Language Technologies Institute

Carnegie Mellon University

Advanced Multimodal Machine Learning

Lecture 10.1: Multimodal Fusion and New Directions

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Lecture Objectives

- Recap: multimodal fusion
- Kernel methods for fusion
  - Multiple Kernel Learning
- Transformers through the lens for kernel
- New directions in multimodal machine learning
  - Representation
  - Alignment
  - Fusion

Quick Recap: Multimodal Fusion

Multimodal fusion

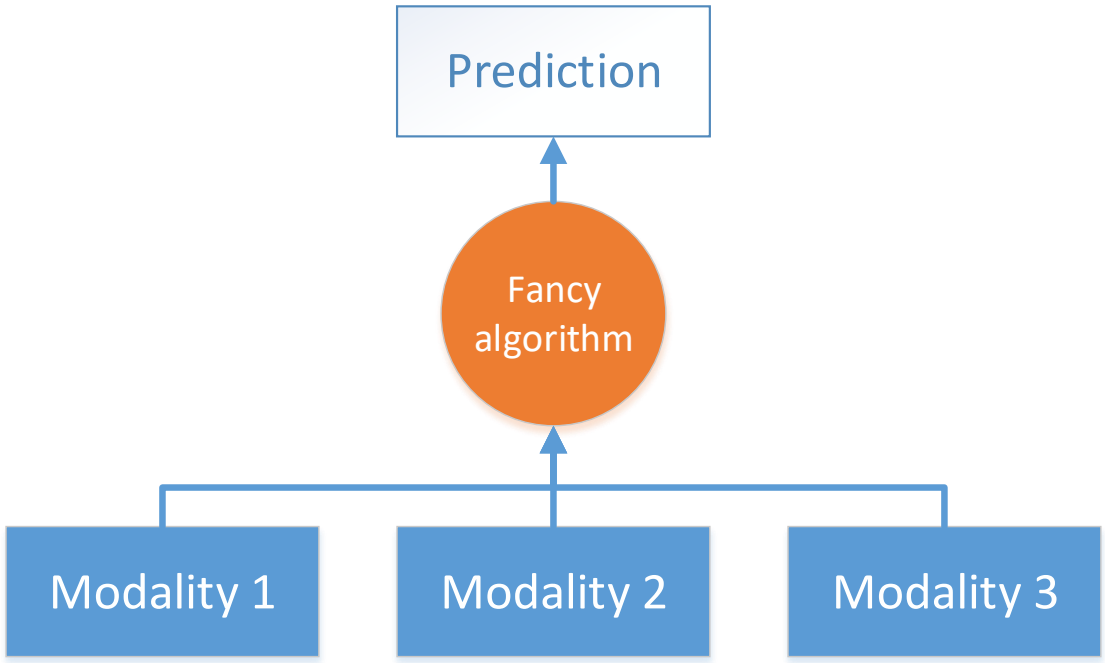
- Process of joining information from two or more modalities to perform a prediction
- Examples
  - Audio-visual speech recognition
  - Audio-visual emotion recognition
  - Multimodal biometrics
  - Speaker identification and diarization
  - Visual/Media Question answering

(a) answer-phone (a) get-out-car (a) fight-person

Multimodal Fusion

Two major types:

- Model Free
  - Early, late, hybrid
- Model Based
  - Neural Networks
  - Graphical models
  - Kernel Methods



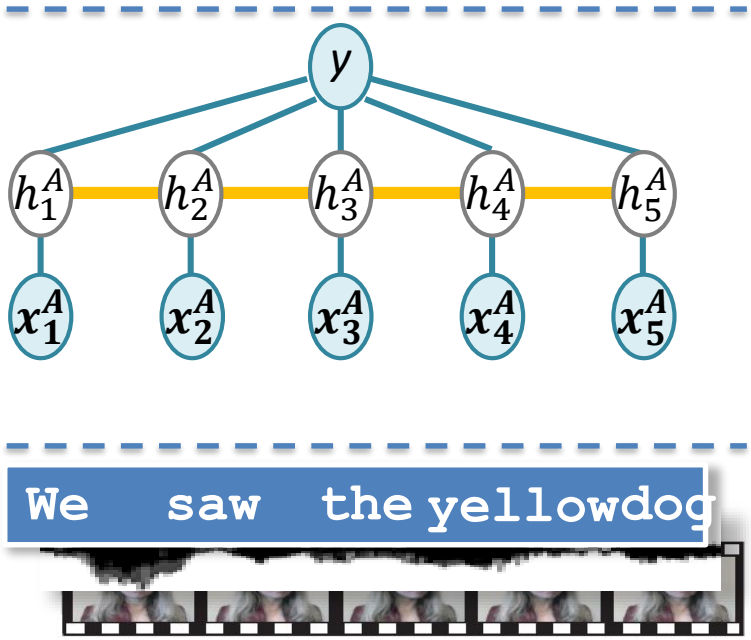
Graphical Model: Learning Multimodal Structure

Modality-***private*** structure

- Internal grouping of observations

Modality-***shared*** structure

- Interaction and synchrony



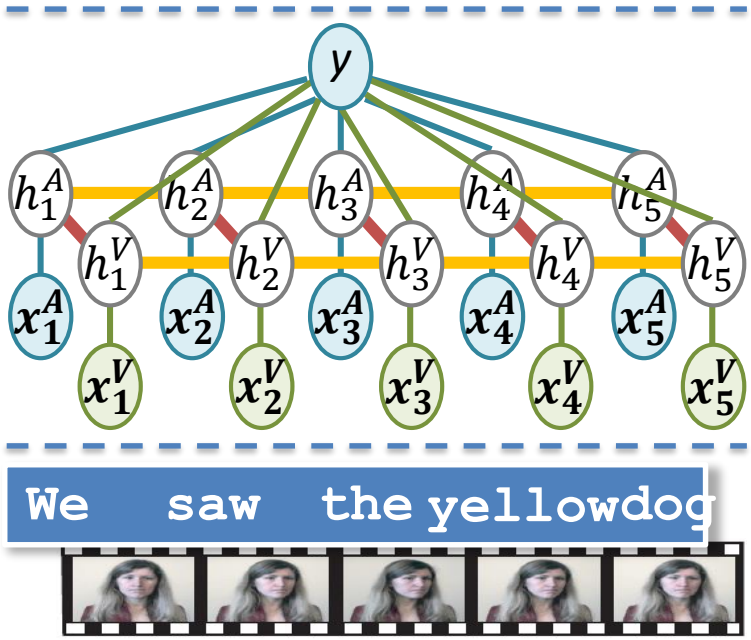
Multi-view Latent Variable Discriminative Models

Modality-***private*** structure

- Internal grouping of observations

Modality-***shared*** structure

- Interaction and synchrony



$$p(y|\, \boldsymbol{x^A}, \boldsymbol{x^V}; \boldsymbol{\theta}) = \sum_{\boldsymbol{h^A}, \boldsymbol{h^V}} p\left(y, \boldsymbol{h^A}, \boldsymbol{h^V} \middle| \boldsymbol{x^A}, \boldsymbol{x^V}; \boldsymbol{\theta}\right)$$

➢ Approximate inference using loopy-belief

Multimodal Fusion: Multiple Kernel Learning

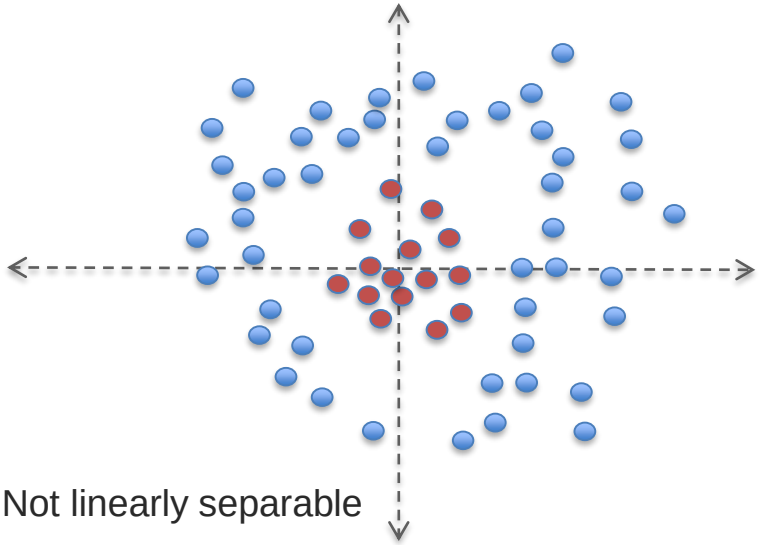
What is a Kernel function?

- A kernel function: Acts as a similarity metric between data points

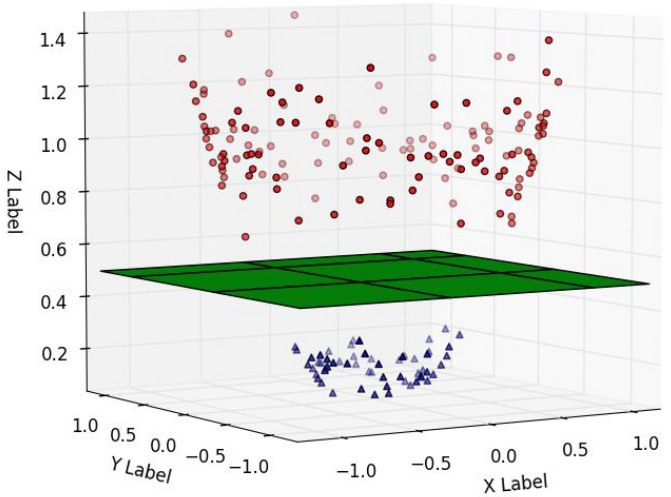
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle, \text{ where } \phi: D \to Z$$

  - Kernel function performs an inner product in feature map space ϕ
  - Inner product (a generalization of the dot product) is often denoted as $\langle ., . \rangle$ in SVM papers
  - $\boldsymbol{x} \in \mathbb{R}^D$ (but not necessarily), but $\phi(\boldsymbol{x})$ can be in any space – same, higher, lower or even in an infinite dimensional space

Non-linearly separable data



Not linearly separable



Same data, but now linearly separable

- Want to map our data to a linearly separable space
- Instead of $\boldsymbol{x}$, want $\phi(\boldsymbol{x})$, in a separable space ($\phi(\boldsymbol{x})$ is a feature map)
- What if $\phi(\boldsymbol{x})$ is much higher dimensional? We do not want to learn more parameters and mapping could become very expensive

Radial Basis Function Kernel (RBF)

- Arguably the most popular SVM kernel
- $K(x_i, x_j) = \exp -\frac{1}{2\sigma^2}\|x_i - x_j\|^2$
- $\phi(\boldsymbol{x}) = ?$
  - It is infinite dimensional and fairly involved, no easy way to actually perform the mapping to this space, but we know what an inner product looks like in it
- σ = ?
  - a hyperparameter
  - With a really low sigma the model becomes close to a KNN approach (potentially very expensive)

Some other kernels

- Other kernels exist
  - Histogram Intersection Kernel
    - good for histogram features
  - String kernels
    - specifically for text and sentence features
  - Proximity distribution kernel
  - (Spatial) pyramid matching kernel

Kernel CCA

If we remember CCA it used only inner products in definitions when dealing with data, that means we can again use kernels

$$(w_1^*, w_2^*) = \underset{w_1, w_2}{\operatorname{argmax}} \frac{w_1' \Sigma_{12} w_2}{\sqrt{w_1' \Sigma_{11} w_1 w_2' \Sigma_{22} w_2}} = \underset{w_1' \Sigma_{11} w_1 = w_2' \Sigma_{22} w_2 = 1}{\operatorname{argmax}} w_1' \Sigma_{12} w_2$$

We can now map into a high-dimensional non-linear space instead

$$(\alpha_1^*, \alpha_2^*) = \underset{\alpha_1, \alpha_2}{\operatorname{argmax}} \frac{\alpha_1' K_1 K_2 \alpha_2}{\sqrt{(\alpha_1' K_1^2 \alpha_2)\ (\alpha_1' K_2^2 \alpha_2)}} = \underset{\alpha_1' K_1^2 \alpha_1 = \alpha_2' K_2^2 \alpha_2 = 1}{\operatorname{argmax}} \alpha_1' K_1 K_2 \alpha_2,$$

[Lai et al. 2000]

Different properties of different signals

How do we deal with heterogeneous or multimodal data?

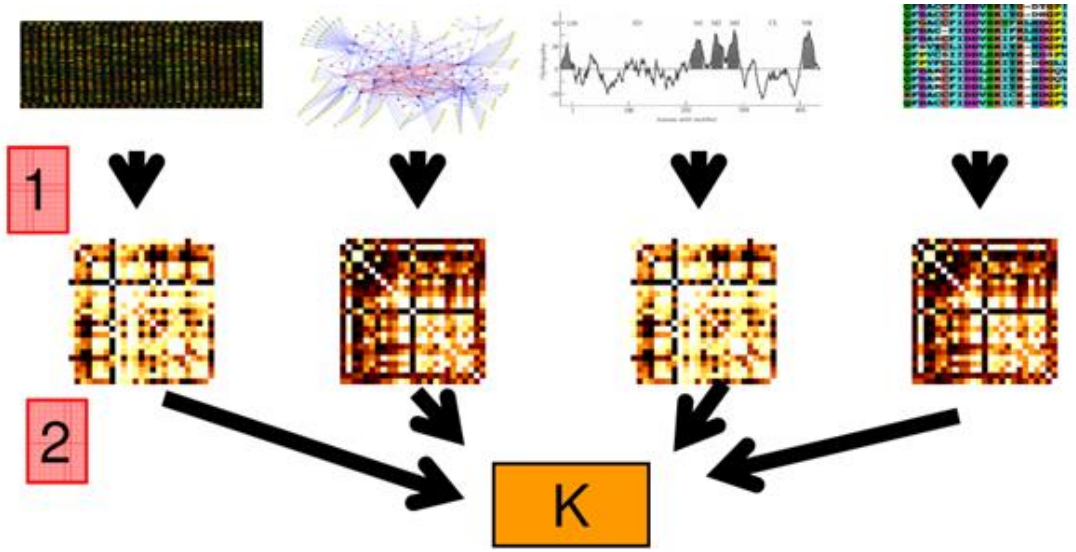
- The data of interest is not in a joint space so appropriate kernels for each modality might be different

Multiple Kernel Learning (MKL) is a way to address this

- Was popular for image classification and retrieval before deep learning approaches came around (winner of 2010 VOC challenge, ImageClef 2011 challenge)
- MKL - fell slightly out of favor when deep learning approaches became popular
- Still useful when large datasets are not available

Multiple Kernel Learning

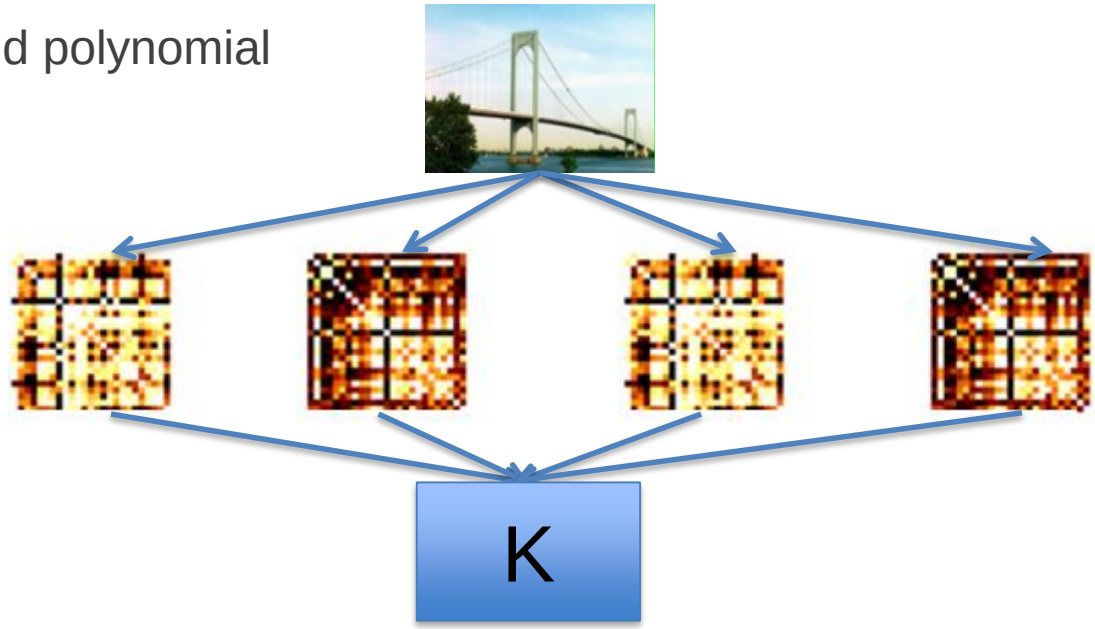
- Instead of providing a single kernel and validating which one works optimize in a family of kernels (or different families for different modalities)
- Works well for unimodal and multimodal data, very little adaptation is needed



