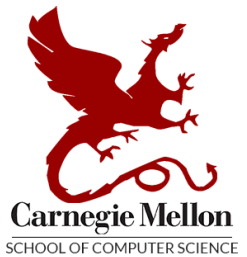
What is Multimodal Machine Learning?

Jonathan Francis
Research Scientist, Bosch Research Pittsburgh (CR/RTC3.1-NA)
PhD Candidate, School of Computer Science, Carnegie Mellon University



Multimodal Machine Learning
Definition

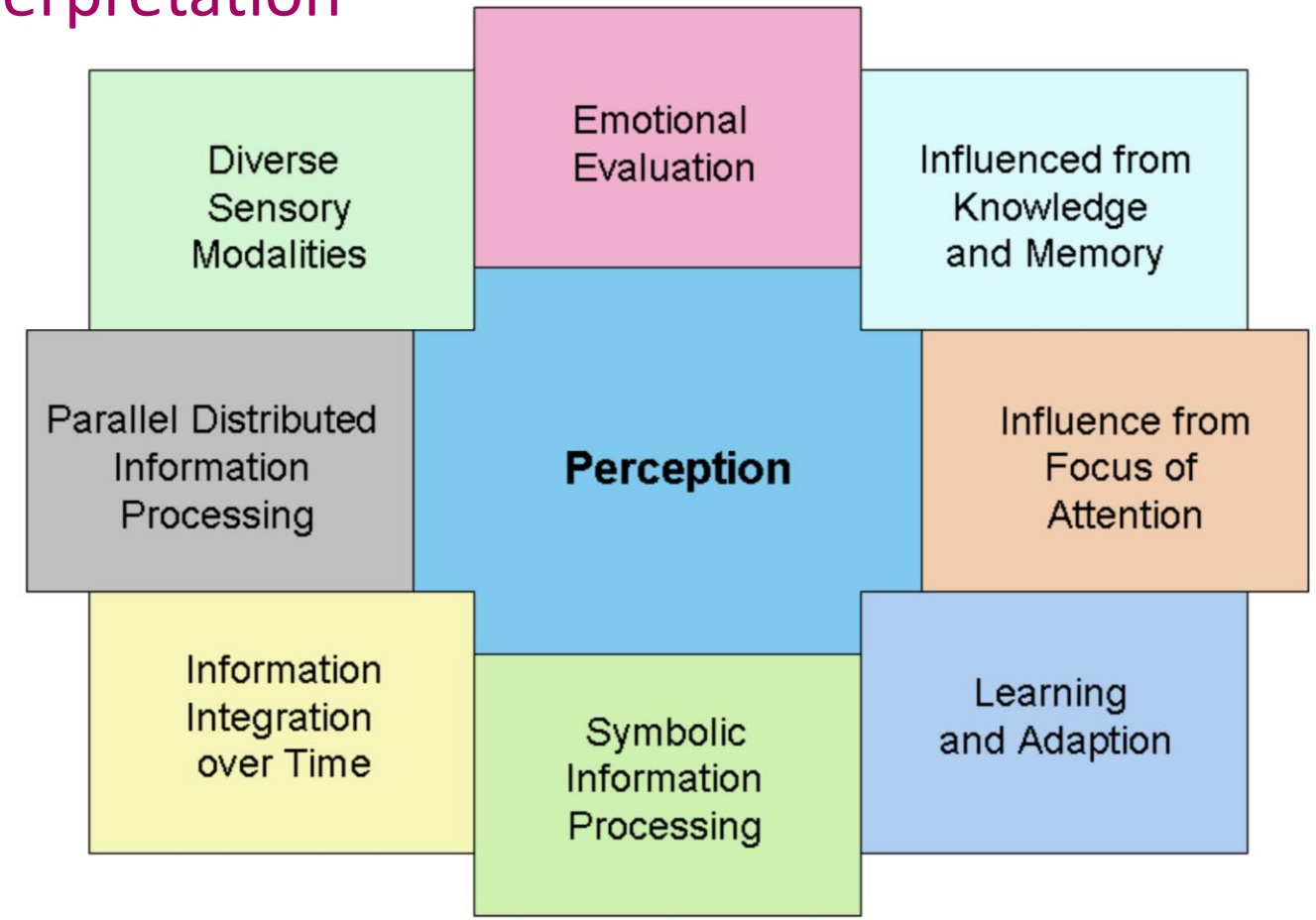
Modality

The way in which something happens or is experienced

- Modality refers to a certain type of information and/or the representation format in which information is stored
- *Sensory modality:* one of the primary forms of sensation, such as vision or touch; channel of communication

