massively pre-train dual encoders and train with a contrastive loss.
CLIP training

1. Contrastive pre-training
   - Text Encoder
   - Image Encoder
   - Input images and text

2. Create dataset classifier from label text
   - Text Encoder
   - Labels (plane, car, dog, bird)
   - A photo of a {object}, A photo of a dog

3. Use for zero-shot prediction
   - Image Encoder
   - Output predictions

Important takeaway: Generalization from natural language supervision + contrastive loss!
MDETR: Modulated Detection for End to End Multimodal Understanding
MDETR
Modulated Detection for End to End Multimodal Understanding

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, Nicolas Carion

Main idea: Only detect objects that are relevant.

Everything is based on finding the alignment between words in the free-form text, and objects in the image.

No longer bottlenecked by pre-trained object detectors!
What is “modulated detection”?

- **Free-form text conditioned detection**
- Output of MDETR for the query “A pink elephant”.

![Elephants](image)
Generic detection vs modulated detection

Text prompt: “blond boy wearing blue shorts. a black surf-board”
Phrase grounding is central to all VL tasks.

How can you answer questions (VQA), describe the image (captioning) or predict entailment (V-NLI) without knowing the relevant parts of the image being asked about?
Architecture

- Pre-requisites
  - DETR: Detection Transformers

- MDETR Components
  - Backbone
  - Text encoder
  - Cross encoder
  - Decoder
DETR - Detection transformer

- End-to-end detection
- Encoder-decoder architecture
Looking inside...
MDETR: Architecture

“A cat with white paws jumps over a fence in front of a yellow tree”

No more pre-defined “class labels”
MDETR: Architecture
Architecture modification for visual question answering

“What is the color of the \textcolor{blue}{sphere} behind the \textcolor{green}{green cylinder}?”
Loss functions

- Soft token prediction
- Contrastive alignment
Losses: Soft token prediction

A cat with white paws jumps over a fence in front of a yellow tree Ø
Losses: Contrastive alignment

- Align embedding of a visual **object** after the decoder to the contextualized representation of the text **token** at the output of the cross-encoder.
- InfoNCE-style

```
    t1  t2  t3
  o1  x   
  o2   x  
  o3  x   x
  o4   
```

“Ball or yellow”
Loss function ablations

<table>
<thead>
<tr>
<th>Model</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported architecture</td>
<td>99.0</td>
</tr>
<tr>
<td>- Contrastive loss</td>
<td>83.2</td>
</tr>
<tr>
<td>- Soft token prediction</td>
<td>87.7</td>
</tr>
</tbody>
</table>
Results

- Synthetic data - CLEVR
- Natural images - Flickr, COCO, Visual Genome
Query: “Any other things that are the same color as the partially visible thing(s)”
## Results on CLEVR and related

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Count</th>
<th>Exist</th>
<th>CLEVR</th>
<th>Comp. Num</th>
<th>Query</th>
<th>Comp. Att</th>
<th>CLEVR-Humans</th>
<th>CoGenT</th>
<th>CLEVR-Ref+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Before FT</td>
<td>After FT</td>
<td>TestA</td>
<td>TestB</td>
<td>Acc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAttNet[70]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>56.6</td>
<td>75.9</td>
<td>98.3</td>
</tr>
<tr>
<td>MGA-Net[74]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FiLM[42]</td>
<td>97.7</td>
<td>94.3</td>
<td>99.1</td>
<td>96.8</td>
<td>99.1</td>
<td>99.1</td>
<td>-</td>
<td>98.3</td>
<td>78.8</td>
<td>-</td>
</tr>
<tr>
<td>MAC [17]</td>
<td>98.9</td>
<td>97.1</td>
<td>99.5</td>
<td>99.1</td>
<td>99.5</td>
<td>99.5</td>
<td>-</td>
<td>98.4</td>
<td>99.8</td>
<td>63.9</td>
</tr>
<tr>
<td>NS-VQA[68]*</td>
<td>99.8</td>
<td>99.7</td>
<td>99.9</td>
<td>99.8</td>
<td>99.8</td>
<td>99.8</td>
<td>67.8</td>
<td>99.8</td>
<td>63.9</td>
<td>-</td>
</tr>
<tr>
<td>OCCAM [60]</td>
<td>99.4</td>
<td>98.1</td>
<td>99.8</td>
<td>99.0</td>
<td>99.9</td>
<td>99.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MDETR</td>
<td>99.7</td>
<td>99.3</td>
<td>99.9</td>
<td>99.4</td>
<td>99.9</td>
<td>99.9</td>
<td>-</td>
<td>59.9</td>
<td>81.7</td>
<td>99.8</td>
</tr>
</tbody>
</table>
Combining Ref Exp style & Flickr style data