[Lanckriet 2004]

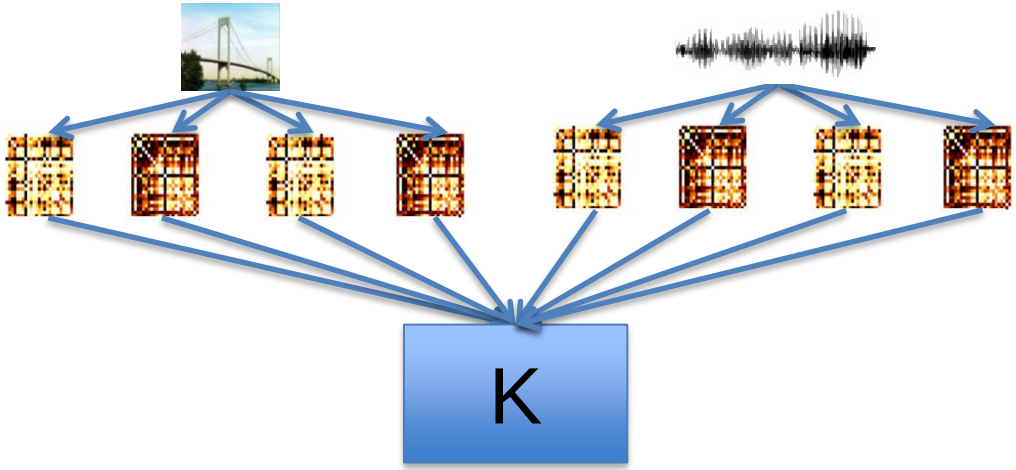
MKL in Unimodal Case

- Pick a family of kernels and learn which kernels are important for the classification case
- For example a set of RBF and polynomial kernels



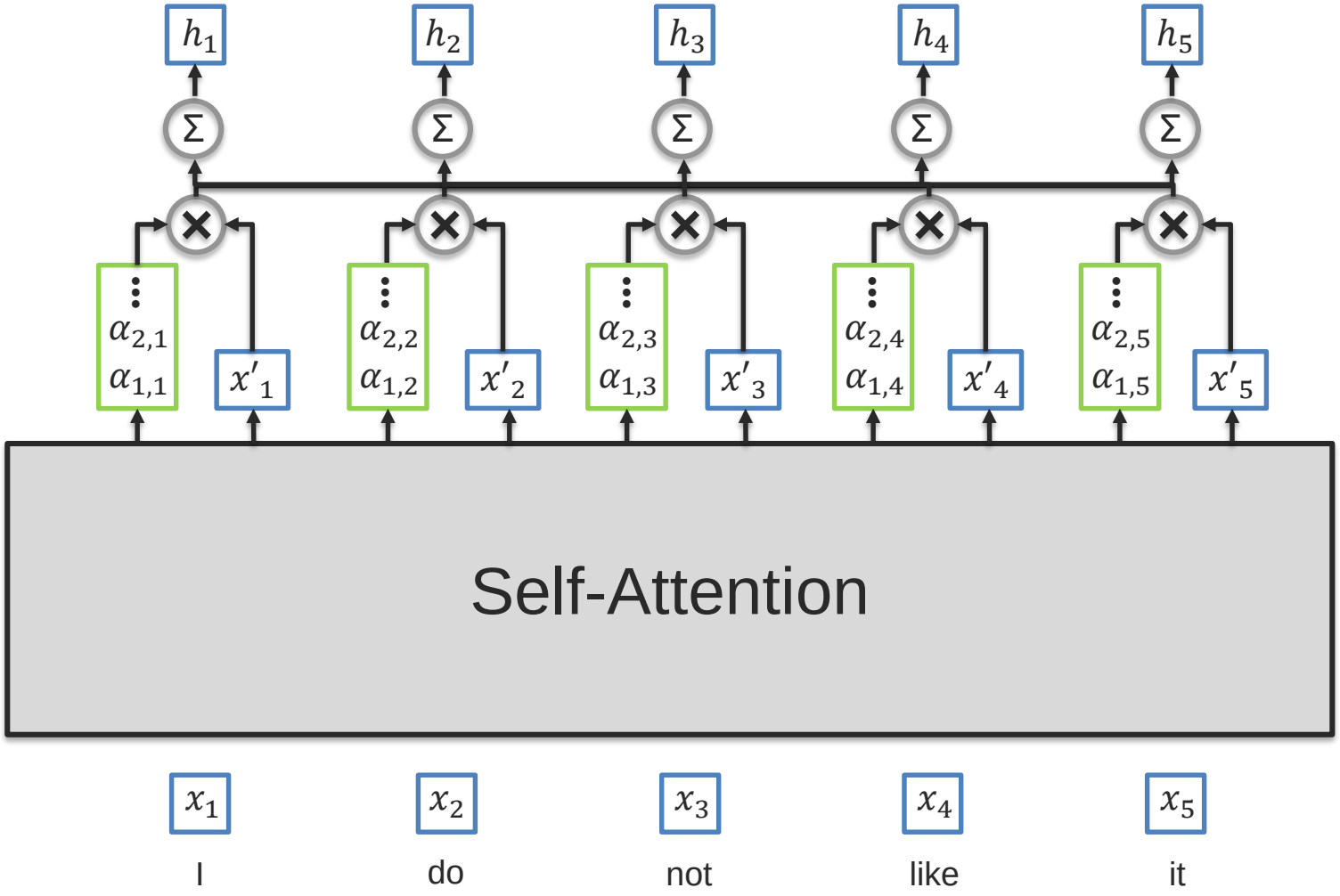
MKL in Multimodal/Multiview Case

- Pick a family of kernels for each modality and learn which kernels are important for the classification case
- Does not need to be different modalities, often we use different views of the same modality (HOG, SIFT, etc.)

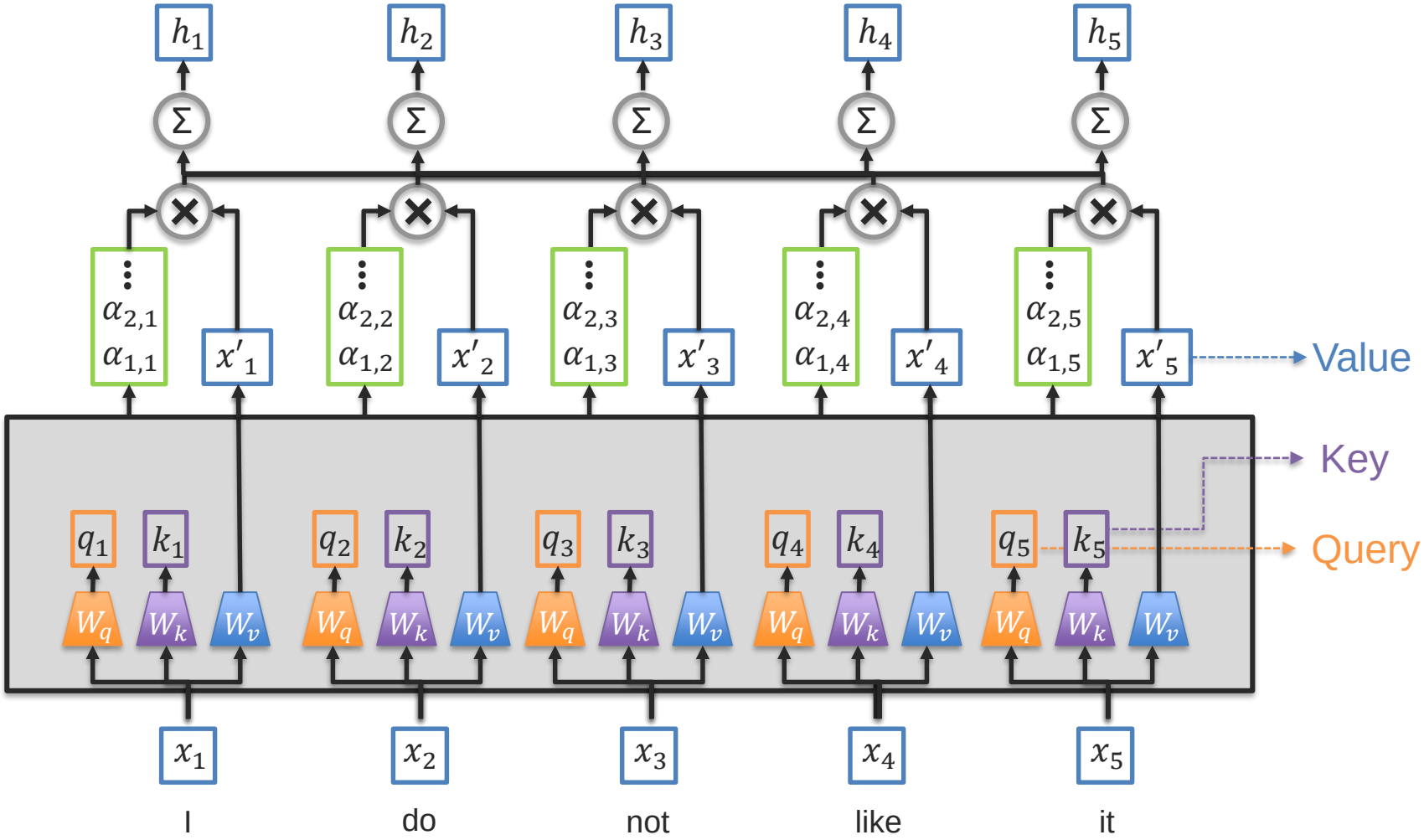


Kernel functions for Transformer networks

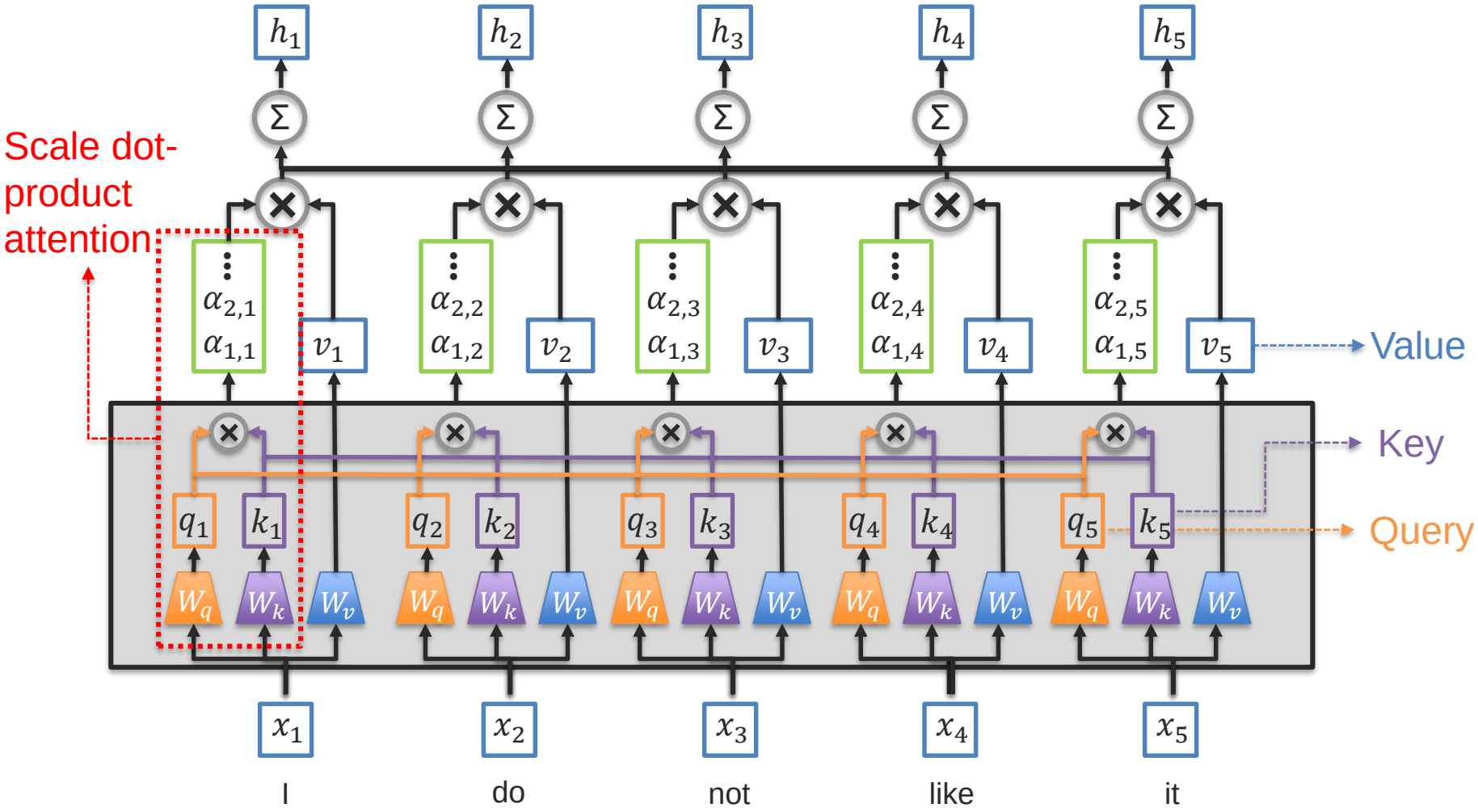
Recap: Self-Attention



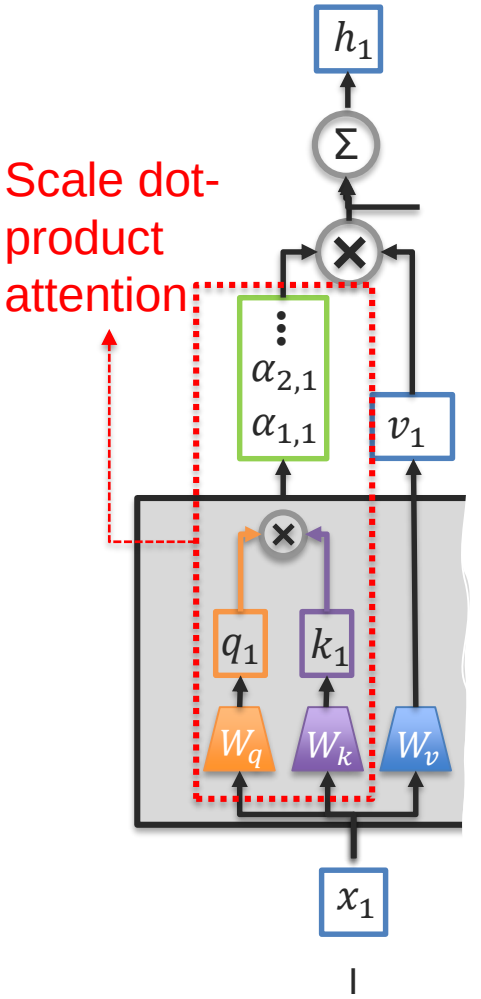
Recap: Transformer Self-Attention



Transformer Self-Attention



Transformer's Attention Function

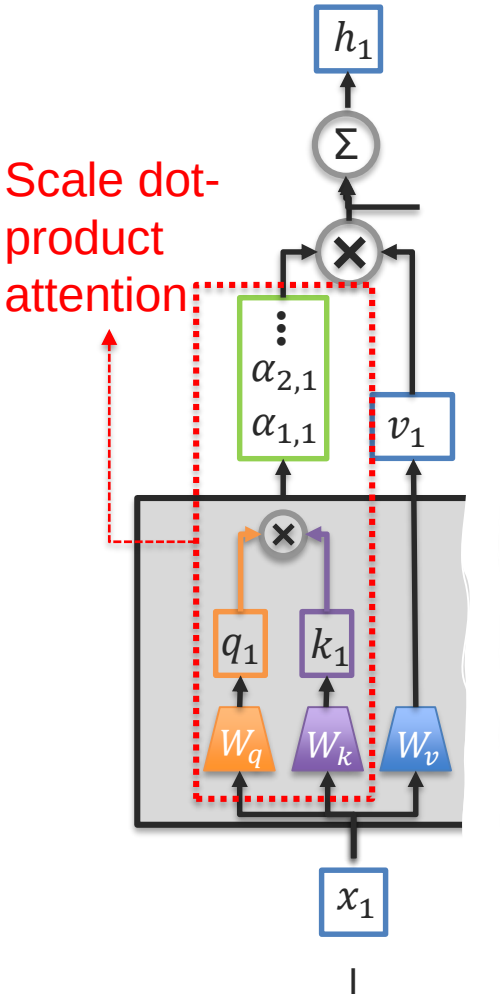


Scale dot-product attention:

$$\boldsymbol{\alpha} = softmax\left(\frac{\boldsymbol{x_q W_q}(\boldsymbol{x_k W_k})^T}{\sqrt{d_k}}\right)$$

How can you interpret it as a kernel similarity function?