Multimodal Machine Learning

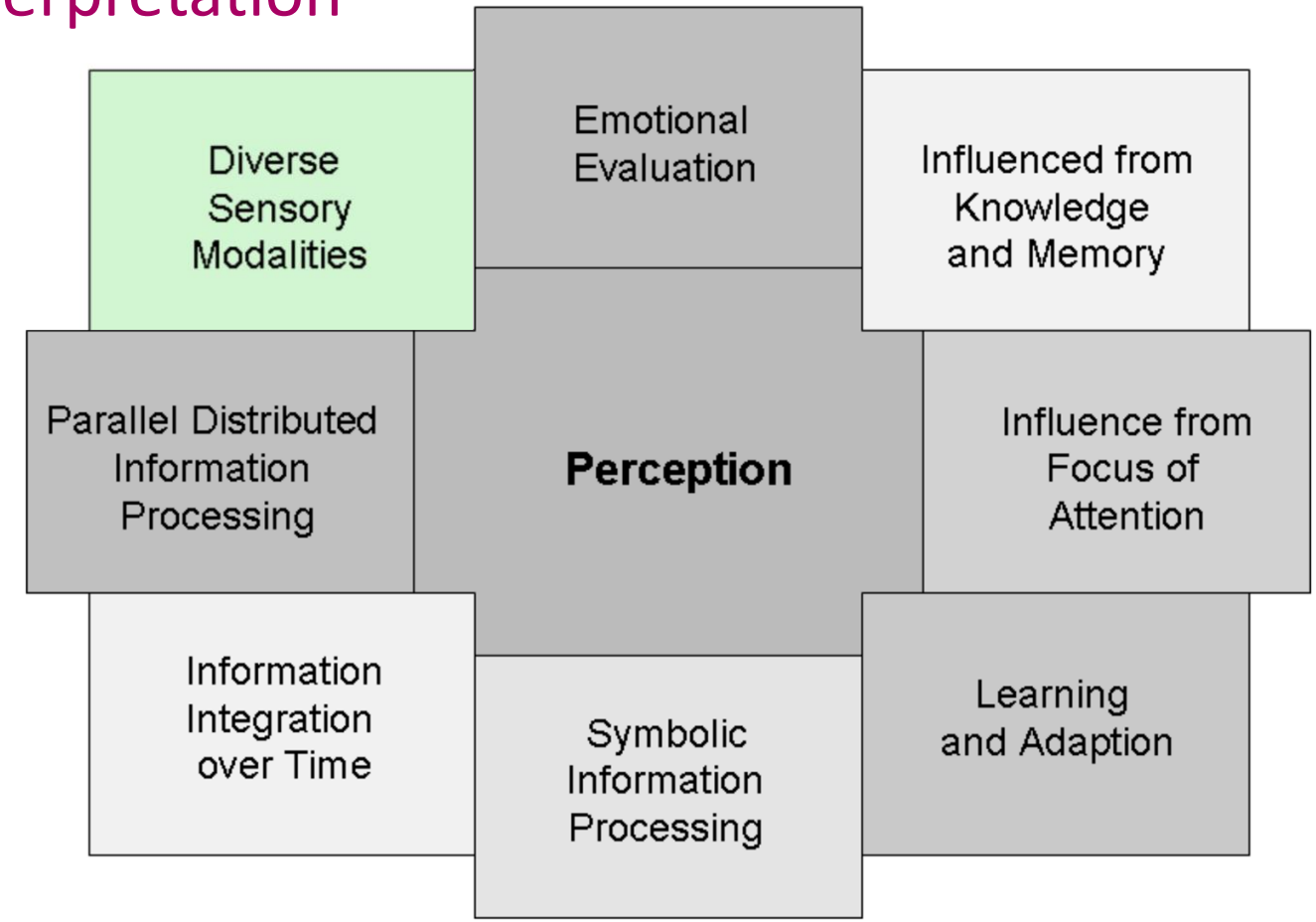
Perception Interpretation



[1] R. Velik, R. Lang, D. Bruckner and T. Deutsch, "Emulating the perceptual system of the brain for the purpose of sensor fusion," 2008 Conference on Human System Interactions, Krakow, 2008, pp. 657-662. doi: 10.1109/HSI.2008.4581518

