Vision and Language
Context: Vision and Language

- **Classification**: Cat
- **Detection**: Cat, Skateboard
- **Captioning**: A cat riding a skateboard
- **Dense Captioning**: Orange spotted cat, Skateboard with red wheels, Cat riding a skateboard, Brown hardwood flooring

@Feifei
Vision and Language: Multi-domain alignment / Retrieval

Context: Vision and Language

Language description/complexity

Vision and Language: from keywords to sentence/caption ...
Vision and Language

1. Multimodal embedding
   • Deep nets to align image+text
   • Learning

2. VQA framework
   • Task modeling
   • Fusion in VQA
   • Reasoning in VQA
Deep semantic-visual embedding

<table>
<thead>
<tr>
<th>Images</th>
<th>Multimodal space</th>
<th>Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image of a cat wearing a suit]</td>
<td>![Multimodal space diagram]</td>
<td>A cat wearing a suit with a red tie and white shirt.</td>
</tr>
</tbody>
</table>

Tools and Models:
- ConvNet ViT
- RNN Transformer
Deep semantic-visual embedding

Image captioning

Multimodal space

Semantic of distance
Retrieval by NN search

Images

A cat wearing a suit with a red tie and white shirt.

Texts

Image generation
Deep semantic-visual embedding

2D Semantic visual space example:
- Distance in the space has a semantic interpretation
- Retrieval is done by finding nearest neighbors
Deep semantic-visual embedding

- Designing image and text embedding architectures
- **Learning** scheme for these deep hybrid nets
Deep semantic-visual embedding

Embedding with pairs: image+label


Dataset: ImagNet1k = 1M image/label
Deep semantic-visual embedding

Embedding with pairs: image+caption

[Finding beans in burgers, M. Engilberge et al, CVPR 2018]

Dataset: 1M image/caption or similar size

θ₀, 2 and φ are the trained parameters
Deep semantic-visual embedding

Embedding with pairs: image+caption

[Learning transferable visual models from natural language supervision, ICML 2021]

Dataset: **500M pairs from Internet = no manual labeling**
Deep semantic-visual embedding

Learning strategy
Training set: \( A = \{(I_n, S_n)\}_n \) of image/caption pairs

Triplet loss: A variant of the standard margin based loss (SVM)

- Triplet \((y, z, z')\)
- Anchor: \(y\) (E.g image representation)
- Positive: \(z\) (E.g associated caption representation)
- Negative: \(z'\) (E.g contrastive caption representation)
- Margin parameter \(\alpha\)

\[
\text{loss}(y, z, z') = \max\{0, \alpha - <y, z> + <y, z'>\}
\]
Learning strategy: triplet loss

\[
\text{loss}(y, z, z') = \max \{ 0, \alpha + d(y, z) - d(y, z') \}
\]
Learning strategy: triplet loss

**Hard negative** margin-based loss:

*Loss for a batch* $\mathcal{B} = \{(I_n, S_n)\}_{n \in B}$ *of image/sentence pairs:*

$$
\mathcal{L}(\Theta; \mathcal{B}) = \frac{1}{|B|} \sum_{n \in B} \left( \max_{m \in C_n \cap B} \text{loss} (x_n, v_n, v_m) + \max_{m \in D_n \cap B} \text{loss} (v_n, x_n, x_m) \right)
$$

**With:**

- $C_n$ (resp. $D_n$) set of indices of caption (resp. image) unrelated to $n$-th element.
Learning strategy: hard negative triplet loss

Mining hard negative contrastive example:

\[
\mathcal{L}(\Theta; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \left( \max_{m \in C_n \cap \mathcal{B}} \text{loss} (x_n, v_n, v_m) + \max_{m \in D_n \cap \mathcal{B}} \text{loss} (v_n, x_n, x_m) \right)
\]
Learning strategy: hard negative triplet loss

Mining hard negative contrastive example:

\[
\mathcal{L}(\Theta; B) = \frac{1}{|B|} \sum_{n \in B} \left( \max_{m \in C_n \cap B} \text{loss} (x_n, v_n, v_m) + \max_{m \in D_n \cap B} \text{loss} (v_n, x_n, x_m) \right)
\]
CLIP learning: Contrastive loss over the whole batch

