Multimodal Deep Learning
with an NLP focus

CS224n
What is multimodality?

**multimodal adjective**

- having or involving several modes, modalities, or maxima

- multimodal distributions
- multimodal therapy

In our case, focusing on NLP: text + one or more other *modality* (images, speech, audio, olfaction, others). We’ll mostly focus on images as the other modality.
Why does multimodality matter?

A range of very good reasons:

- **Faithfulness**: Human experience is multimodal
- **Practical**: The internet & many applications are multimodal
- **Data efficiency and availability**:
  - Efficiency: Multimodal data is rich and “high bandwidth” (compared to language; quoting LeCun, “an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts”), so better for learning?
  - Scaling: More data is better, and we’re running out of high quality text data.

Multimodality is one of the main frontiers of the new foundation model revolution.
Multimodal brains

McGurk effect (McGurk & MacDonald, 1976)

https://www.youtube.com/watch?v=2k8fHR9jKVM
Multimodal applications

Let’s say we’re dealing with two modalities – text, and images:

● Retrieval (image <-> text)
● Captioning (image -> text)
● Generation (text -> image)
● Visual question answering (image+text -> text)
● Multimodal classification (image+text -> label)
● Better understanding/generation (image+text -> label/text)
Multimodal is hot right now

.. and/but has been “the next big thing” for almost a decade!

Language Is Not All You Need: Aligning Perception with Language Models

Shaohan Huang,* Li Dong,* Wenhui Wang,* Yaru Hao,* Saksham Singhal,* Shuming Ma*  
Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi  
Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei†  
Microsoft
Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. Multimodal foundation models
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?
Cross-modal “Visual-Semantic Embeddings”

WSABI (Weston et al 2010), DeVise (Frome et al 2013), Cross-Modal Transfer (Socher et al 2013)

\[
loss(image, label) = \sum_{j \neq label} \max[0, \text{margin} - t_{label}M \bar{v}(image) + t_{j}M \bar{v}(image)]
\]

Frome et al. 2013

Socher et al. 2013
Multimodal distributional semantics (Bruni et al., 2014)

Algorithm:

- Obtain visual “word vector” via BOVW:
  - Identify keypoints and get their descriptors
  - Cluster these and map to counts
- Concatenate with textual word vector
- Apply SVD to “fuse” information

This approach was shown to lead to better word representations on human similarity judgment datasets.
Neural version (KB, 2014; Lazaridou et al., 2015)

Kiela & Bottou, 2014

Lazaridou et al., 2015

\[ \mathcal{L}_{\text{ling}}(w_t) = \text{maximize context prediction} \]

\[ \mathcal{L}_{\text{vision}}(w_t) = \text{maximize similarity} \]
Beyond words: Sentence level alignment

Grounded Compositional Semantics (Socher et al., 2013)
Visual-Semantic Embeddings (Kiros et al., 2014; Faghri et al., 2015)
Visual-Semantic Alignments (Karpathy & Li, 2015)
Grounded Sentence Representations (Kiela et al., 2016)

Hinge/margin-like loss as in WSABI/DeViSE.
Image to text: Captioning

Show and tell (Vinyals et al., 2015)

Show, attend and tell (Xu et al., 2016)
Attention as visual-semantic alignment

A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.
Text to image: Conditional image synthesis

Generative adversarial nets (Goodfellow et al. 2014)

As training progresses, the generator gets closer to producing output that can fool the discriminator.

Finally, if generator training goes well, the discriminator gets worse at telling the difference between real and fake. It starts to classify fake data as real, and its accuracy decreases.

Source: Google

Reed et al., 2016
Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. Multimodal foundation models
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?
Problems with multimodality

If it’s so important, why isn’t every system multimodal from first principles?

- One modality can dominate other modalities.
- Additional modalities can add noise.
- Full coverage over modalities is not guaranteed.
- We are (were) not ready.
- It’s complicated.
Features

Featurizing text: Batch\textsubscript{size} x Sequence\textsubscript{length} x Hidden\textsubscript{size}.