(c) “the man in the red shirt carrying baseball bats”

(a) “one small boy climbing a pole with the help of another boy on the ground”
MDETR: Pretraining

- Images from Flickr30k, COCO, Visual Genome
- Combine training examples across different datasets for the same image.
- => 1.3m aligned image-text pairs
- 40 epochs

“the woman in the grey shirt with a watch on her wrist. the older woman wearing a blue sweater. the other woman in a gray coat and scarf.”
Phrase grounding on Flickr30k - Qualitative results

(a) “one small boy climbing a pole with the help of another boy on the ground”

(b) “A man talking on his cellphone next to a jewelry store”

(c) “A man in a white t-shirt does a trick with a bronze colored yo-yo”
### Phrase grounding on Flickr30k - Quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>Val</th>
<th>Test</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@5</td>
<td>R@10</td>
<td>R@10</td>
</tr>
<tr>
<td><strong>ANY-BOX-PROTOCOL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAN [21]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>69.7</td>
<td>84.2</td>
</tr>
<tr>
<td>VisualBert[25]</td>
<td>68.1</td>
<td>84.0</td>
<td>86.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VisualBert†[25]</td>
<td>70.4</td>
<td>84.5</td>
<td>86.3</td>
<td>71.3</td>
<td>85.0</td>
<td>86.5</td>
</tr>
<tr>
<td>MDETR-R101</td>
<td>78.9</td>
<td>88.8</td>
<td>90.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MDETR-R101†*</td>
<td><strong>82.5</strong></td>
<td><strong>92.9</strong></td>
<td><strong>94.9</strong></td>
<td><strong>83.4</strong></td>
<td><strong>93.5</strong></td>
<td><strong>95.3</strong></td>
</tr>
<tr>
<td>MDETR-ENB3†*</td>
<td><strong>82.9</strong></td>
<td><strong>93.2</strong></td>
<td><strong>95.2</strong></td>
<td><strong>84.0</strong></td>
<td><strong>93.8</strong></td>
<td><strong>95.6</strong></td>
</tr>
<tr>
<td>MDETR-ENB5†*</td>
<td><strong>83.6</strong></td>
<td><strong>93.4</strong></td>
<td><strong>95.1</strong></td>
<td><strong>84.3</strong></td>
<td><strong>93.9</strong></td>
<td><strong>95.8</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>MERGED-BOXES-PROTOCOL</strong></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CITE [43]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>61.9</td>
<td>-</td>
</tr>
<tr>
<td>FAOG [66]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>68.7</td>
<td>-</td>
</tr>
<tr>
<td>SimNet-CCA [45]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>71.9</td>
<td>-</td>
</tr>
<tr>
<td>MDETR-R101†*</td>
<td><strong>82.4</strong></td>
<td><strong>92.6</strong></td>
<td><strong>94.5</strong></td>
<td><strong>83.3</strong></td>
<td><strong>92.1</strong></td>
<td><strong>93.8</strong></td>
</tr>
</tbody>
</table>
Referring expressions

(a) “brown bear”

(b) “zebra facing away”

(c) “the man in the red shirt carrying baseball bats”