Multimodal Machine Learning

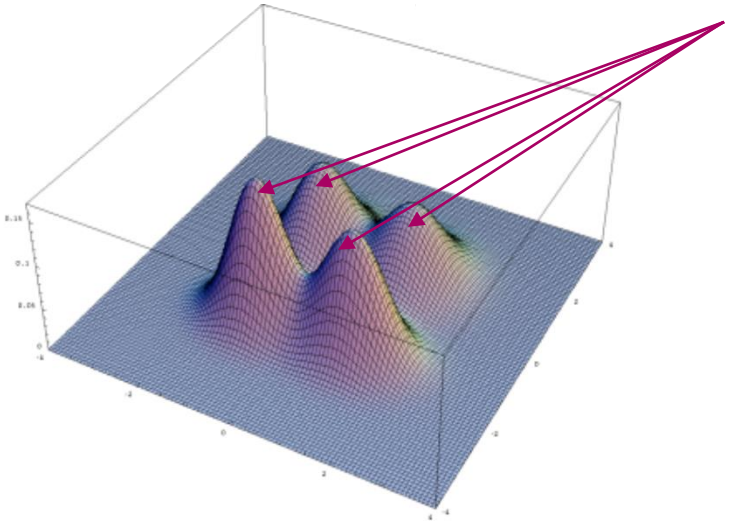
Perception Interpretation



[1] R. Velik, R. Lang, D. Bruckner and T. Deutsch, "Emulating the perceptual system of the brain for the purpose of sensor fusion," 2008 Conference on Human System Interactions, Krakow, 2008, pp. 657-662. doi: 10.1109/HSI.2008.4581518

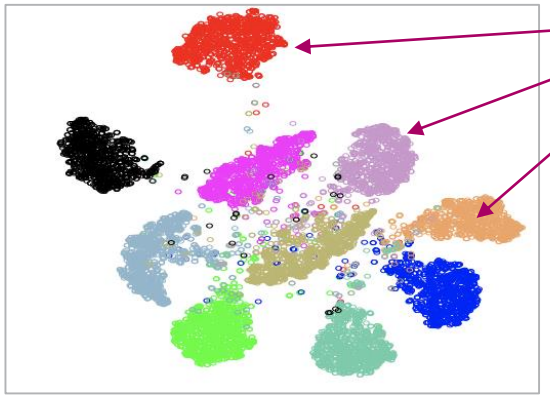
Multimodal Machine Learning
Distributional Interpretation

Given some event horizon, observations yield multiple modes, i.e., distinct ‘peaks’ (local maxima) in the probability density function



Data distribution with multimodal features

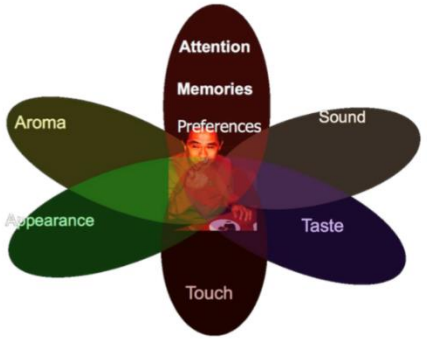
Given some embedding space, interdependencies between projected samples (possibly from different distributions) yield distinct ‘clusters’ or alignments



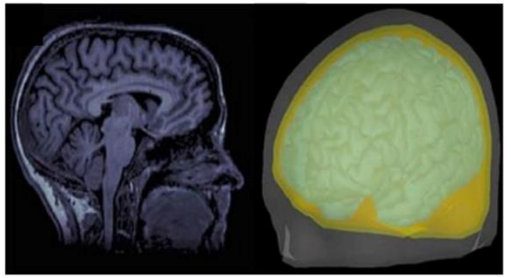
(Projected) data distribution higher-dimensional clustering behaviour