Featurizing images:

- **Sparse “region” features:**
  - Object detectors
- **Dense features:**
  - ConvNet layer(s) or feature maps
  - Vision Transformer layers

Anderson et al., 2018
Region features

- R-CNN (Girshick et al., 2014); Fast R-CNN (Girshick, 2015); Faster R-CNN (Ren et al., 2015); YOLO (you only look once) vX.
“Off the shelf” ConvNet features (Razavian et al., 2014)
Vision Transformers (Dosovitskiy et al., 2020)
Multimodal fusion

Similarity

- Inner product: $uv$

Linear / sum

- Concat: $W[u, v]$
- Sum: $Wu + Vv$
- Max: $\max(Wu, Vv)$

Multiplicative

- Multiplicative: $Wu \odot Vv$
- Gating: $\sigma(Wu) \odot Vv$
- LSTM-style: $\tanh(Wu) \odot Vv$

Attention

- Attention: $\alpha Wu + \beta Vv$
- Modulation: $[\alpha u, (1-\alpha)v]$

Bilinear

- Bilinear: $uWv$
- Bilinear gated: $uW\sigma(v)$
- Low-rank bilinear: $uU^TVv = P(Uu \odot Vv)$
- Compact bilinear: $\text{FFT}^{-1}(\text{FFT}(\Psi(x, h_1, s_1)) \odot \text{FFT}(\Psi(x, h_2, s_2)))$
Early middle and late

Suppose we have a binary classifier MLP and two input vectors.

Early - mix inputs:
  \[ \sigma(W_2 \sigma(W_1[u,v] + b_1) + b_2) \]

Middle - concatenate features:
  \[ \sigma(W_2 [\sigma(W_1[v] + b_1), \sigma(W'[v] + b'_1)] + b_2) \]

Late - combine final scores:
  \[ \frac{1}{2} (\sigma(W_2 \sigma(W_1[u] + b_1) + b_2) + \sigma(V_2 \sigma(V_1[u] + b'_1) + b'_2)) \]
FiLM (Perez et al., 2017)

Modulate one modality, layerwise, by the other.

\[ \gamma_{i,c} = f_c(x_i) \]

\[ \beta_{i,c} = h_c(x_i) \]

\[ \text{FiLM}(F_{i,c} \mid \beta_{i,c}, \gamma_{i,c}) = \gamma_{i,c} F_{i,c} + \beta_{i,c} \]
Outline

1. Early models
2. Features and fusion
3. **Contrastive models**
4. Multimodal foundation models
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?
CLIP (Radford et al. 2021)

Exact same contrastive loss as earlier, but.. Transformers and *web data*!

1. Contrastive pre-training

2. Create dataset classifier from label text

3. Use for zero-shot prediction
CLIP Robustness

IMHO one of the best papers ever written in our field: extremely thorough, worth a close read.

Generalizes MUCH better →
ALIGN (Jia et al., 2021)

Same idea, but EVEN MORE data (JFT at 1.8B image-text pairs vs CLIP’s 300m).
Aligned datasets

HUGE open source datasets of image-text pairs now exist.

Used to train eg StableDiffusion (Rombach et al., 2022).

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS
by: Romain Beaumont, 31 Mar, 2022

https://laion.ai/blog/laion-5b/
Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. **Multimodal foundation models**
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?
How do we make this multimodal? →

Randomly mask 15% of tokens

Input

Use the output of the masked word's position to predict the masked word

Possible classes:
- Aardvark (0.1%)
- Improvisation (10%)
- Zyzzyva (0%)

FFNN + Softmax

Alammar (2018), Illustrated Bert
Visual BERTs: VisualBERT

A person hits a ball with a tennis racket

VisualBERT Li et al. 2019
Visual BERTs: ViLBERT

(a) Masked multi-modal learning

(b) Multi-modal alignment prediction
Visual BERTs: LXMERT

Learning Cross-Modality Encoder Representations from Transformers
Visual BERTs: Supervised Multimodal Bitransformers

Figure 1: Illustration of the multimodal bitransformer architecture.
Visual BERTs: PixelBert

Pixel-BERT

Sentence Encoder

The woman held a black umbrella

[CLS] The ... a [MASK] umbrella [SEP]

Embedding

Position
Token
Semantic

Cross-Modality Alignment

[CLS] the ... a [MASK] umbrella [SEP]

CNN-based Visual Encoder

CNN Backbone

Pixel Feature Embedding

Random Sampling

Semantic Embedding

Pre-Training Tasks

[MATCH] Image-Text Matching (ITM)