Vision and Language: Multi-domain alignment / Retrieval
## Cross-modal retrieval

<table>
<thead>
<tr>
<th>Query</th>
<th>Closest elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>A plane in a cloudy sky</td>
<td><img src="image1" alt="Image of a plane in a cloudy sky" /> <img src="image2" alt="Image of a plane in a cloudy sky" /> <img src="image3" alt="Image of a plane in a cloudy sky" /> <img src="image4" alt="Image of a plane in a cloudy sky" /> <img src="image5" alt="Image of a plane in a cloudy sky" /></td>
</tr>
<tr>
<td>A dog playing with a frisbee</td>
<td><img src="image1" alt="Image of a dog playing with a frisbee" /> <img src="image2" alt="Image of a dog playing with a frisbee" /> <img src="image3" alt="Image of a dog playing with a frisbee" /> <img src="image4" alt="Image of a dog playing with a frisbee" /> <img src="image5" alt="Image of a dog playing with a frisbee" /></td>
</tr>
</tbody>
</table>
| ![Image of a herd of sheep standing on top of snow covered field.](image1) | **1. A herd of sheep standing on top of snow covered field.**  
2. There are sheep standing in the grass near a fence.  
3. some black and white sheep a fence dirt and grass |
Cross-modal retrieval and localization

Visual grounding examples:

Generating multiple heat maps with different textual queries

Finding beans in burgers: Deep semantic-visual embedding with localization, M. Engilberge et al, CVPR 2018
Cross-modal retrieval and localization

Visual grounding examples
Zero-shot classification with CLIP

(2) Create dataset classifier from label text

(3) Use for zero-shot prediction
Vision and Language

1. Multimodal embedding
   - Deep nets to align text+image
   - learning

2. VQA framework
   - Task modeling
   - Fusion in VQA
   - Reasoning in VQA
Visual Question Answering

Question Answering:
What does Claudia do?
Visual Question Answering:

What does Claudia do?
Visual Question Answering

Visual Question Answering:
What does Claudia do?

Sitting at the bottom
Standing at the back
...
Visual Question Answering:

What does Claudia do?

+ Sitting at the bottom
  Standing at the back

Solving this task interesting for:
- Study of deep learning models in a multimodal context
- Improving human-machine interaction
- One step to build visual assistant for blind people
VQA: one step further for (visual) reasoning

VQA - Visual Question Answering

Images

Questions

What color is the cat's tie?

Multimodal space

Answers

red

COCOQA 15756
What does the man ride while wearing a black wet suit?
Ground truth: surfboard
IMG+BOW: jacket (0.35)
2-VIS+LSTM: surfboard (0.53)
BOW: tie (0.30)

DAQUAR 2136
What is right of table?
Ground truth: shelves
IMG+BOW: shelves (0.33)
2-VIS+BLSTM: shelves (0.28)
LSTM: shelves (0.20)

Does it appear to be rainy?
Does this person have 20/20 vision?

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

What color is the fire Hydrant on the left?

Green
VQA: one step further for (visual) reasoning

VQA

What color is the fire Hydrant on the right?

Yellow
Who is wearing glasses?

Different answers

Similar images

man \quad \text{\textcolor{green}{man}} \quad \text{\textcolor{green}{woman}}

@VQA workshop, CVPR 2017

⇒ Need very good Visual and Question (deep) representations
⇒ Full scene understanding
⇒ Need High level multimodal interaction modeling
⇒ Merging operators, attention and reasoning
Vanilla VQA scheme: 2 deep + fusion
Question: Is the lady with the blue fur wearing glasses?

VQA: the output space

VQA: one step further for (visual) reasoning

Yes
VQA: the output space

VQA Dataset [Antol et al. 2015]

- released for the VQA Challenge Workshop at CVPR 2016
- Each pair (image, question) is associated with 10 correct answers

Figure: Example of an (image,question,answers) triplet from VQA dataset
VQA: one step further for (visual) reasoning

VQA: the output space

Evaluation metric

\[ acc_{vqa}(answer) = \min \left( 1, \frac{\# \text{ humans that provided that answer}}{3} \right) \] (2)

Volumes:

- Train set: 82,783 images, 248,349 questions and answers
- Val set: 40,504 images, 121,512 questions and answers
- Test set: 81,434 images, 244,302 questions

Output space representation:

=> Classify over the most frequent answers (3000/95%)
VQA: the output space

**Question**: Is the lady with the blue fur wearing glasses?
VQA processing

Image
- Convolutional Network (VGG, ResNet,....)
- Detection system (EdgeBoxes, Faster-RCNN, ...)

Question
- Bag-of-words
- Recurrent Network (RNN, LSTM, GRU, SRU, ...)

Learning
- Fixed answer vocabulary
- Classification (cross-entropy)
Fusion in VQA
VQA: fusion

Concatenation & projection: \( y = \mathbf{W} \begin{bmatrix} q \\ v \end{bmatrix} \)

Element-wise sum: \( y = (\mathbf{W}q) + (\mathbf{V}v) \)

Element-wise product: \( y = (\mathbf{W}q) \odot (\mathbf{V}v) \)

Multi-layer perceptron: \( y = \text{MLP} \left( \begin{bmatrix} q \\ v \end{bmatrix} \right) \)
VQA: fusion

Is the lady with the purple fur wearing glasses?