RefCOCO  RefCOCO+  RefCOCOg
## Results for referring expressions on RefCOCO

<table>
<thead>
<tr>
<th>Method</th>
<th>Detection backbone</th>
<th>Pre-training image data</th>
<th>RefCOCO val</th>
<th>RefCOCO testA</th>
<th>RefCOCO testB</th>
<th>RefCOCO+ val</th>
<th>RefCOCO+ testA</th>
<th>RefCOCO+ testB</th>
<th>RefCOCOg val</th>
<th>RefCOCOg test</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAttNet[69]</td>
<td>R101</td>
<td>None</td>
<td>76.65</td>
<td>81.14</td>
<td>69.99</td>
<td>65.33</td>
<td>71.62</td>
<td>56.02</td>
<td>66.58</td>
<td>67.27</td>
</tr>
<tr>
<td>ViLBERT[34]</td>
<td>R101</td>
<td>CC (3.3M)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>72.34</td>
<td>78.52</td>
<td>62.61</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VL-BERT_L[54]</td>
<td>R101</td>
<td>CC (3.3M)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>72.59</td>
<td>78.57</td>
<td>62.30</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UNITER_L[6]*</td>
<td>R101</td>
<td>CC, SBU, COCO, VG (4.6M)</td>
<td>81.41</td>
<td>87.04</td>
<td>74.17</td>
<td>75.90</td>
<td>81.45</td>
<td>66.70</td>
<td>74.86</td>
<td>75.77</td>
</tr>
<tr>
<td>VILLA_L[9]*</td>
<td>R101</td>
<td>CC, SBU, COCO, VG (4.6M)</td>
<td>82.39</td>
<td>87.48</td>
<td>74.84</td>
<td>76.17</td>
<td>81.54</td>
<td>66.84</td>
<td>76.18</td>
<td>76.71</td>
</tr>
<tr>
<td>ERNIE-VIL_L[68]</td>
<td>R101</td>
<td>CC, SBU (4.3M)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>75.95</td>
<td>82.07</td>
<td>66.88</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MDETR</td>
<td>R101</td>
<td>COCO, VG, Flickr30k (200k)</td>
<td><strong>86.75</strong></td>
<td><strong>89.64</strong></td>
<td><strong>81.47</strong></td>
<td><strong>79.52</strong></td>
<td><strong>84.72</strong></td>
<td><strong>69.76</strong></td>
<td><strong>81.64</strong></td>
<td><strong>80.98</strong></td>
</tr>
<tr>
<td>MDETR</td>
<td>ENB3</td>
<td>COCO, VG, Flickr30k (200k)</td>
<td><strong>87.51</strong></td>
<td><strong>90.38</strong></td>
<td><strong>82.90</strong></td>
<td><strong>81.13</strong></td>
<td><strong>85.52</strong></td>
<td><strong>72.96</strong></td>
<td><strong>83.35</strong></td>
<td><strong>83.45</strong></td>
</tr>
</tbody>
</table>
Results for segmentation on PhraseCut

(a) Query: “street lamp”
(b) Query: “major league logo”
(c) Query: “zebras on savanna”

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>PhraseCut M-IoU</th>
<th>Pr@0.5</th>
<th>Pr@0.7</th>
<th>Pr@0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMI[4]</td>
<td>R101</td>
<td>21.1</td>
<td>22.0</td>
<td>11.6</td>
<td>1.5</td>
</tr>
<tr>
<td>HULANet[63]</td>
<td>R101</td>
<td>41.3</td>
<td>42.4</td>
<td>27.0</td>
<td>5.7</td>
</tr>
<tr>
<td>MDETR</td>
<td>R101</td>
<td><strong>53.1</strong></td>
<td><strong>56.1</strong></td>
<td><strong>38.9</strong></td>
<td><strong>11.9</strong></td>
</tr>
<tr>
<td>MDETR</td>
<td>ENB3</td>
<td><strong>53.7</strong></td>
<td><strong>57.5</strong></td>
<td><strong>39.9</strong></td>
<td><strong>11.9</strong></td>
</tr>
</tbody>
</table>
MDETR: Architecture (GQA)

“Is the **bookcase** to the left of the **nightstand** in the photograph?”
Question answering: results on GQA