Transformers

black

Masked Language Model (MLM)

Elementwise Sum

[::] Special Token
[V] Visual Token

Misnomer: they mean segment embedding

PixelBert Huang et al. 2020
UNITER

Chen, Yi, Lu, et al. 2020
ViLT (Kim et al. 2021)

Feeding data directly to the transformer.
So many models

<table>
<thead>
<tr>
<th>VL-PTM</th>
<th>Text encoder</th>
<th>Vision encoder</th>
<th>Fusion scheme</th>
<th>Pre-training tasks</th>
<th>Multimodal datasets for pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion Encoder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniter [2020]</td>
<td>BERT</td>
<td>Faster R-CNN</td>
<td>Single stream</td>
<td>MLM+ITM+MR + WRA + MRFR + MRC</td>
<td>CC+COCO+VG+SBU</td>
</tr>
<tr>
<td>OSCAR [2020c]</td>
<td>BERT</td>
<td>Faster R-CNN</td>
<td>Single stream</td>
<td>MLM+ITM</td>
<td>CC+COCO+SBU+Flickr30k + VQA</td>
</tr>
<tr>
<td>InterBert [2020]</td>
<td>BERT</td>
<td>Faster R-CNN</td>
<td>Single stream</td>
<td>MLM+ITM</td>
<td>CC+COCO+SBU</td>
</tr>
<tr>
<td>VL-BERT [2019]</td>
<td>BERT</td>
<td>Faster R-CNN</td>
<td>Dual stream</td>
<td>MLM+MRC+ITM</td>
<td>CC+COCO+SBU</td>
</tr>
<tr>
<td>LXMERT [2019]</td>
<td>BERT</td>
<td>Faster R-CNN</td>
<td>Dual stream</td>
<td>MLM+MRC+ITM</td>
<td>CC+COCO+SBU</td>
</tr>
<tr>
<td>Unified VLP [2020]</td>
<td>UniLM</td>
<td>Faster R-CNN</td>
<td>Single stream</td>
<td>MLM+seq2seq LM</td>
<td>CC</td>
</tr>
<tr>
<td>UNIMO [2020b]</td>
<td>BERT, RoBERTa</td>
<td>Faster R-CNN</td>
<td>Single stream</td>
<td>MLM+seq2seq LM+MRC+MRFR+CMCL</td>
<td>COCO+CC+VG+VQA</td>
</tr>
<tr>
<td>VL-T5 [2021]</td>
<td>T5, BART</td>
<td>Faster R-CNN</td>
<td>Single stream</td>
<td>MLM+VQA+ITM+VG+GC</td>
<td>COCO+VQ</td>
</tr>
<tr>
<td>XGPT [2021]</td>
<td>Transformer</td>
<td>Faster R-CNN</td>
<td>Single stream</td>
<td>MLM+VQA+ITM+VG+GC</td>
<td>COCO+VQ</td>
</tr>
<tr>
<td>ALBEF [2021a]</td>
<td>BERT</td>
<td>Faster R-CNN + Swin transformer</td>
<td>Dual stream</td>
<td>MLM+ITM+MFR</td>
<td>COCO+VG</td>
</tr>
<tr>
<td>SimVLM [2021b]</td>
<td>ViT</td>
<td>Single stream</td>
<td>MLM+ITM+CMCL</td>
<td>CC+COCO+VG+SBU</td>
<td></td>
</tr>
<tr>
<td>WenLan [2021]</td>
<td>EfficientNet</td>
<td>Dual stream</td>
<td>PrefixLM</td>
<td>C4+ALIGNN</td>
<td></td>
</tr>
<tr>
<td>ViLT [2021]</td>
<td>Linear Projection</td>
<td>Single stream</td>
<td>MLM+ITM</td>
<td>CC+COCO+VG+SBU</td>
<td></td>
</tr>
<tr>
<td>Dual Encoder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP [2021]</td>
<td>GPT2</td>
<td>ViT, ResNet</td>
<td>Single stream</td>
<td>CMCL</td>
<td>self-collected</td>
</tr>
<tr>
<td>ALIGN [2021]</td>
<td>BERT</td>
<td>EfficientNet</td>
<td>Single stream</td>
<td>CMCL</td>
<td>self-collected</td>
</tr>
<tr>
<td>DeCLIP [2021b]</td>
<td>GPT2, BERT</td>
<td>ViT, ResNet, RegNetY-64GF</td>
<td>Single stream</td>
<td>CMCL+MLM+CL</td>
<td>CC+self-collected</td>
</tr>
<tr>
<td>Fusion Encoder+Dual Encoder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VLMo [2021a]</td>
<td>BERT</td>
<td>ViT</td>
<td>Single stream</td>
<td>MLM+ITM+CMCL</td>
<td>CC+COCO+VG+SBU</td>
</tr>
<tr>
<td>FLAVA [2021]</td>
<td>ViT</td>
<td>Single stream</td>
<td>MMM+ITM+CMCL</td>
<td>CC+COCO+VG+SBU+RedCaps</td>
<td></td>
</tr>
</tbody>
</table>