Transformer's Attention Function



Scale dot-product attention:

$$\boldsymbol{\alpha} = softmax\left(\frac{\boldsymbol{x_q W_q}(\boldsymbol{x_k W_k})^T}{\sqrt{d_k}}\right)$$

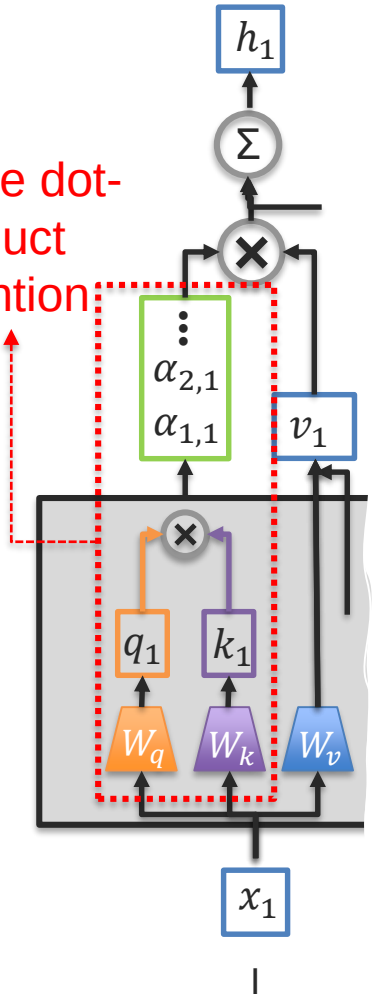
Kernel-formulated attention:

$$\boldsymbol{\alpha} = \frac{k(\boldsymbol{x_q}, \boldsymbol{x_k})}{\sum_{\{\boldsymbol{x'_k}\}} k(\boldsymbol{x_q}, \boldsymbol{x'_k})}$$

What is the impact of the kernel function?

Tsai et al., Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel, EMNLP 2019

Transformer's Attention Function



Scale dot-product attention

Conventional Transformer

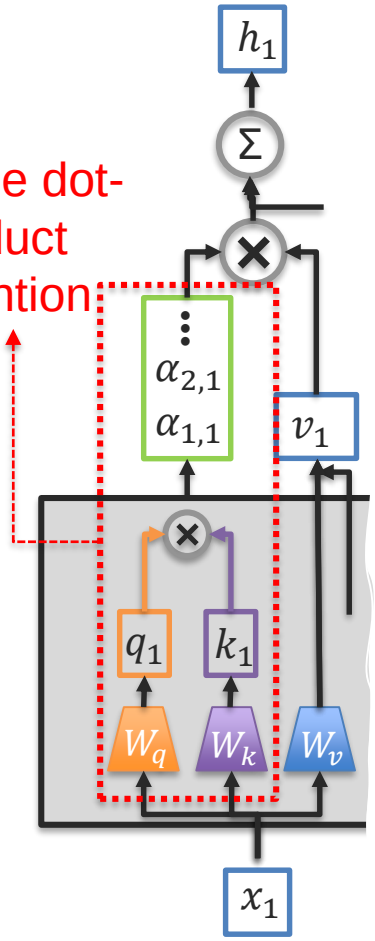
What is the impact of the kernel function?

Type	Kernel Form	NMT (BLEU↑) Asym. ($W_q \neq W_k$)	NMT (BLEU↑) Sym. ($W_q = W_k$)
Linear	$\langle f_a W_q, f_b W_k \rangle$	not converge	not converge
Polynomial	$\left(\langle f_a W_q, f_b W_k \rangle\right)^2$	32.72	32.43
Exponential	$\exp\left(\frac{\langle f_a W_q, f_b W_k \rangle}{\sqrt{d_k}}\right)$	33.98	33.78
RBF	$\exp\left(-\frac{\|f_a W_q - f_b W_k\|^2}{\sqrt{d_k}}\right)$	**34.26**	34.14

What is the best way to integrate the position embedding?

Tsai et al., Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel, EMNLP 2019

Transformer's Attention Function



Scale dot-product attention

What is the best way to integrate the position embedding?

	PE Incorporation	Kernel Form	NMT (BLEU↑)
Vaswami et al →	Direct-Sum	$k_{\exp}\left(f_q + t_q, f_k + t_k\right)$	33.98
	Look-up Table	$L_{t_q - t_k, f_q} \cdot k_{\exp}\left(f_q, f_k\right)$	34.12
Transformer XL →	Product Kernel	$k_{\exp}\left(f_q, f_k\right) \cdot k_{f_q}\left(t_q, t_k\right)$	33.62
Proposed →	Product Kernel	$k_F\left(f_q, f_k\right) \cdot k_T\left(t_q, t_k\right)$	**34.71**

Proposed →

$$\text{with } k_F(f_q, f_k) = \exp\left(\frac{\langle f_q W_F, f_k W_F \rangle}{\sqrt{d_k}}\right)$$

$$\text{and } k_T(t_q, t_k) = \exp\left(\frac{\langle t_q W_T, t_k W_T \rangle}{\sqrt{d_k}}\right),$$

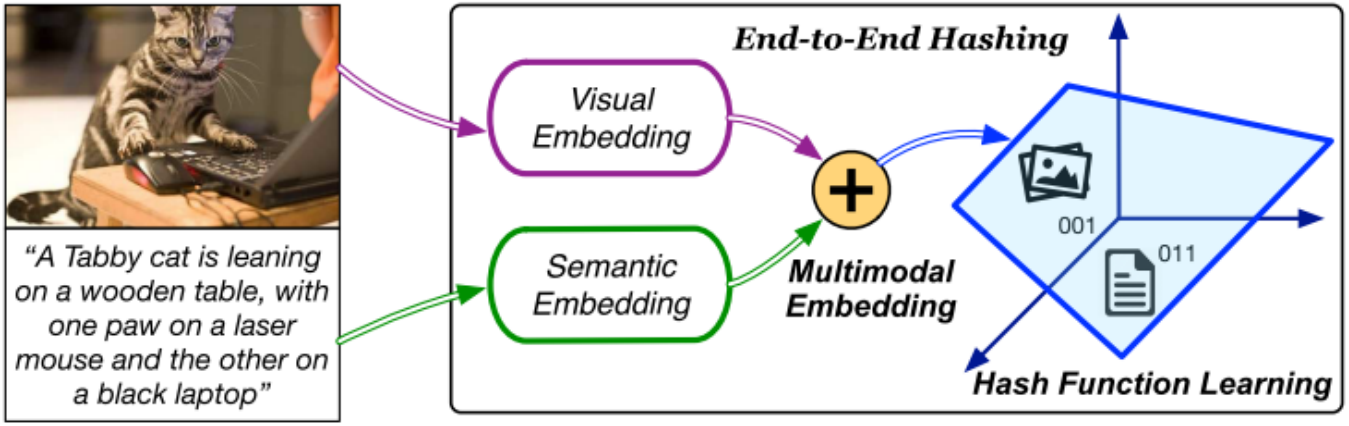
Same weight matrices!

Tsai et al., Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel, EMNLP 2019

New(-ish) Directions: Representation

Representation 1: Hash Function Learning

- We talked about coordinated representations, but mostly enforced “simple” coordination
- We can make embeddings more suitable for retrieval
  - Enforce a Hamming space (binary n-bit space)

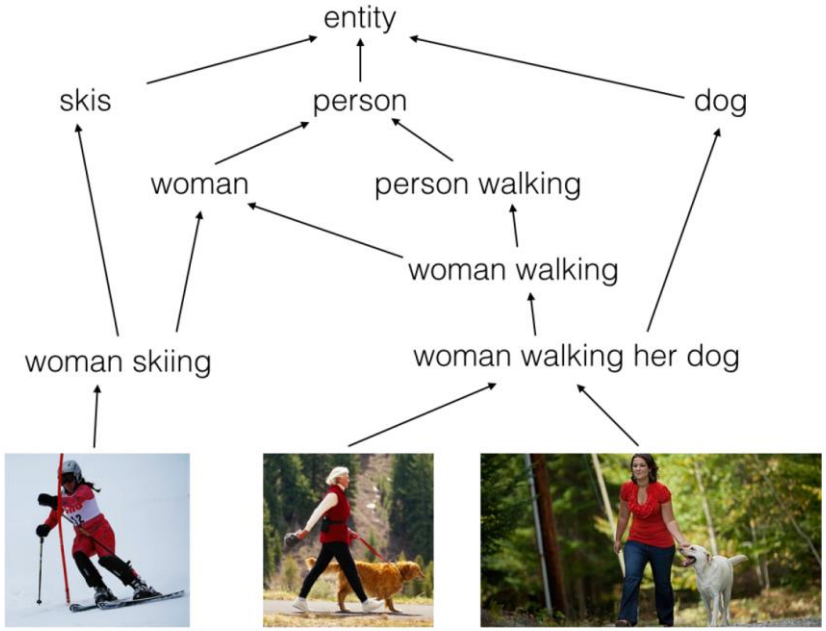


[Cao et al. Deep visual-semantic hashing for cross-modal retrieval, KDD 2016]

Representation 2: Order-Embeddings

- We talked about coordinated representations, but mostly enforced “simple” coordination
  - Can we take it further?
- Replaces symmetric similarity

$$x \preceq y \text{ if and only if } \bigwedge_{i=1}^{N} x_i \geq y_i$$



- Enforce approximate structure when training the embedding

[Vendrov et al. Order-embeddings of images and language, ICLR 2016]

Representation 3: Hierarchical Multimodal LSTM

General Image Captioning:

- *'A man is standing in front of towers.'*

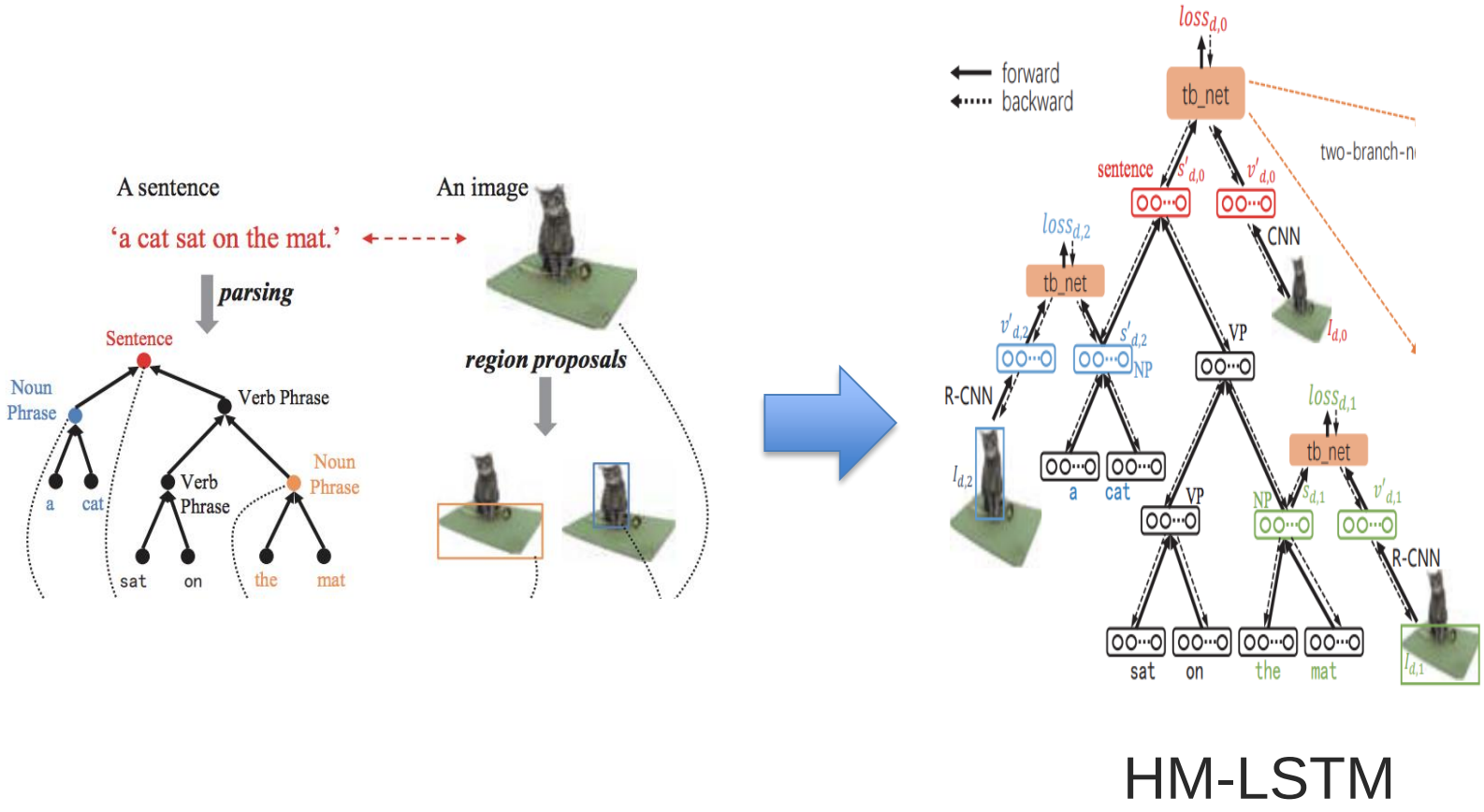
Region-oriented, detailed, and phrase-level Captioning:

- *'a man with a blue hat and sunglasses'*
- *'a girl in red jacket and black dress'*
- *'several white towers with golden spire'*

- Uses these region-based phrases to hierarchically build sentences

Niu, Zhenxing, et al. "Hierarchical multimodal lstm for dense visual-semantic embedding." *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017.

Representation 3: Hierarchical Multimodal LSTM

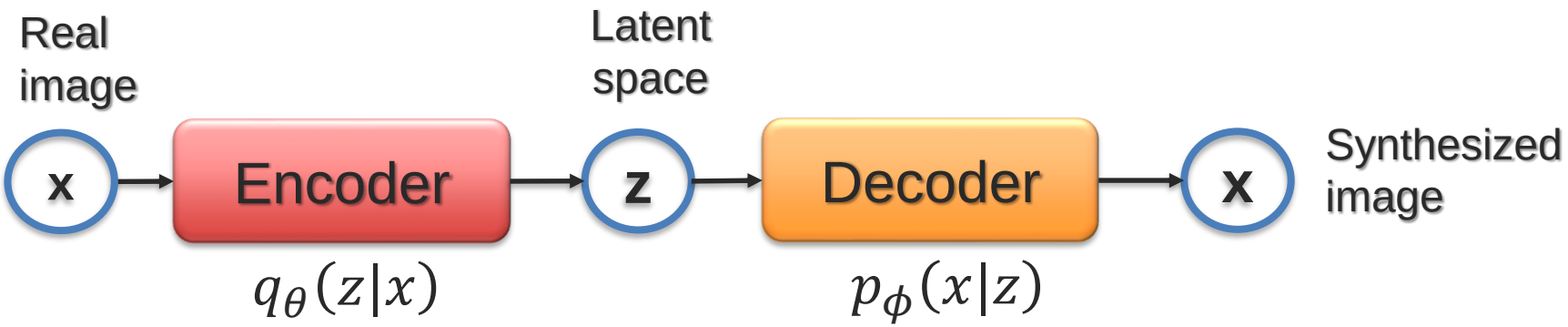


HM-LSTM

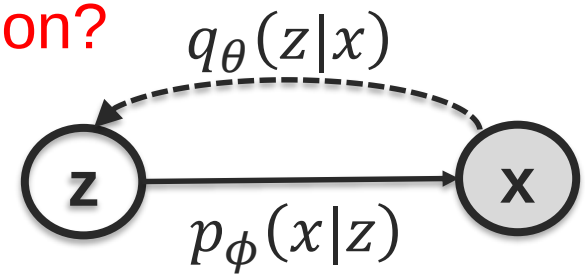
Niu, Zhenxing, et al. "Hierarchical multimodal lstm for dense visual-semantic embedding." *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017.

Representation 4: Multimodal VAE (MVAE)

Variational Autoencoder (VAE):



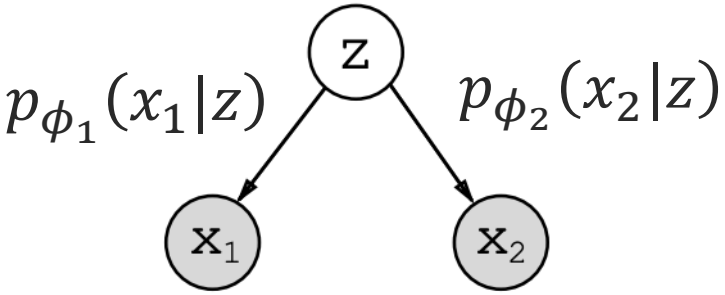
Graphical model representation?