GRU

Resnet

$q$

$v$

Fusion
Elementwise sum
Elementwise product
Concatenation + MLP

$y$ -> Softmax -> no

Concatenation & projection: $y = W \begin{bmatrix} q \\ v \end{bmatrix}$

Element-wise sum: $y = (Wq) + (Vv)$

Element-wise product: $y = (Wq) \odot (Vv)$

Multi-layer perceptron: $y = MLP \left( \begin{bmatrix} q \\ v \end{bmatrix} \right)$
VQA: bilinear fusion

[Kim, Jin-Hwa et al. Hadamard Product for Low-rank Bilinear Pooling, ICLR 2017]

Bilinear model:

score for class \( k \) = bilinear combination of dimensions in \( q \) and \( v \)

\[
y^k = \sum_{i=1}^{d_q} \sum_{j=1}^{d_v} T^{ijk} q^i v^j
\]

\[
y = \mathcal{T} \times_1 q \times_2 v
\]
VQA: bilinear fusion

\[ y^k = \sum_{i=1}^{d_q} \sum_{j=1}^{d_v} T^{ijk} q^i v^j \]

Learn the 3-ways tensor coeff.

• Different than the Signal Proc. Tensor analysis (representation)

Problem: \( q, v \) and \( y \) are of dimension \( \sim 2000 \)

=> 8 billion free parameters in the tensor

Need to reduce the tensor size:

• Idea: structure the tensor to reduce the number of parameters
Multiple ways of learning a merging function between two vector spaces

- Linear projections
- Deep fusions
- Bilinear models, simplified by:
  - sketching techniques,
  - tensor decompositions framework
- higher-order fusion
Reasoning in VQA

Q: Are there an equal number of large things and metal spheres?
VQA: reasoning

What is reasoning (for VQA)?

Attentional reasoning

Relational reasoning

Iterative reasoning

Compositional reasoning
VQA: reasoning

What is reasoning (for VQA)?

**Attentional reasoning:** given a certain context (i.e. Q), focus only on the relevant subparts of the image

**Relational reasoning**

**Iterative reasoning**

**Compositional reasoning**
VQA: attentional reasoning

Idea: focusing only on parts of the image relevant to Q

- Each region scored according to the question

What is sitting on the desk in front of the boys?

- Representation = sum of all (weighted) embeddings
What is sitting on the desk in front of the boys?
VQA: attentional reasoning

What is sitting on the desk in front of the boys?
VQA: attentional reasoning

Attention blocks are Transformers like (Cross attention)

Attentional glimpse in most of recent strategies [MLB, MCB, MUTAN…]
VQA: one step further for (visual) reasoning

VQA: attentional reasoning

What is sitting on the desk in front of the boys?

What are on the shelves in the background?

Tucker Decomposition with Structured Sparsity

Laptops

Books
Focusing on multiple regions: Multi-glimpse attention

Where is the smoke coming from?
VQA: attentional reasoning with Multi-glimpse attention

Multi Attention blocks are Multi-head Transformers like (Cross attention)

Focus on the train

Focus on the smoke
VQA: one step further for (visual) reasoning

VQA: attentional reasoning with Multi-glimpse attention

(a) Question: Where is the woman? - Answer: on the elephant

(b) Question: Where is the smoke coming from? - Answer: train
VQA: attentional reasoning

From fixed grid to adaptive grid selection

=> Additional region detection learning process
What is reasoning (for VQA)?

**Attentional reasoning**: given a certain context (i.e. Q), focus only on the relevant subparts of the image.

**Relational reasoning**: object detection + mutual relationships (spatial, semantic,...), merging both with Q

Iterative reasoning

Compositional reasoning
Determine the answer using relevant objects and relationships

Question: Are both men wearing ties?

Answer: No
VQA: reasoning

What is reasoning (for VQA)?

*Attentional reasoning*: given a certain context (i.e. Q), focus only on the relevant subparts of the image

*Relational reasoning*: object detection + mutual relationships (spatial, semantic,...), merging both with Q

*Iterative reasoning*: refining the attention step-by-step, each time extracting a different piece of information from the image

*Generalized in Compositional reasoning*
Iterative Reasoning

At least 3 elementary steps are required to answer the question

- Find bicycles
- Find the bicycle that has a basket
- Find what is in this basket

**Stacked attention**: iteratively refining visual attention and question representation

What are sitting in the basket on a bicycle?

Zichao Yang *et. al.*, *Stacked Attention Networks for Image Question Answering*, CVPR 2016
Multimodal Relational Reasoning for VQA

MUREL system:
• Vector representation for Attention process
• Spatial and semantic contexts to model relations between image regions
• Iterative process / Multistep reasoning

Cadene et al., MuRel: Multimodal Relational Reasoning for Visual Question Answering CVPR 2019
MuRel: Multimodal Relational Reasoning for VQA
MuRel: Multimodal Relational Reasoning for VQA
MuRel: Multimodal Relational Reasoning for VQA

VQA: one step further for (visual) reasoning