Du et al. 2022 VLP Survey
Recommended paper


Figure 3: Visualisation of the (a) single-stream, (b) dual-stream intra-modal and (c) dual-stream inter-modal Transformer layers. (d) shows our gated bimodal layer. The inter-modal layer attends across modalities, while the intra-model layer attends within each modality. Ours can attend to either or both.

Figure 4: Visualisation of the score matrix for (a) single-stream, (b) text–text, (c) vision–vision, (d) text–vision, and (e) vision–text interactions. Shades of green denote the text modality, while purple ones denote the vision modality. Dual-stream scores are sub-matrices of the single-stream scores matrix.
FLAVA (Singh et al., 2021)

Holistic approach to multimodality.

One foundation model spanning V&L, CV and NLP.

Jointly pretrained on:

- unimodal text data (CCNews + BookCorpus)
- unimodal image data (ImageNet)
- public paired image-text data (70M)

All data/models are publicly released.
The PMD dataset

- 70M image-text pairs from public sources

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Example Image</th>
<th>Example Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO</td>
<td><img src="image1.png" alt="COCO Image" /></td>
<td>A close up view of a pizza sitting on a table with a soda in the back.</td>
</tr>
<tr>
<td>Visual Genome</td>
<td><img src="image2.png" alt="Visual Genome Image" /></td>
<td>A lenovo laptop rebooting</td>
</tr>
<tr>
<td>SBU captions</td>
<td><img src="image3.png" alt="SBU Captions Image" /></td>
<td>Front view of basket 13, from the sidewalk in front of the basket.</td>
</tr>
<tr>
<td>Localized narratives</td>
<td><img src="image4.png" alt="Localized Narratives Image" /></td>
<td>The woman is touching a utensil in front of her on the grill stand.</td>
</tr>
<tr>
<td>WIT</td>
<td><img src="image5.png" alt="WIT Image" /></td>
<td>Typocerus balteatus, Subfamily: Flower Longhorns</td>
</tr>
<tr>
<td>RedCaps</td>
<td><img src="image6.png" alt="RedCaps Image" /></td>
<td>Deigdoh falls in india</td>
</tr>
<tr>
<td>CC12M</td>
<td><img src="image7.png" alt="CC12M Image" /></td>
<td>Jumping girl in a green summer dress stock illustration</td>
</tr>
<tr>
<td>YFCC filtered</td>
<td><img src="image8.png" alt="YFCC Filtered Image" /></td>
<td>In the kitchen at the Muse Nissim de Camondo</td>
</tr>
</tbody>
</table>
Problem to solve

FLAVA for multi-domain joint pretraining
(\textit{global contrastive, \texttt{MMM}, \texttt{MIM}, \texttt{MLM}, ...})

- image-text pairs
- unpaired images
- unpaired text

visual recognition (e.g. ImageNet)
language understanding (e.g. GLUE)
multimodal reasoning (e.g. VQA)
How does FLAVA work?
How does FLAVA work?

This cat was wonderful! He was making his daily cleaning on an ancient grave as to say “I am the boss here!”
How does FLAVA work?

This cat was wonderful! He was making his daily cleaning on an ancient grave as to say “I am the boss here!”
How well does it work?

- On average, over 35 tasks, FLAVA obtains impressive performance
SimVLM (Wang et al., 2022)

Slowly moving from contrastive/discriminative to generative.
CoCa Contrastive Captioner (Yu et al., 2022)

Best of both (contrastive and generative) worlds.