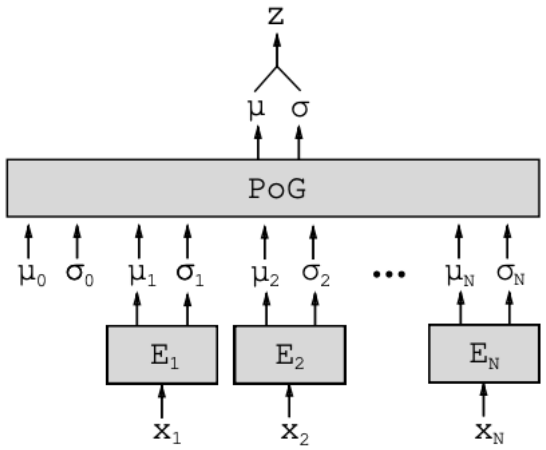


With multimodal observations?

Representation 4: Multimodal VAE (MVAE)

Multimodal variational autoencoder (MVAE)





What will be the encoder?

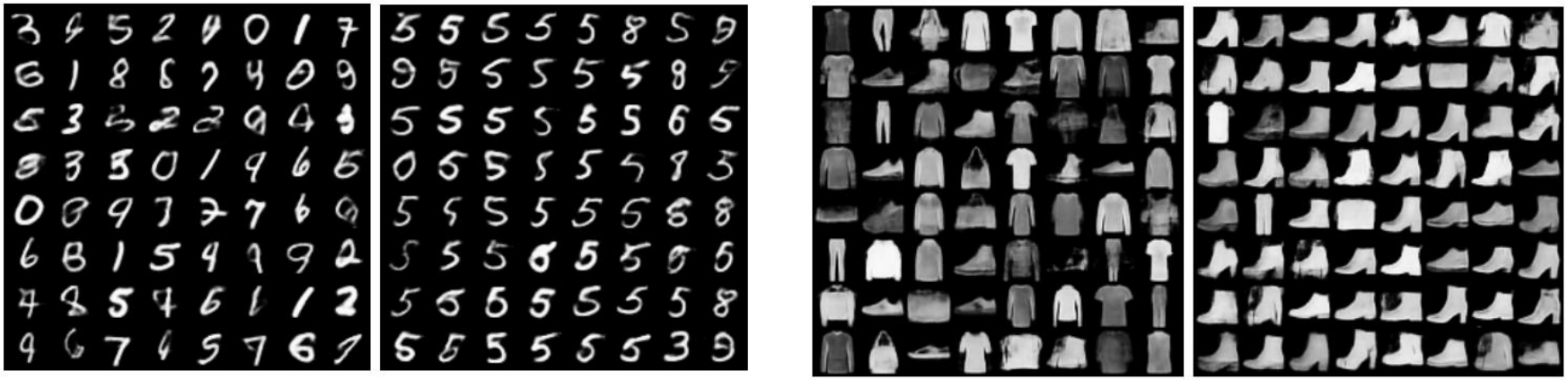
Product of expert (PoG) to combine the variational parameters from the unimodal encoders

[Wu, Mike, and Noah Goodman. “Multimodal Generative Models for Scalable Weakly-Supervised Learning.”, NIPS 2018]

Representation 4: Multimodal VAE (MVAE)

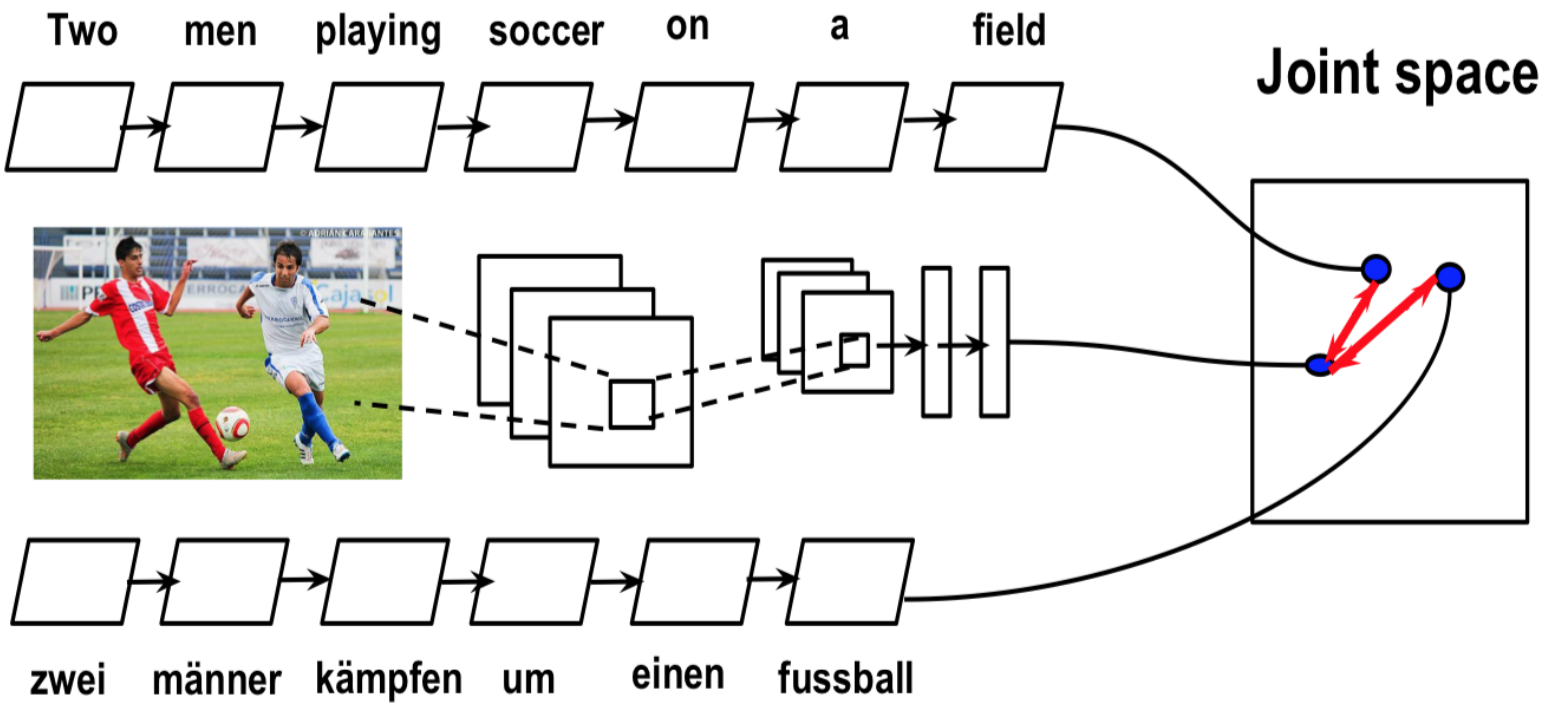
"Mulitmodal" datasets: Transform unimodal datasets into "multi-modal" problems by treating labels as a second modality

$z \sim p(z)$ $z \sim p(z|x_2 = 5)$ $z \sim p(z)$ $z \sim p(z|x_2 = ankle\ boot)$



[Wu, Mike, and Noah Goodman. "Multimodal Generative Models for Scalable Weakly-Supervised Learning.", NIPS 2018]

Representation 5: Multilingual Representations



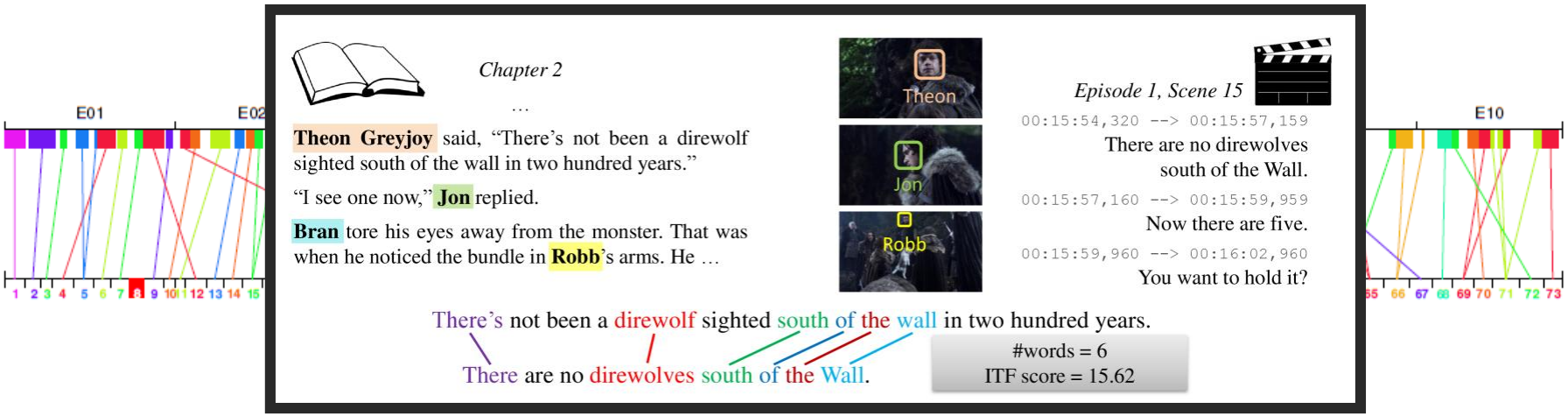
Goal: map image and its descriptions (not translations) in both languages close to each other.

[Gella et al. " Image Pivoting for Learning Multilingual Multimodal Representations", ACL 2017]

New(-ish) Directions: Alignment

Alignment 1: Books to scripts/movies

- Aligning very different modalities
- Books to scripts/movies

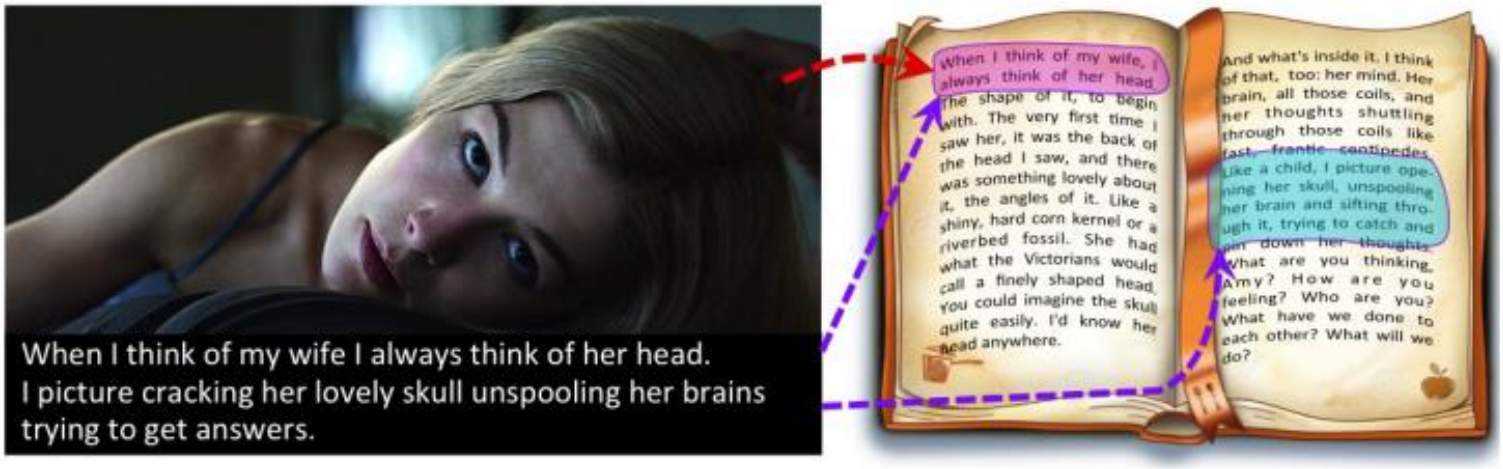


- Hand-crafted similarity based approach

[Tapaswi et al. Book2Movie: Aligning Video scenes with Book chapters, CVPR 2015]

Alignment 2: Books to scripts/movies

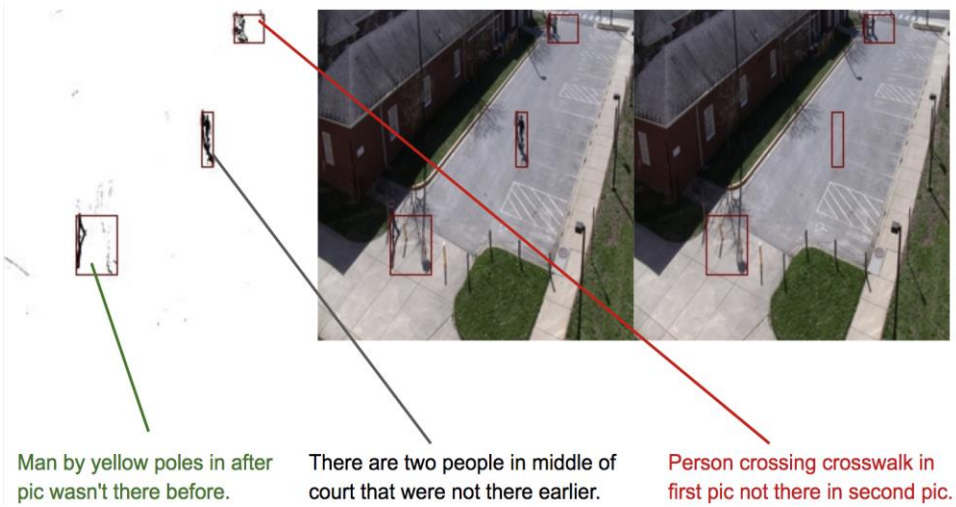
- Aligning very different modalities
- Books to scripts/movies

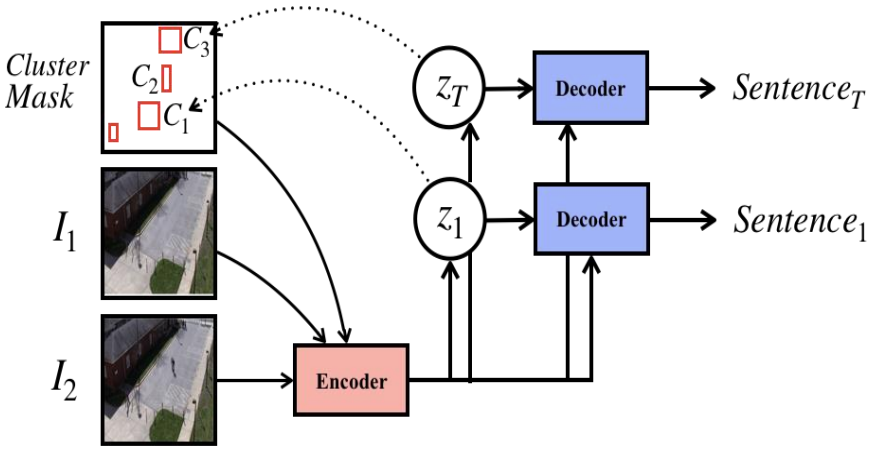


- Supervision based approach

[Zhu et al. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, ICCV 2015]

Alignment 3: Spot-The-Diff





- '*Spot-the-diff*': a new task and a dataset for succinctly describing all the differences between two similar images.
- Proposes a new model that captures visual salience through a latent alignment between clusters of differing pixels and output sentences.

[Jhamtani and Berg-Kirkpatrick. Learning to Describe Differences Between Pairs of Similar Images., EMNLP 2018]

Alignment 4: Textual Grounding

A woman in a green shirt is getting ready to throw her bowling ball down the lane...

Two women wearing hats covered in flowers are posing.