- Additional object queries specialized for question types answer, + type of question in REL, OBJ, GLOBAL, CAT, ATTR.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training img data</th>
<th>Test-dev</th>
<th>Test-std</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoVie [39]</td>
<td>-</td>
<td>-</td>
<td>57.10</td>
</tr>
<tr>
<td>LXMERT[55]</td>
<td>VG, COCO (180k)</td>
<td>60.0</td>
<td>60.33</td>
</tr>
<tr>
<td>VL-T5 [7]</td>
<td>VG, COCO (180k)</td>
<td>-</td>
<td>60.80</td>
</tr>
<tr>
<td>MMN [5]</td>
<td>-</td>
<td>-</td>
<td>60.83</td>
</tr>
<tr>
<td>OSCAR [27]</td>
<td>VG, COCO, Flickr, SBU (4.3M)</td>
<td>61.58</td>
<td>61.62</td>
</tr>
<tr>
<td>MDETR-R101</td>
<td>VG, COCO, Flickr30k (200k)</td>
<td>62.48</td>
<td>61.99</td>
</tr>
<tr>
<td>MDETR-ENB5</td>
<td>VG, COCO, Flickr30k (200k)</td>
<td>62.95</td>
<td>62.45</td>
</tr>
<tr>
<td>NSM [18]</td>
<td>-</td>
<td>-</td>
<td>63.17</td>
</tr>
<tr>
<td>VinVL [71]</td>
<td>VG, COCO, Objects365, SBU, Flickr30k, CC, VQA, OpenImagesV5 (5.65M)</td>
<td>65.05</td>
<td>64.65</td>
</tr>
</tbody>
</table>
Interpretable predictions

Given this image and the question:

“What is on the table?”

Predicted answer: “laptop”
Another example

Query: “What color is the train?”

Predicted answer: “red”
Some additional examples

"A green umbrella. A pink striped umbrella. A plain white umbrella"

"A flowery top. A blue dress. An orange shirt"

"A car. An electricity box"
Limits to zero-shot detection

- Training data has no “negative examples” - i.e. when the text does not correspond to any object in the image
- Model will always try to find something (usually salient objects in the image)
Results for detection on LVIS

- Performs well with as low as 1 sample/class, performance drops with more annotated data probably due to class imbalance.
- Due to overlaps between COCO/LVIS/..., we report results on the subset of 5k validation images that our model has never seen during training.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>AP</th>
<th>AP50</th>
<th>AP_r</th>
<th>AP_c</th>
<th>AP_f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN</td>
<td>100%</td>
<td>33.3</td>
<td>51.1</td>
<td>26.3</td>
<td>34.0</td>
<td>33.9</td>
</tr>
<tr>
<td>DETR</td>
<td>1%</td>
<td>4.2</td>
<td>7.0</td>
<td>1.9</td>
<td>1.1</td>
<td>7.3</td>
</tr>
<tr>
<td>DETR</td>
<td>10%</td>
<td>13.7</td>
<td>21.7</td>
<td>4.1</td>
<td>13.2</td>
<td>15.9</td>
</tr>
<tr>
<td>DETR</td>
<td>100%</td>
<td>17.8</td>
<td>27.5</td>
<td>3.2</td>
<td>12.9</td>
<td>24.8</td>
</tr>
<tr>
<td>MDETR</td>
<td>1%</td>
<td>16.7</td>
<td>25.8</td>
<td>11.2</td>
<td>14.6</td>
<td>19.5</td>
</tr>
<tr>
<td>MDETR</td>
<td>10%</td>
<td>24.2</td>
<td>38.0</td>
<td>20.9</td>
<td>24.9</td>
<td>24.3</td>
</tr>
<tr>
<td>MDETR</td>
<td>100%</td>
<td>22.5</td>
<td>35.2</td>
<td>7.4</td>
<td>22.7</td>
<td>25.0</td>
</tr>
</tbody>
</table>
Conclusion
Key takeaways

- Remove dependence on pre-trained object detectors
- No longer restricted by fixed vocabulary of object classes (often 1600 classes, 400 attributes)
- Can detect anything referred to in free-form text
- Novel combinations of categories and attributes (pink elephant!)
- Interpretable predictions
Thank you!

Website: https://ashkamath.github.io/mdetr_page/
Colab: https://colab.research.google.com/github/ashkamath/mdetr/blob/colab/notebooks/MDETR_demo.ipynb
Code: https://github.com/ashkamath/mdetr
Email : aish@nyu.edu