Pretraining

Zero-shot, frozen-feature or finetuning
Frozen (Tsimpoukelli, Menick, Cabi, et al., 2021)

Kind of like MMBT but with a better LLM (T5) and a better vision encoder (NF-ResNet).

Multi-Modal Few-Shot Learners!
Flamingo (Alayrac et al., 2022)

80b param model based on Chinchilla. Multi-image.

![Flamingo diagram]

Performance relative to SOTA

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RareAct</th>
<th>NextQA</th>
<th>iVQA</th>
<th>Flick30K</th>
<th>STAR</th>
<th>MSVDQA</th>
<th>OKVQA</th>
<th>HatefulMemes</th>
<th>VixWiz</th>
<th>VATEX</th>
<th>VQAv2</th>
<th>COCO</th>
<th>VixDia</th>
<th>MSRv2TOQA</th>
<th>T很久QVA</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
<td>133%</td>
<td>74%</td>
<td>80%</td>
<td>100%</td>
<td>73%</td>
<td>80%</td>
<td>88%</td>
<td>88%</td>
<td>88%</td>
<td>84%</td>
<td>66%</td>
<td>16%</td>
<td>70%</td>
<td>40%</td>
<td>86%</td>
</tr>
</tbody>
</table>

VixWiz (visual question answering for the blind)

**INPUT**
Image + question

**OUTPUT**
Answer

HatefulMemes (multimodal hateful content classification)

**INPUT**
Image + OCR

**OUTPUT**
Hateful Classification

VATEX (event captioning)

**INPUT**
Video

**OUTPUT**
Event description

A kid performing a kickflip with a skateboard

Textual content:

- a flamingo. They are found in the Caribbean.
- This is a flamingo. They are mainly found in Chile.
- This is a chinchilla. They are very popular in Japan.
Perceiver Resampler

```python
def perceiver_resampler(x_f, time_embeddings, x, num_layers):
    # Add the time position embeddings and flatten.
    x_f = x_f + time_embeddings
    x_f = flatten(x_f) # [T, S, d] -> [T * S, d]
    # Apply the Perceiver Resampler layers.
    for t in range(num_layers):
        # Attention.
        x = x + attention_i(q=x, kv=concat([x_f, x]))
        # Feed forward.
        x = x + ffw_i(x)
    return x
```

---

The Perceiver Resampler model. **"""**
Gated XATTN

Inject visual info directly into a frozen LM via cross-attention (remember FiLM?).

def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn,  # xattn gating parameter - init at 0.
    alpha_dense,  # ffw gating parameter - init at 0.
)

    """Applies a GATED XATTN-DENSE layer."""

    # 1. Gated Cross Attention
    y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
    # 2. Gated Feed Forward (dense) Layer
    y = y + tanh(alpha_dense) * ffw(y)

    # Regular self-attention + FFW on language
    y = y + Frozen_attention(q=y, kv=x)
    y = y + Frozen_ffw(y)

    return y  # output visually informed language features
Why is this funny?

Original image from Karpathy as a “visual Turing test” →

This is a picture of Barack Obama. He is a former president of the United States.

What is he doing?

He is looking at the scale.

How many people are there in this picture?

There are at least 5 people in this picture.

Where is Obama’s foot positioned?

Obama’s foot is positioned on the right side of the scale.

Where was this picture taken?

It was taken in a school.

What happens as a result?

The scale shows a higher weight.

How many mirrors are there?

There are at least 2 mirrors in this picture.

Is the person on the scale aware of it?

I think he is not aware of it.

What is the person standing on?

The person is standing on a rug.

Do you think that’s why people are laughing?

I think so.
BLIP/BLIP2 (Li et al., 2023)

Freeze it all (CLIP-ViT / OPT decoder / FlanT5 encoder-decoder)

**BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models**

Bootstrapping from a **Decoder-based** Large Language Model (e.g. OPT)

- Input Image
- Image Encoder
- Learned Queries
- Q-Former
- Fully Connected
- Output Text: *a cat wearing sunglasses*

Bootstrapping from an **Encoder-Decoder-based** Large Language Model (e.g. FlanT5)

- Input Image
- Image Encoder
- Learned Queries
- Q-Former
- Fully Connected
- LLM Encoder
- LLM Decoder
- Prefix Text: *a cat*
- Suffix Text: *wearing sunglasses*
1. The Audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.