Young man wearing a hooded jacket sitting on snow in front of mountain area.

second bike from right in front

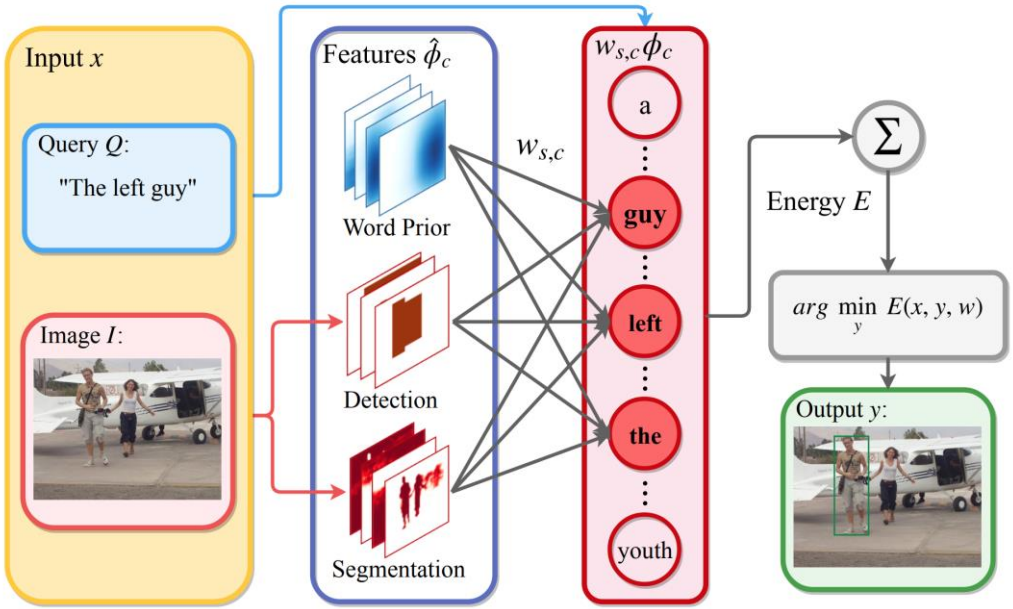
painting next to the two on the left

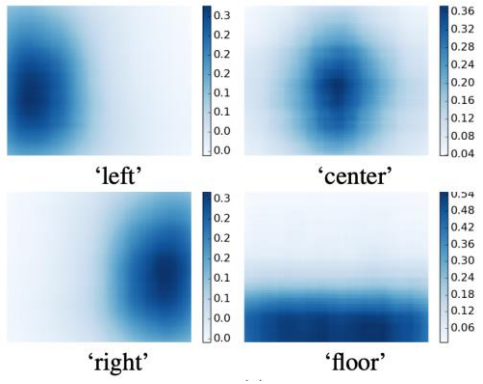
person all the way to the right

[Yeh, Raymond, et al. “Interpretable and globally optimal prediction for textual grounding using image concepts.”, NIPS 2017.]

Alignment 4: Textual Grounding

- Formulate the bounding box prediction as an energy minimization
  - The energy function is defined as a linear combination of a set of "image concepts" $\phi_c(x, w_r) \in \mathbb{R}^{W \times H}$

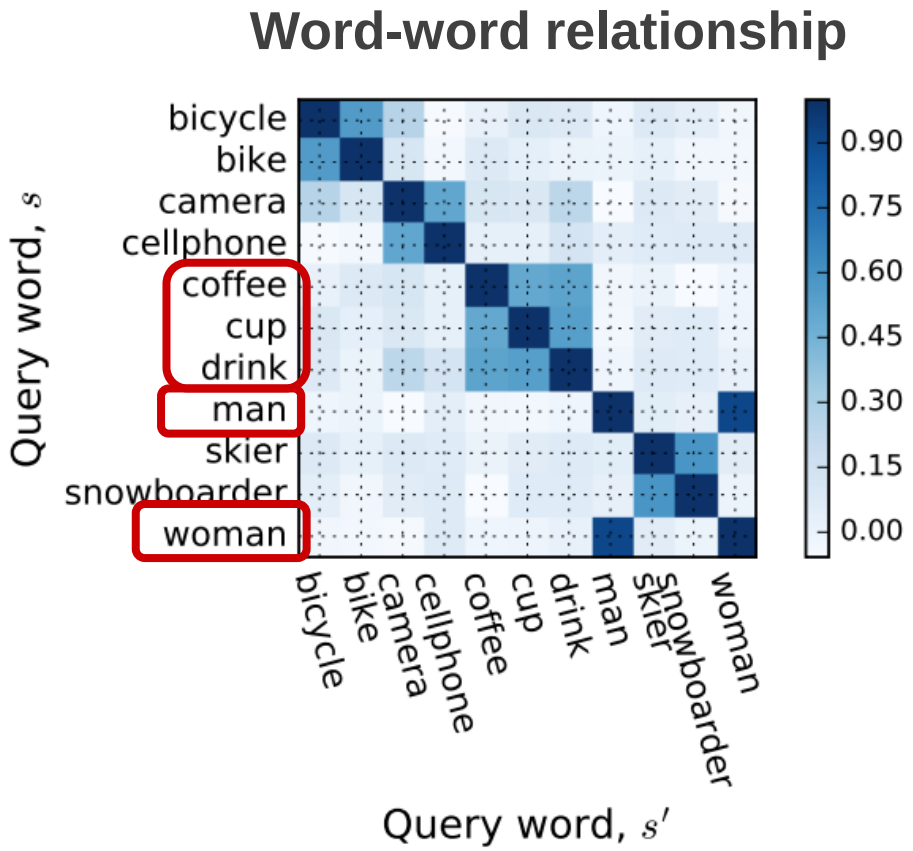




Word priors

[Yeh, Raymond, et al. "Interpretable and globally optimal prediction for textual grounding using image concepts.", NIPS 2017.]

Alignment 4: Textual Grounding



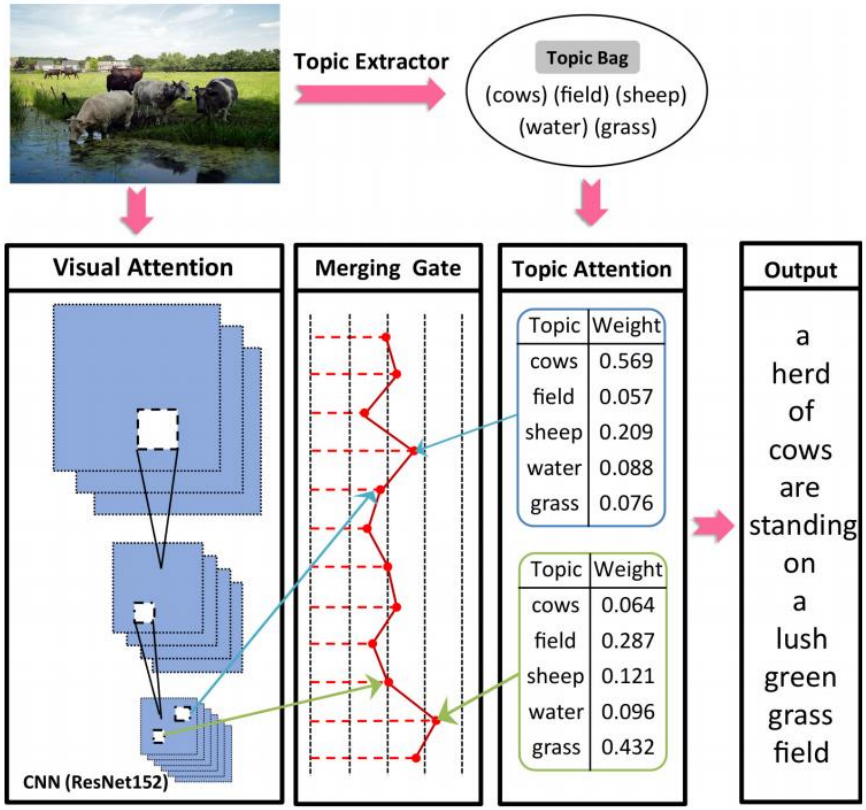
$$\cos(w_s, w_{s'})$$

$$w_s = [w_{s,1}; w_{s,1}; \ldots; w_{s,|C|}]$$

[Yeh, Raymond, et al. “Interpretable and globally optimal prediction for textual grounding using image concepts.”, NIPS 2017.]

Alignment 5: Comprehensive Image Captions

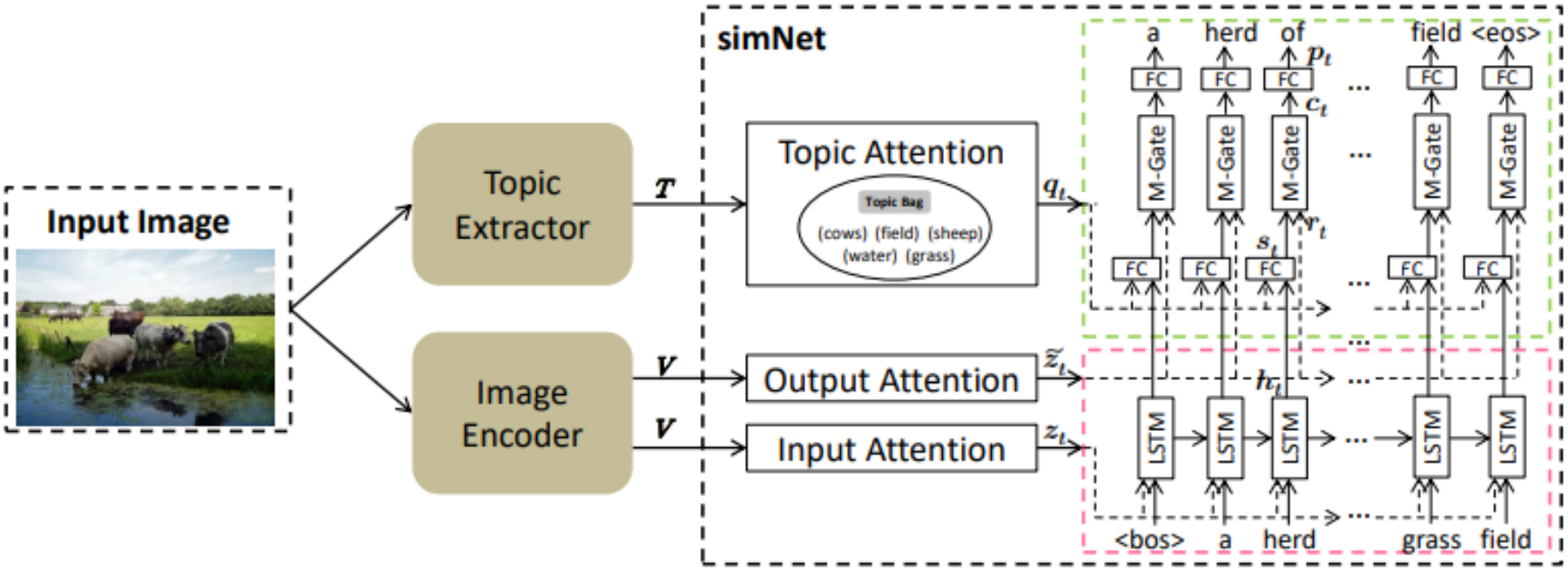
- Merging attention from text and visual modality for image captioning
- Strike a balance between details (visual driven) and coverage of objects (text/topic driven)



[Liu et al. simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions, 2018]

Alignment 5: Comprehensive Image Captions

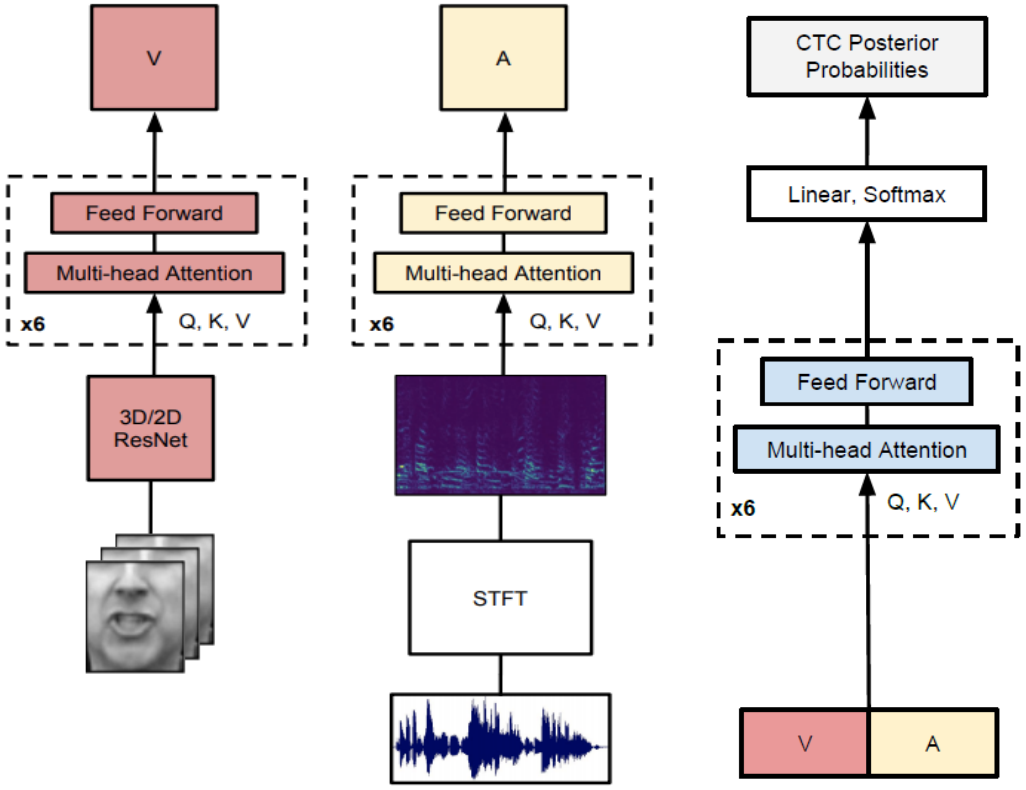
- Merging attention from text and visual modality for image captioning



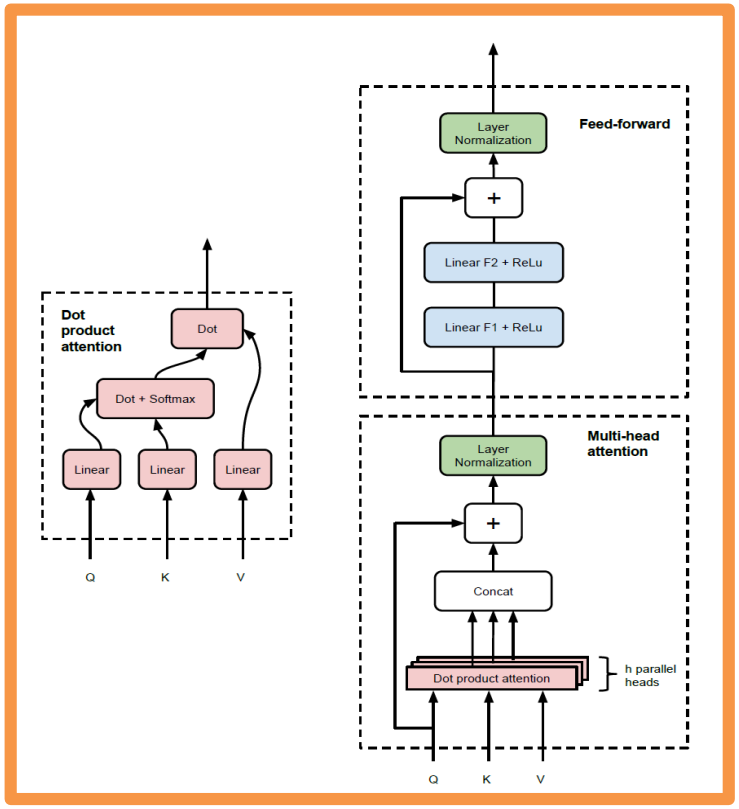
[Liu et al. simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions, 2018]

New(-ish) Directions: Fusion

Fusion 1a: Multi-Head Attention for AVSR



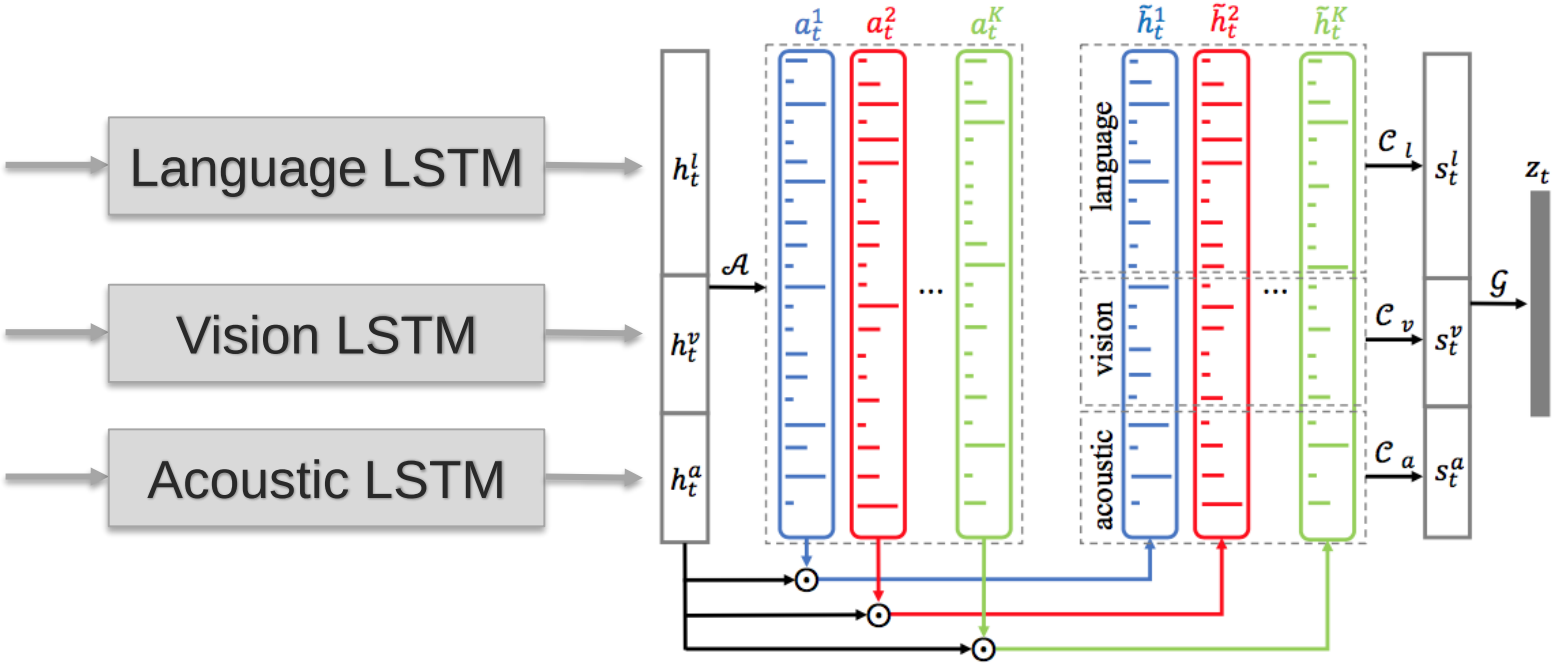
Multi-head Attention



Afouras, Triantafyllos, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Deep audio-visual speech recognition." *arXiv preprint arXiv:1809.02108* (Sept 2018).

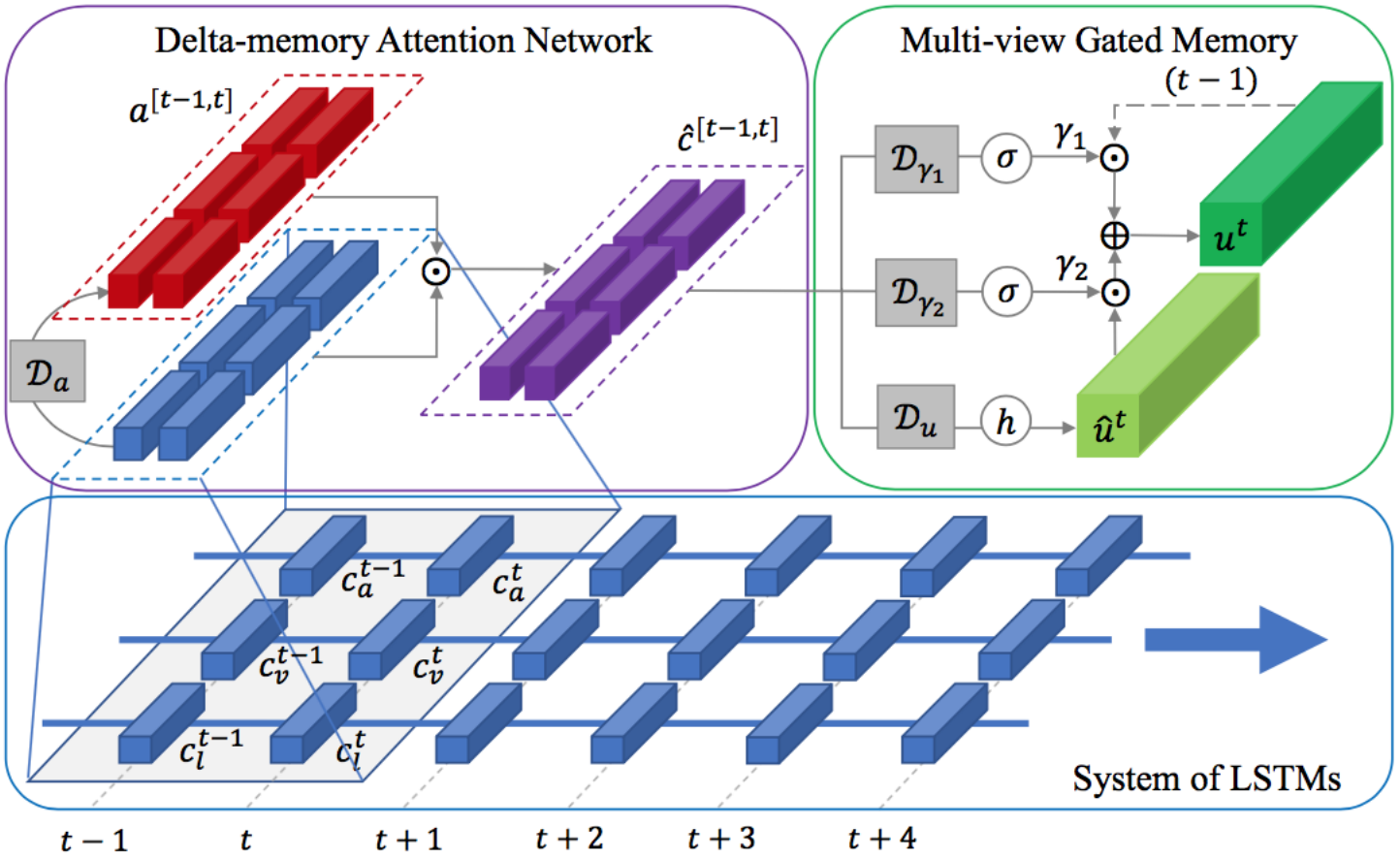
Fusion 1b: Fusion with Multiple Attentions

- Modeling Human Communication – Sentiment, Emotions, Speaker Traits



[Zadeh et al., Human Communication Decoder Network for Human Communication Comprehension, AAAI 2018]

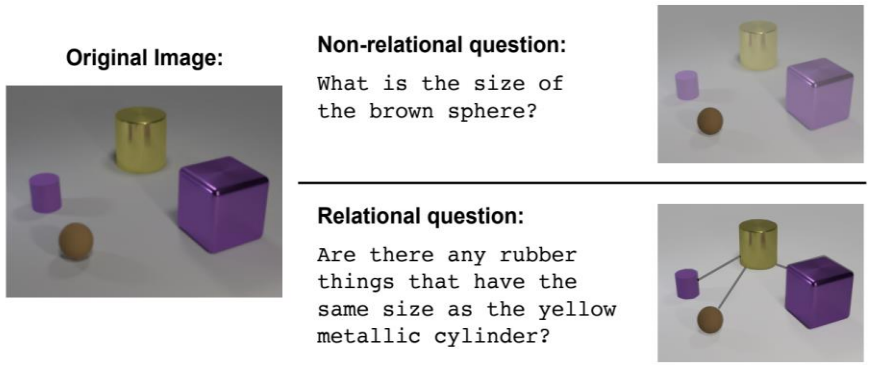
Fusion 2: Memory-Based Fusion



[Zadeh et al., Memory Fusion Network for Multi-view Sequential Learning, AAAI 2018]

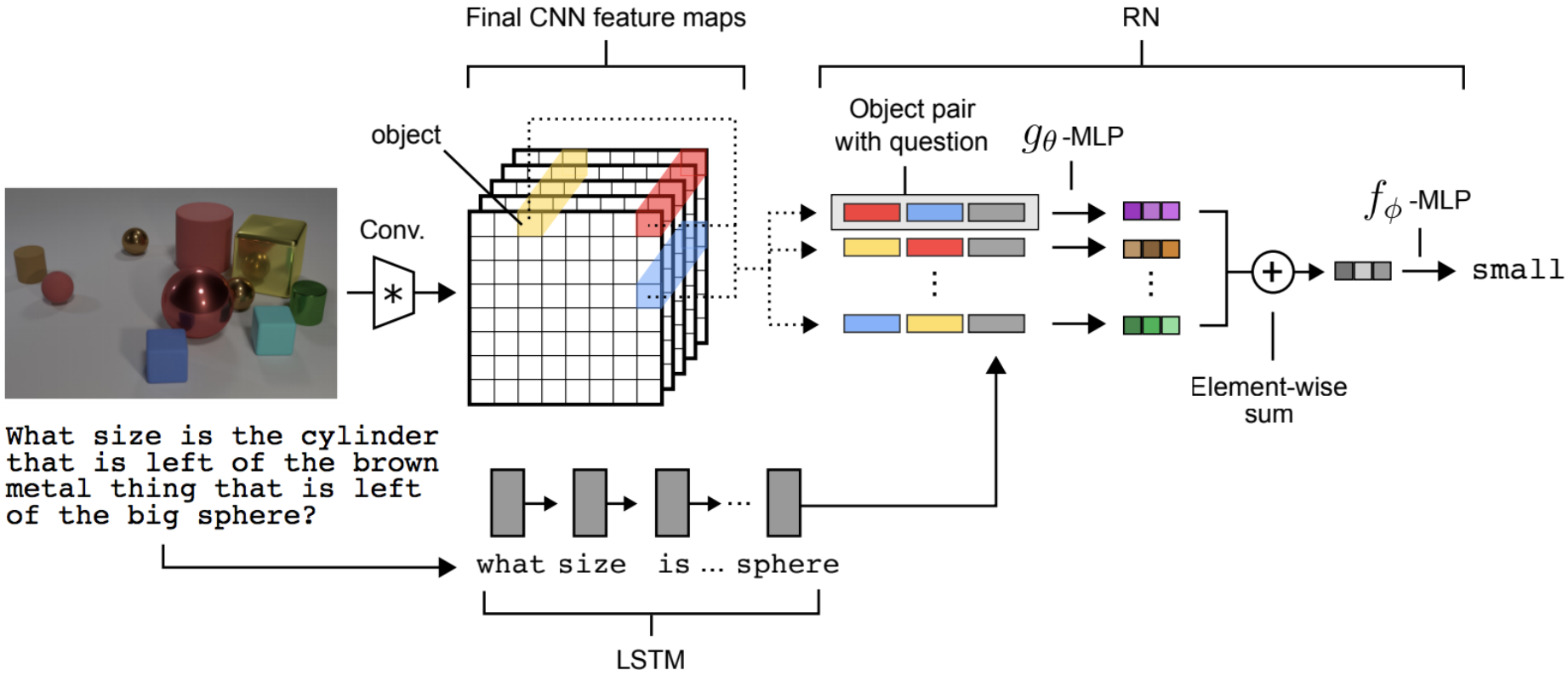
Fusion 3: Relational Questions

- Aims to improve relational reasoning for Visual Question Answering
- Current deep learning architectures – unable to capture reasoning capabilities on their own
- Proposes a Relation Network (RN) that augments CNNs for better reasoning



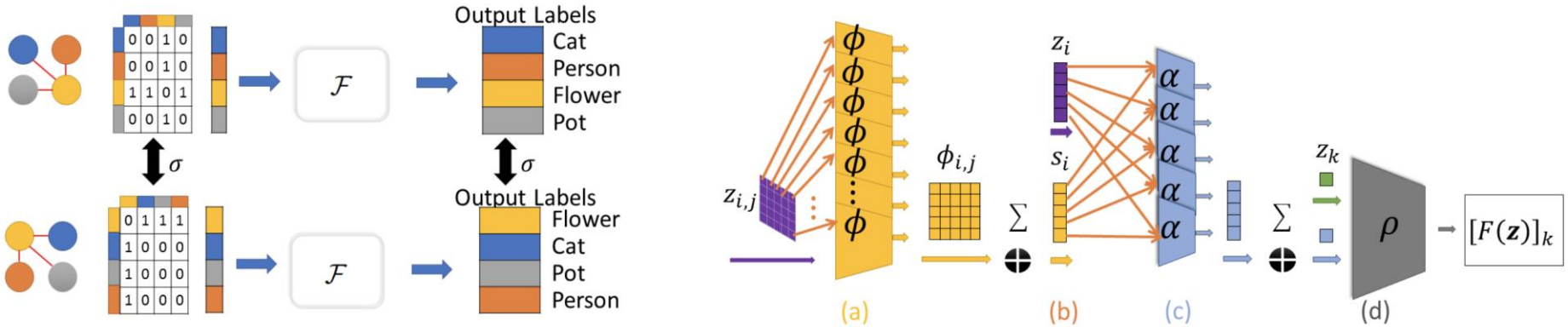
Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in neural information processing systems* (pp. 4967-4976).

Fusion 3: Relational Questions



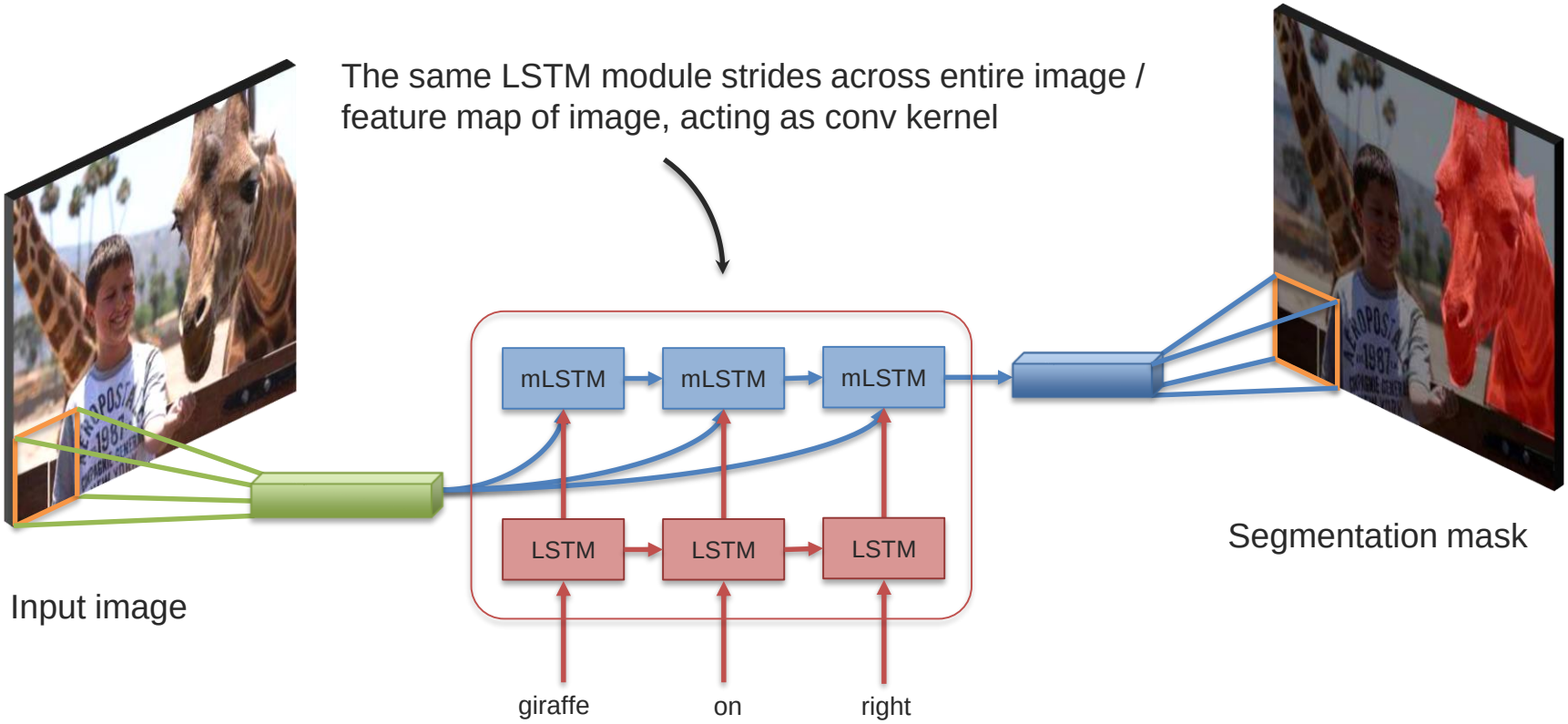
Fusion 4: Structured Prediction

- Scene-graph prediction: The output structure is invariant to specific permutations.
- The paper describe a model that satisfies the permutation invariance property, and achieve state-of-the-art results on the competitive Visual Genome benchmark



[Herzig et al. Mapping Images to Scene Graphs with Permutation-Invariant Structured Prediction, NIPS 2018]