2. The Great Wall of China was built in 221 BC by Emperor Qin Shi Huang to protect the emperor's capital from invaders from the north.

3. It is a genus of flowering plants in the orchid family, orchidaceae. It is native to Eastern Asia and South East Asia. It is a perennial flowering plant that grows up to a height of 1 m.

4. Is this photo unusual?
   - Yes, it's a house that looks like it's upside down.
   - How could someone get out of the house?
   - It has a slide on the side of the house.

5. What are shown in the photo?
   - A man and a chicken.
   - What does the man feel and why?
   - He is scared of the chicken because it is flying at him.

6. What are the ingredients I need to make this?
   - Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.
   - What is the first step?
   - Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

7. Which city is this?
   - Singapore.
   - Why do you think so?
   - The city has a statue of a merlion.

8. What happened at the end of this movie?
   - The Titanic sank.
   - Did Leonardo DiCaprio's character survive?
   - No, he drowned.

9. What is in the photo?
   - A pizza that looks like a cat.
   - What is the nose made of?
   - A slice of pepperoni.
Multimodal “Chain of Thought” (Zhang et al., 2023)

Providing a rationale helps give the right answer.

**Question:** Which property do these two objects have in common?  
**Context:** Select the better answer.  
**Options:** (A) soft  (B) salty

**Rationale**
Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.

**Answer**
The answer is (B).
KOSMOS-1 (Huang et al., 2023)

LLMs => MLLMs == FM

Figure 5: Multimodal Chain-of-Thought prompting enables KOSMOS-1 to generate a rationale first, then to tackle complex question-answering and reasoning tasks.
Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. Multimodal foundation models
5. **Evaluation**
6. Beyond images: Other modalities
7. Where to next?
COCO - Common Objects in Context

Super impactful datasets (Lin et al. 2014; Chen et al. 2015)

Main multimodal tasks:

- Image captioning
- Image-caption retrieval

Similar datasets:

- Flickr30k, ConceptualCaptions, VisualGenome, SBU, RedCaps, LAION

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

cocodataset.org
VQA - Visual Question Answering (Antol et al., 2015)

- The dominant task in vision and language.
  - VQA/VQAv2 citations: 4305+1684
  - COCO Captions: 1647
  - Flickr30k: 1228

- At first the “V” in VQA was found to not matter all that much, so a follow-up VQAv2 dataset was created (Goyal et al., 2017).

- There’s also GQA (Hudson & Manning, 2019).
CLEVR (Johnson et al., 2016)

Compositional language and elementary visual reasoning diagnostics in a controlled setting.

Hand crafted for measuring compositionality.

Q: Are there an equal number of large things and metal spheres? Q: What size is the cylinder that is left of the brown metal 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 metal things?

Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.
Hateful Memes (Kiela et al., 2020)

Motivated by the shortcomings of other V&L datasets: we need something that is harder, more realistic, and requires true multimodal reasoning and understanding.
Hateful Memes

Highly trained annotators, so: decent quality but small and expensive

Key concept: benign confounders

A “challenge set” for the community to do zero-shot/finetuning from pretrained
Hateful Memes

Findings in the paper:

- Big gap with human performance.
- Region features (as opposed to grid) seem to help.
- Earlier fusion is better than middle, is better than late.
- Multimodal pretraining doesn’t really work.
Hateful Memes Competition

After the paper came a $100k competition on an unseen test set:

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Unseen Dev Acc.</th>
<th>Unseen Dev AUROC</th>
<th>Unseen Test Acc.</th>
<th>Unseen Test AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal</td>
<td>Image-Region</td>
<td>61.48</td>
<td>53.54</td>
<td>60.28±0.18</td>
<td>54.64±0.80</td>
</tr>
<tr>
<td></td>
<td>Text BERT</td>
<td>60.37</td>
<td>60.88</td>
<td>63.60±0.54</td>
<td>62.65±0.40</td>
</tr>
<tr>
<td>Multimodal</td>
<td>Late Fusion</td>
<td>61.11</td>
<td>61.00</td>
<td>64.06±0.02</td>
<td>64.44±1.60</td>
</tr>
<tr>
<td></td>
<td>Concat BERT</td>
<td>64.81</td>
<td>65.42</td>
<td>65.90±0.82</td>
<td>66.28±0.66</td>
</tr>
<tr>
<td></td>
<td>MMBT-Grid</td>
<td>67.78</td>
<td>65.47</td>
<td>66.85±1.61</td>
<td>67.24±2.53</td>
</tr>
<tr>
<td></td>
<td>MMBT-Region</td>
<td>70.04</td>
<td>71.54</td>
<td>70.10±1.39</td>
<td>72.21±0.20</td>
</tr>
<tr>
<td></td>
<td>ViLBERT</td>
<td>69.26</td>
<td>72.73</td>
<td>70.86±0.70</td>
<td>73.39±1.32</td>
</tr>
<tr>
<td></td>
<td>Visual BERT</td>
<td>69.67</td>
<td>71.10</td>
<td>71.30±0.68</td>
<td>73.23±1.04</td>
</tr>
<tr>
<td>Multimodal</td>
<td>ViLBERT CC</td>
<td>70.37</td>
<td>70.78</td>
<td>70.03±1.07</td>
<td>72.78±0.50</td>
</tr>
<tr>
<td>(Multimodal Pretraining)</td>
<td>Visual BERT COCO</td>
<td>70.77</td>
<td>73.70</td>
<td>69.95±1.06</td>
<td>74.59±1.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Winner characteristics: frameworks matter, SOTA pretrained models, ensembles, entities, faces and external knowledge. STILL FAR FROM SOLVED.
Winoground

How good is CLIP really?

Some relevant ideas/findings from NLP:

- Winograd schemas
  "The [trophy] doesn't fit in the [suitcase] because it is too [large/small]"
- Word order may not matter all that much

Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha††  Robin Jia†  Dieuwke Hupkes†  Joelle Pineau††

Adina Williams†  Douwe Kiela†
Winoground

- Examples written by linguist experts
- Using Getty Images API
- Simple way to measure by comparing scores
- In some cases, very difficult and requiring world knowledge
Winoground Findings

SOTA models often perform below chance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Text</th>
<th>Image</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTurk Human</td>
<td>89.50</td>
<td>88.50</td>
<td>85.50</td>
</tr>
<tr>
<td>Random Chance</td>
<td>25.00</td>
<td>25.00</td>
<td>16.67</td>
</tr>
<tr>
<td>VinVL</td>
<td>37.75</td>
<td>17.75</td>
<td>14.50</td>
</tr>
<tr>
<td>UNITER\textsubscript{large}</td>
<td>38.00</td>
<td>14.00</td>
<td>10.50</td>
</tr>
<tr>
<td>UNITER\textsubscript{base}</td>
<td>32.25</td>
<td>13.25</td>
<td>10.00</td>
</tr>
<tr>
<td>ViLLA\textsubscript{large}</td>
<td>37.00</td>
<td>13.25</td>
<td>11.00</td>
</tr>
<tr>
<td>ViLLA\textsubscript{base}</td>
<td>30.00</td>
<td>12.00</td>
<td>8.00</td>
</tr>
<tr>
<td>VisualBERT\textsubscript{base}</td>
<td>15.50</td>
<td>2.50</td>
<td>1.50</td>
</tr>
<tr>
<td>ViLT (ViT-B/32)</td>
<td>34.75</td>
<td>14.00</td>
<td>9.25</td>
</tr>
<tr>
<td>LXMERT</td>
<td>19.25</td>
<td>7.00</td>
<td>4.00</td>
</tr>
<tr>
<td>ViLBERT\textsubscript{base}</td>
<td>23.75</td>
<td>7.25</td>
<td>4.75</td>
</tr>
<tr>
<td>UniIT\textsubscript{ITM finetuned}</td>
<td>19.50</td>
<td>6.25</td>
<td>4.00</td>
</tr>
<tr>
<td>CLIP (ViT-B/32)</td>
<td>30.75</td>
<td>10.50</td>
<td>8.00</td>
</tr>
<tr>
<td>VSE++\textsubscript{COCO} (ResNet)</td>
<td>22.75</td>
<td>8.00</td>
<td>4.00</td>
</tr>
<tr>
<td>VSE++\textsubscript{COCO} (VGG)</td>
<td>18.75</td>
<td>5.50</td>
<td>3.50</td>
</tr>
<tr>
<td>VSE++\textsubscript{Flickr30k} (ResNet)</td>
<td>20.00</td>
<td>5.00</td>
<td>2.75</td>
</tr>
<tr>
<td>VSE++\textsubscript{Flickr30k} (VGG)</td>
<td>19.75</td>
<td>6.25</td>
<td>4.50</td>
</tr>
<tr>
<td>VSRN\textsubscript{COCO}</td>
<td>17.50</td>
<td>7.00</td>
<td>3.75</td>
</tr>
<tr>
<td>VSRN\textsubscript{Flickr30k}</td>
<td>20.00</td>
<td>5.00</td>
<td>3.50</td>
</tr>
</tbody>
</table>