Fusion 5: Recurrent Multimodal Interaction

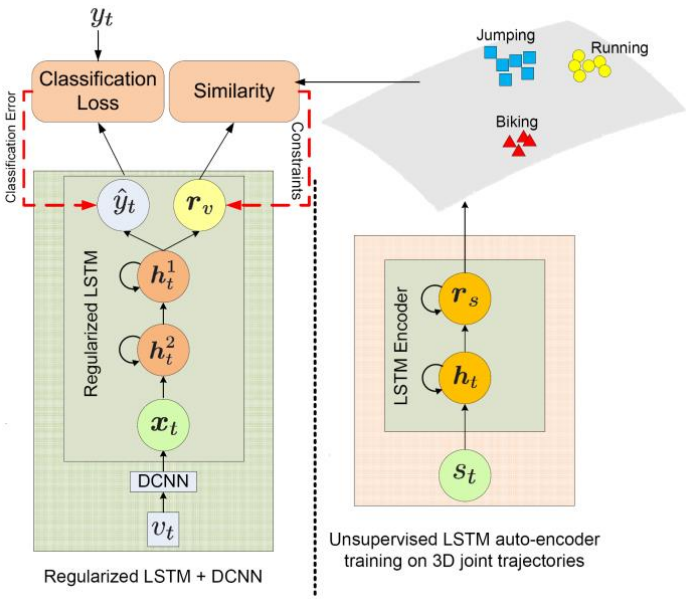


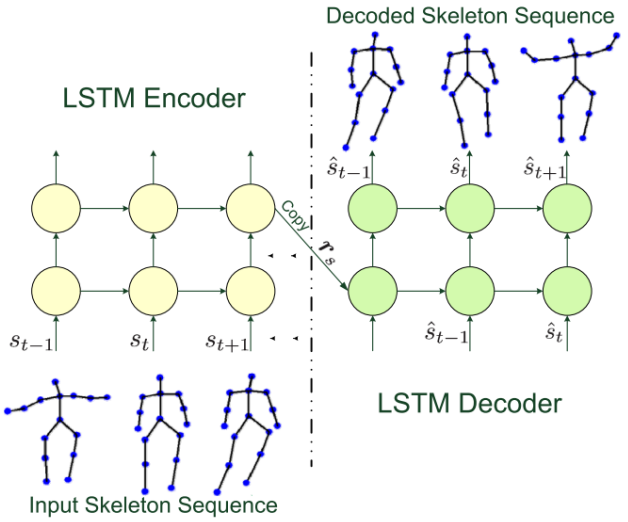
[Liu et al. Recurrent Multimodal Interaction for Referring Image Segmentation, 2017]

New(ish) Directions: Co-Learning

Co-learning 1: Regularizing with Skeleton Seqs

- Better unimodal representation by regularizing using a different modality

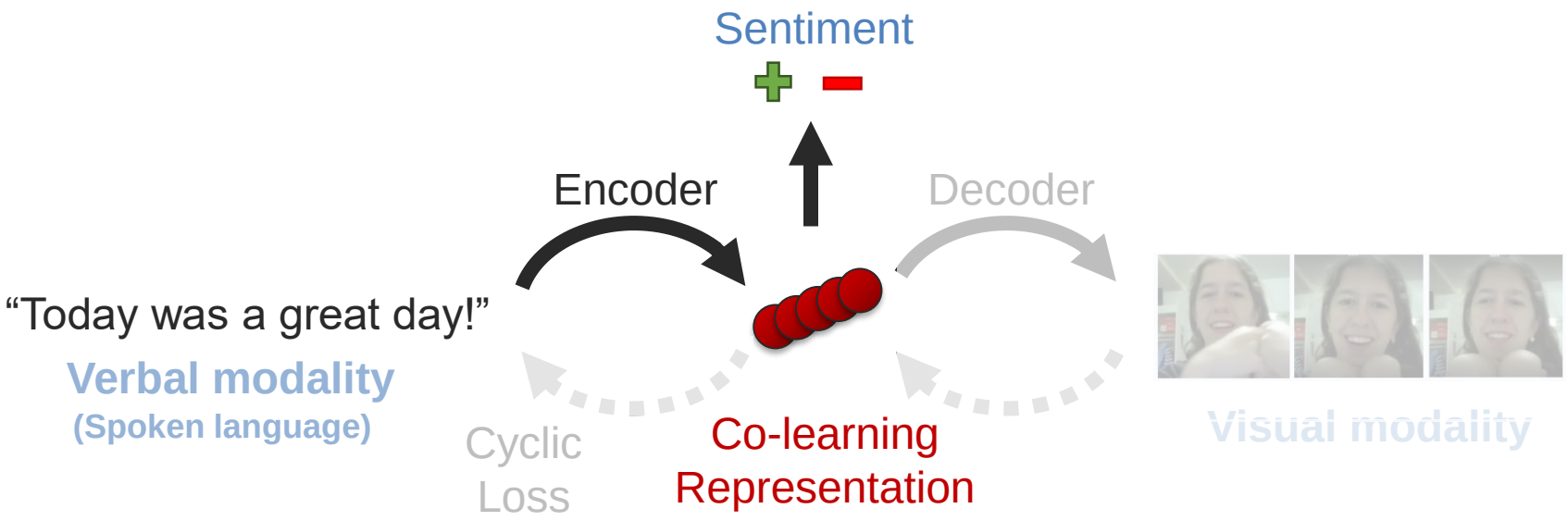




Non parallel data!

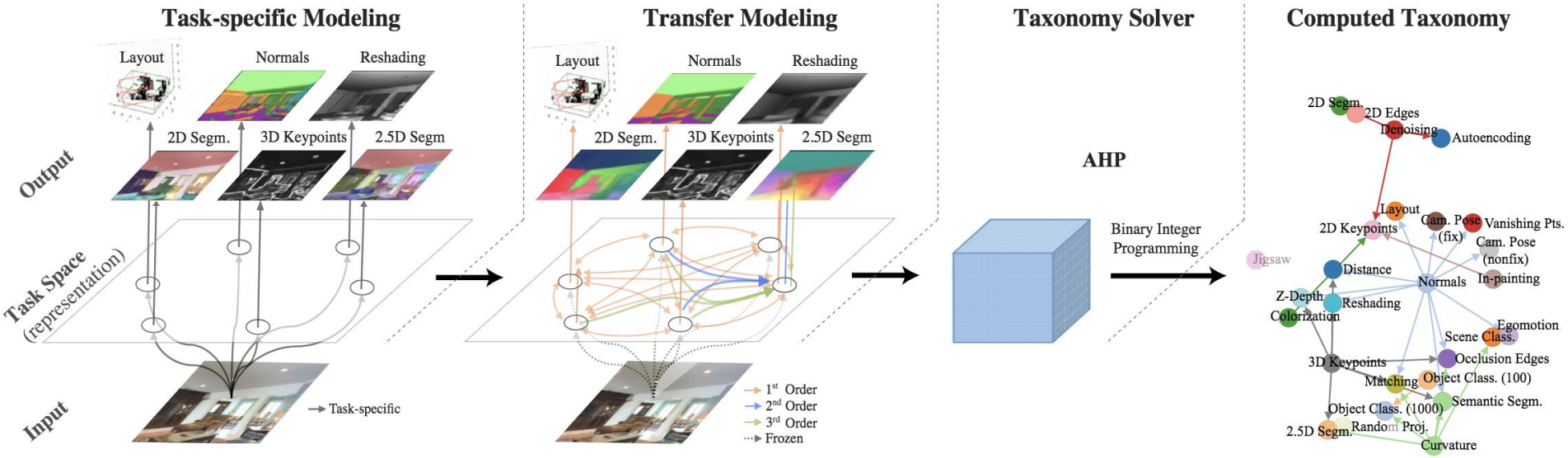
[B. Mahasseni and S. Todorovic, "Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition," in CVPR, 2016]

Co-Learning 2: Multimodal Cyclic Translation



Paul Pu Liang*, Hai Pham*, et al., "Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities", AAAI 2019

Co-learning 3: Taskonomy

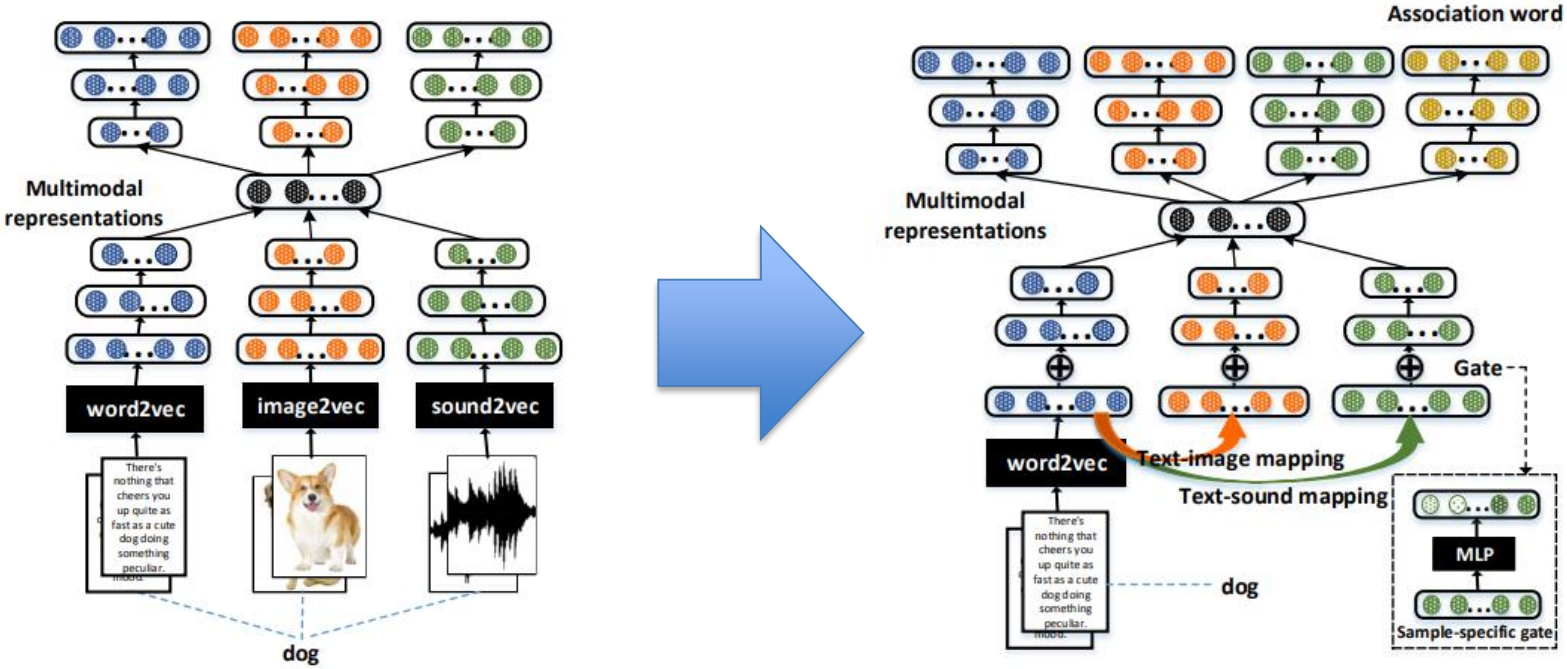


Process overview. The steps involved in creating the taxonomy.

Zamir, Amir R., et al. "Taskonomy: Disentangling Task Transfer Learning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Co-learning 4: Associative Multichannel Autoencoder

- Learning representation through fusion and translation
- Use associated word prediction to address data sparsity



[Wang et al. Associative Multichannel Autoencoder for Multimodal Word Representation, 2018]

Co-learning 5: Grounding Semantics in Olfactory Perception

- Grounding language in vision, sound, and **smell**

Olfactory-Relevant Examples					
MEN		sim	SimLex-999		sim
bakery	bread	0.96	steak	meat	0.75
grass	lawn	0.96	flower	violet	0.70
dog	terrier	0.90	tree	maple	0.55
bacon	meat	0.88	grass	moss	0.50
oak	wood	0.84	beach	sea	0.47
daisy	violet	0.76	cereal	wheat	0.38
daffodil	rose	0.74	bread	flour	0.33

	Chemical Compound				
Smell label	Phenethyl acetate	Isoamyl butyrate	Anisyl butyrate	Myrcene	Syringaldehyde
Melon	✓	✓			
Pineapple	✓				✓
Licorice			✓		
Anise			✓	✓	
Beer				✓	✓

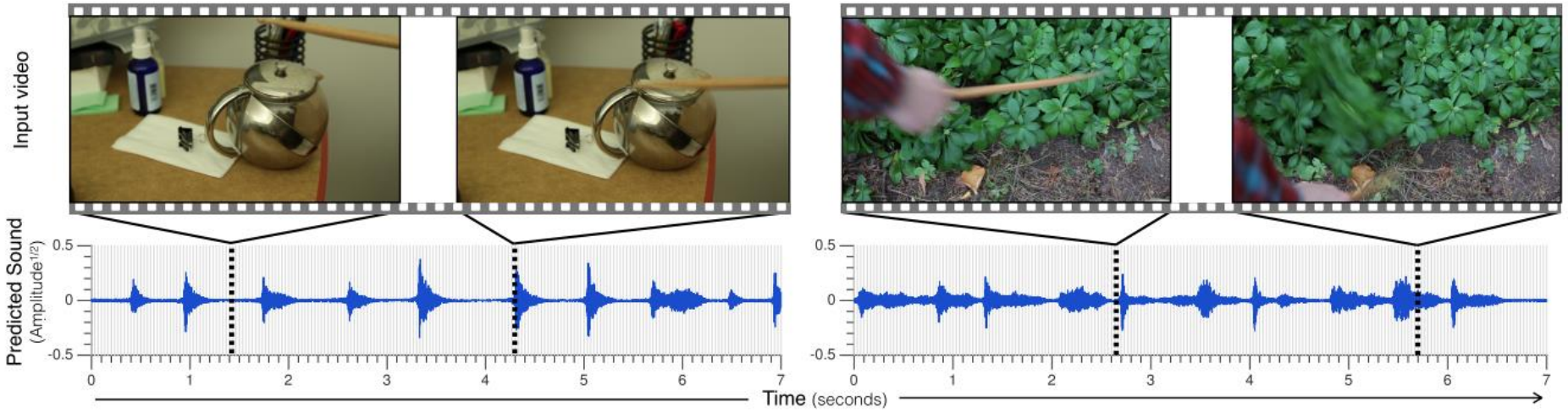
Table 2: A BoCC model.

[Kiela et al., Grounding Semantics in Olfactory Perception, ACL-IJCNLP, 2015]

New(-ish) Directions: Translation

Translation 1: Visually indicated sounds

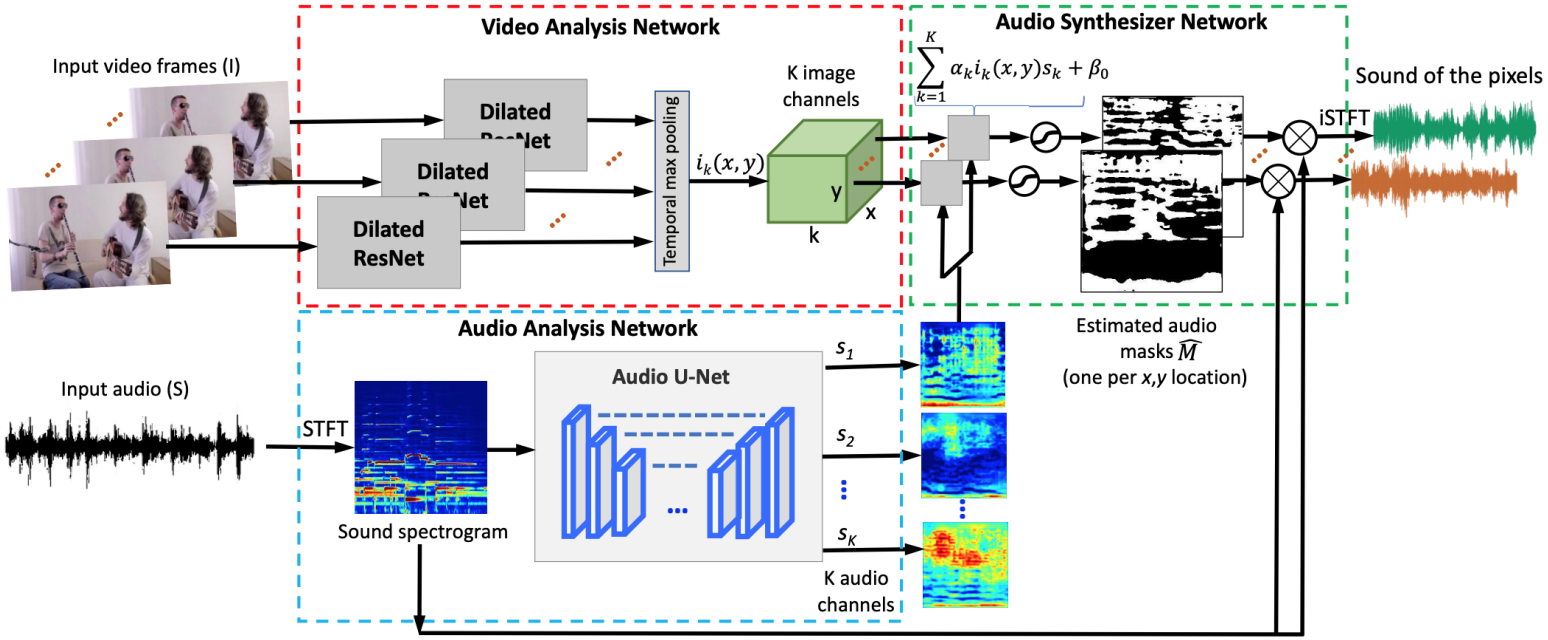
- Sound generation!