STILL FAR FROM SOLVED.
DALL-E2 on Winoground I

More Winoground prompts. 1st run, no cherry-picking. To all I added "digital art". That helps w/ the composition (and aesthetic imo), particularly for less common things 🌿.
DALL-E2 on Winoground II

More Winoground prompts. 1st run, no cherry-picking. To all I added "digital art". That helps w/ the composition (and aesthetic imo), particularly for less common things 🎨.

STILL NOT SOLVED
Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. Multimodal foundation models
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?
Speech / audio

Can EASILY do another full lecture just on this topic.

Recent cool example: Whisper, trained on 680,000 hours of multilingual multitask data.

We can also just treat audio as vision ;)

---

Figure 3: Illustration of the Neural Auditory Embedding method, using a convolutional neural network. The auditory signal is converted to a spectrogram which is fed to the neural network for classification. The pre-softmax layer, FC7, is transferred and taken as the neural audio embedding (NAE) for the given sound file.

Kiela et al., 2017
Video and text and audio

MERLOT (Zellers et al., 2021)
MERLOT Reserve (idem, 2022)

Joint Encoder (Transformer) for all modalities and timesteps

Predict MASKed text and audio

Inputs for segment \( t \)

*jigging, popcorn popping*

*turn the heat on high*

*music*

*add a lid and then*

*sizzling*

*Jiggle it while it pops*

*it’s thin plastic, so it’ll be easy to cut.*

*So I’ll cut it with a circular [MASK]*
Grounded language learning in simulated environments

Hermann, Hill, et al. 2017

Das et al., 2018
Text to 3D

POINT-E (Nichol, Jun, et al., 2022)
Olfactory embeddings (Kiela et al., 2015)

Bag of chemical compounds model.

<table>
<thead>
<tr>
<th>Smell Label</th>
<th>Chemical Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melon</td>
<td>✓</td>
</tr>
<tr>
<td>Pineapple</td>
<td>✓</td>
</tr>
<tr>
<td>Licorice</td>
<td>✓</td>
</tr>
<tr>
<td>Anise</td>
<td>✓</td>
</tr>
<tr>
<td>Beer</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smell Label</th>
<th>Phenethyl acetate</th>
<th>Isoamyl butyrate</th>
<th>Anisyl butyrate</th>
<th>Myrcene</th>
<th>Syringaldehyde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melon</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pineapple</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Licorice</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Anise</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Beer</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2: A BoCC model.
Outline

1. Early models
2. Multimodal fusion
3. Contrastive models
4. Multimodal foundation models
5. Other modalities
6. Evaluation
7. Where to next?
One foundation model to rule them all

There will modality-agnostic foundation models that can read and generate many modalities.

These models can be bigger and be trained on vastly more data. Parameters will be shared in interesting ways.

Automatic alignment from unpaired unimodal data will become a big topic.
Multimodal scaling laws

We are just beginning to understand multimodal scaling laws, lots of interesting work to do here in understanding trade-offs.

**Scaling Laws for Generative Mixed-Modal Language Models**

Armen Aghajanyan*,†, Lili Yu*,†, Alexis Conneau†, Wei-Ning Hsu†

Karen Hambardzumyan◊, Susan Zhang†, Stephen Roller†, Naman Goyal†

Omer Levy† & Luke Zettlemoyer†, ◊
Retrieval augmented generative multimodal models

These can be multimodal

Lewis et al., 2021
Better evaluation and benchmarking

We need better measurement.
Thanks for listening!

Thank you!

Email: dkiela@stanford.edu
Twitter: @douwekiela