[Owens et al. Visually indicated sounds, CVPR, 2016]

Translation 2: The Sound of Pixels

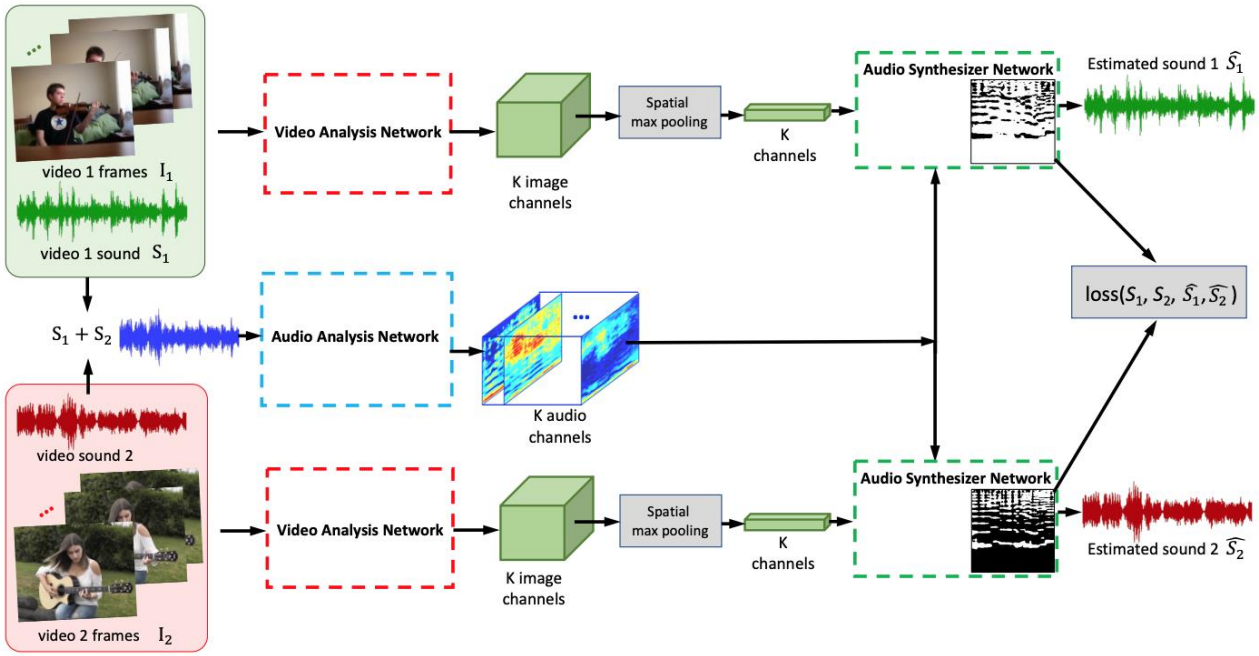
- Propose a system that learns to localize the sound sources in a video and separate the input audio into a set of components coming from each object by leveraging unlabeled videos.



[Zhao, Hang, et al. "The sound of pixels.", ECCV 2018] https://youtu.be/2eVDLEQlKD0

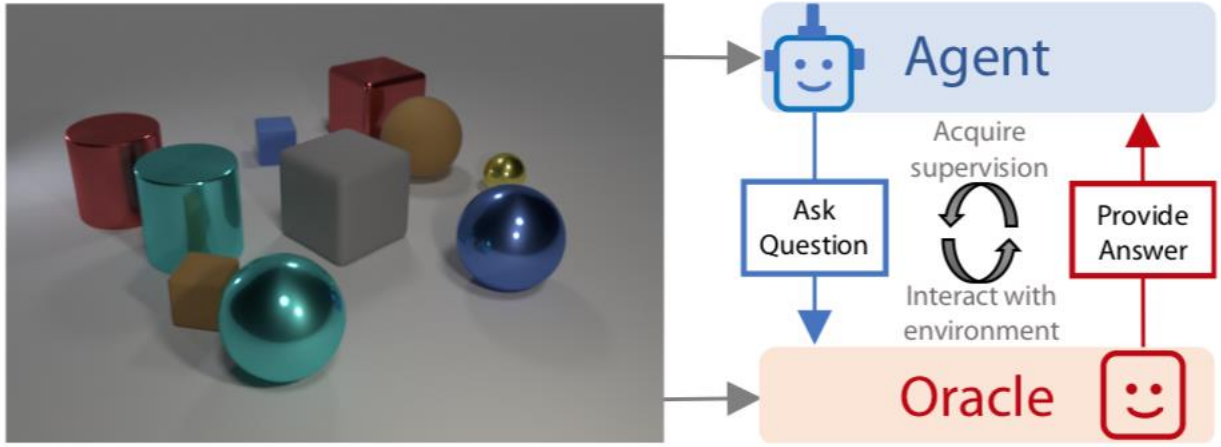
Translation 2: The Sound of Pixels

- Trained in a **self-supervised** manner by learning to separate the sound source of a video from the audio mixture of multiple videos conditioned on the visual input associated with it.



[Zhao, Hang, et al. “The sound of pixels.”, ECCV 2018]

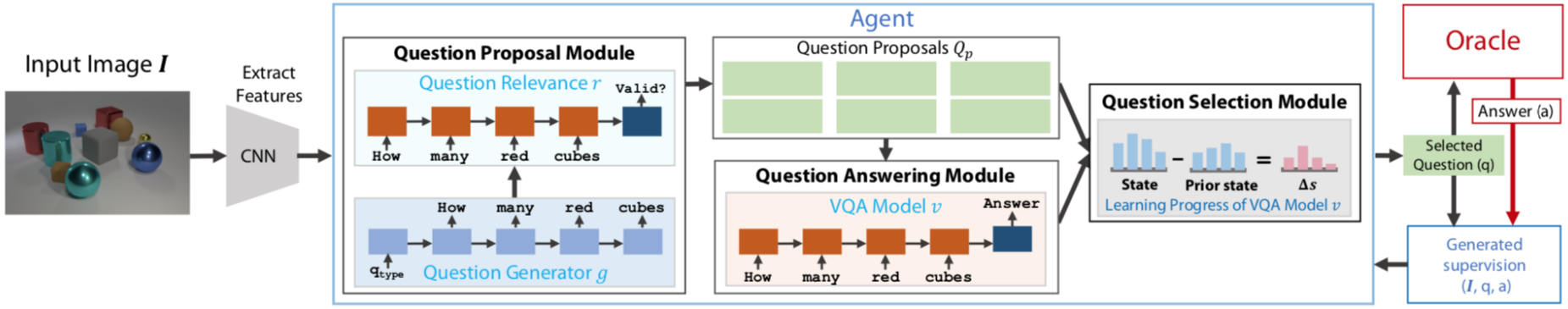
Translation 3: Learning-by-asking (LBA)



- an **agent** interactively learns by asking questions to an **oracle**
- standard VQA training has a fixed dataset of questions
- in LBA the agent has the potential to learn more quickly by asking "good" questions (like a bright student in a class)

[Misra et al. "Learning by Asking Questions", CVPR 2018]

Translation 3: Learning-by-asking (LBA)



Training:

- Given on the input image the model decides what questions to ask
- Answers are obtained by human supervised oracle

Testing:

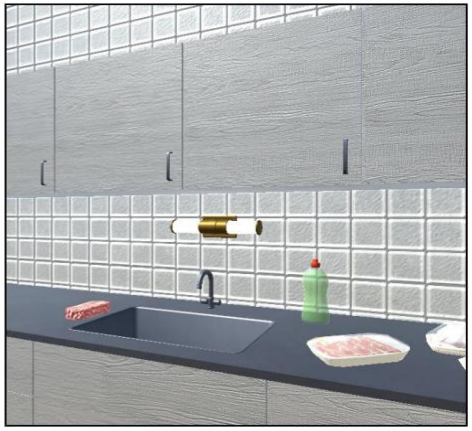
- LBA is evaluated exactly like VQA

[Misra et al. "Learning by Asking Questions", CVPR 2018]

Translation 4: Navigation

- Goal prediction
  - Highlight the goal location by generating a probability distribution over the environment panoramic image
- Interpretability
  - Explicit goal prediction modeling makes the approach more interpretable

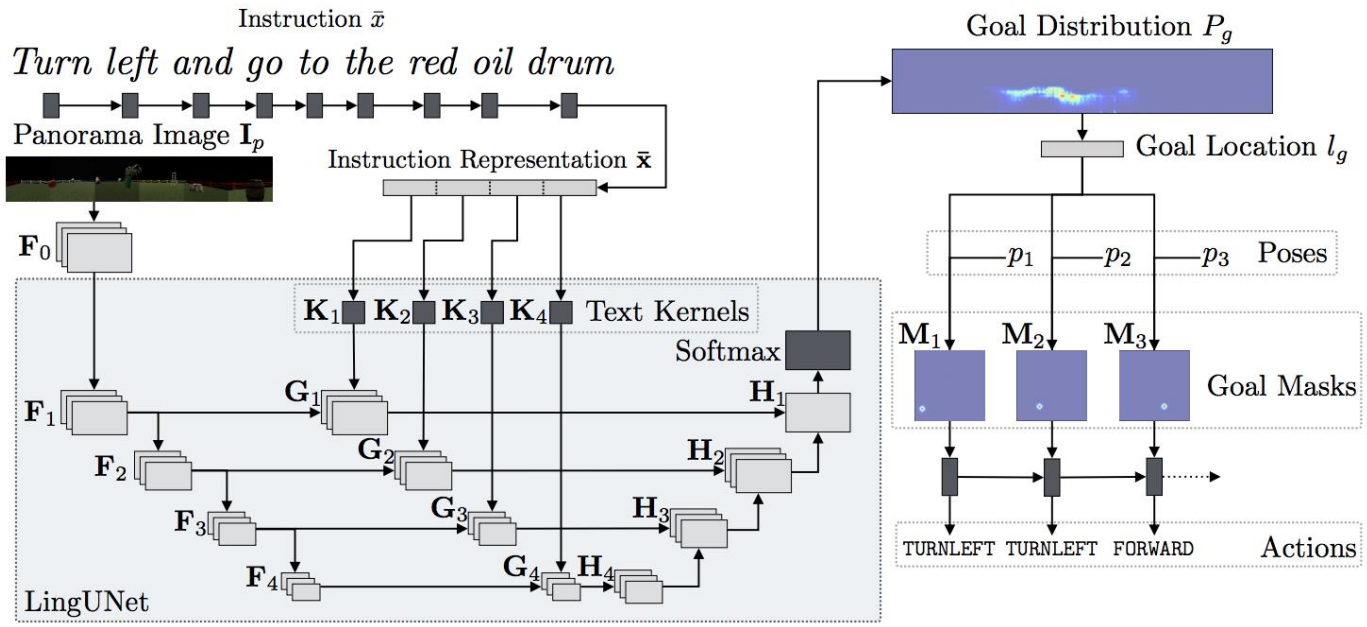
After reaching the hydrant head towards the blue fence and pass towards the right side of the well.



Put the cereal, the sponge, and the dishwashing soap into the cupboard above the sink.

[Misra et al. Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction, EMNLP 2018]

Translation 4: Navigation

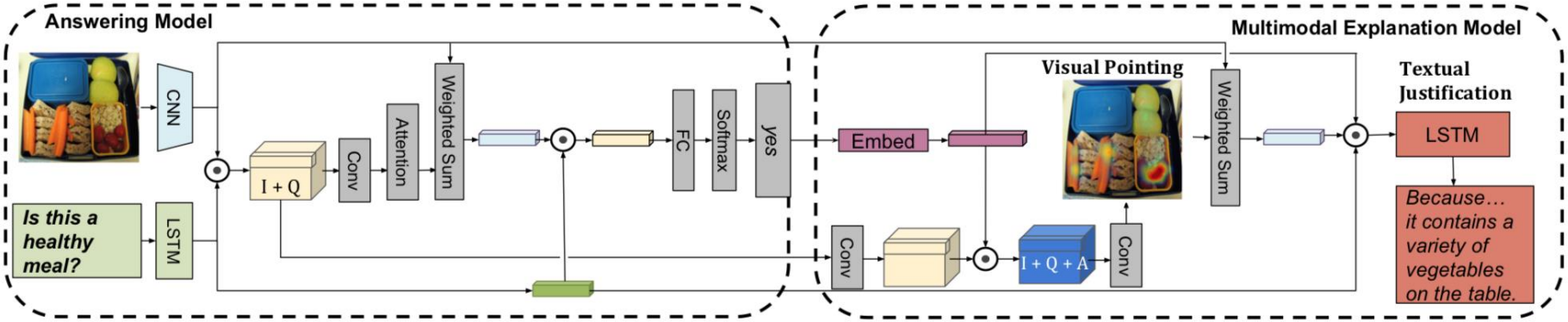


- The paper proposes to decompose instruction execution into: 1) goal prediction and 2) action generation

[Misra et al. Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction, EMNLP 2018]

Translation 5: Explanations for VQA and ACT

Pointing and Justification Architecture

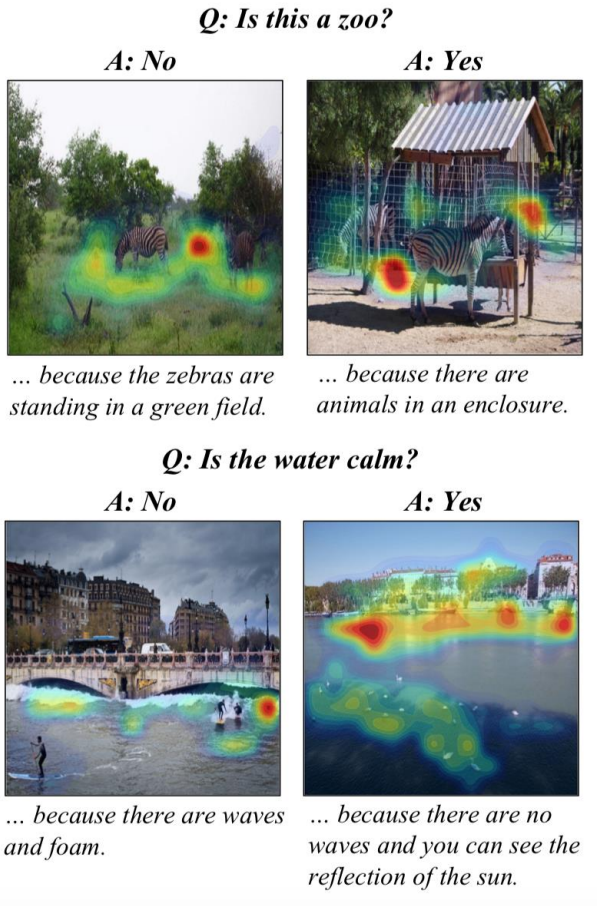


- **Answering Model**: predicts an answer given the image and the question
- **Multimodal Explanation Model**: generates visual and textual explanations given the answer, question, and image

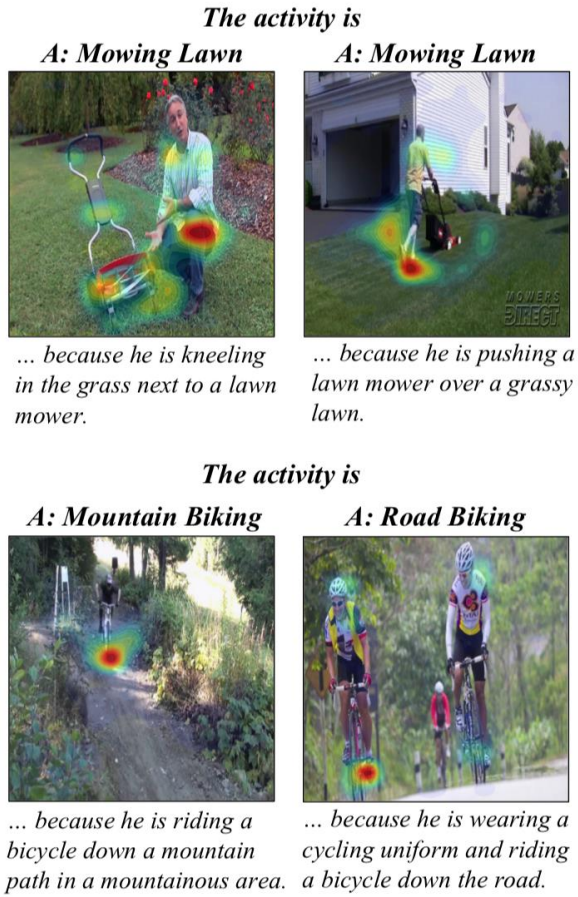
[Park et al. "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence", CVPR 2018]

Translation 5: Explanations for VQA and ACT

VQA-X:



ACT-X:



[Park et al. "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence", CVPR 2018]