Multimodal Machine Learning

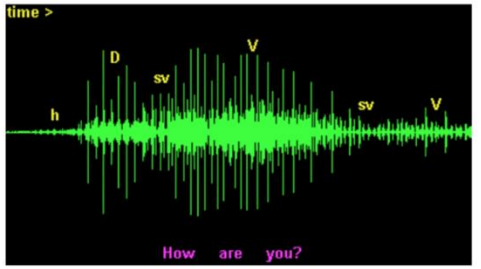
Research Areas



Psychology

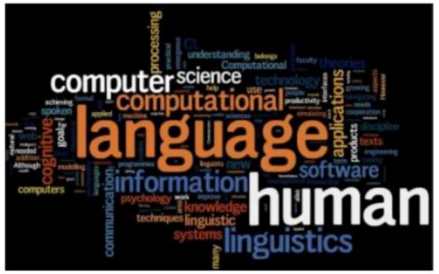


Medical



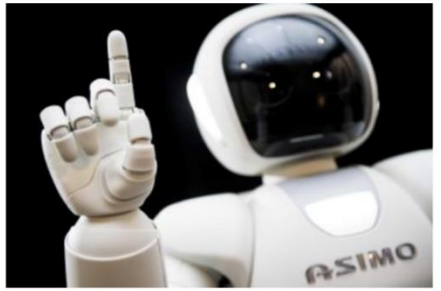
Speech

Vision

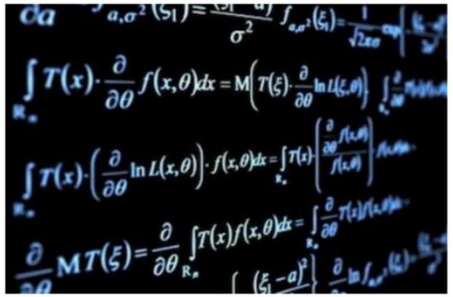


Language

Multimedia



Robotics



Learning

BOSCH

Multimodal Machine Learning

Taxonomy of Core Challenges

#1. Representation

- *Joint*
  - Neural networks
  - Graphical models
  - Sequential
- *Coordinated*
  - Similarity
  - Structured

#2. Alignment

- *Explicit*
  - Unsupervised
  - Supervised
- *Implicit*
  - Graphical models
  - Neural Networks

#3. Fusion

- *Model Agnostic*
  - Early
  - Late
  - Hybrid
- *Model-based*
  - *Kernel*
  - *Graphical Model*
  - *Neural Networks*

#4. Translation

- *Example-based*
  - Retrieval
  - Combination
- *Model-based*
  - Grammar
  - Encode-decoder
  - Online prediction

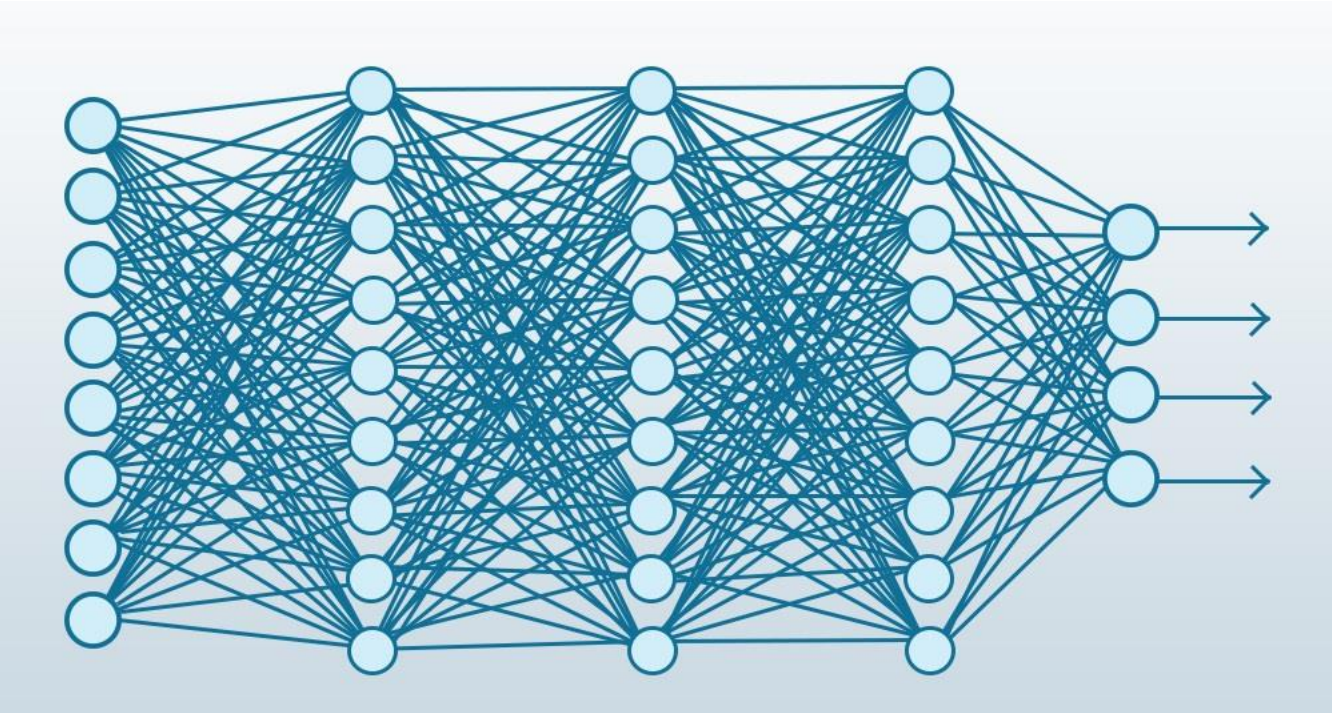
#5. Co-Learning

- *Parallel data*
  - Co-training
  - Transfer learning
- *Non-parallel data*
  - Zero-shot
  - Grounding
- *Hybrid Data*
  - *Bridging*

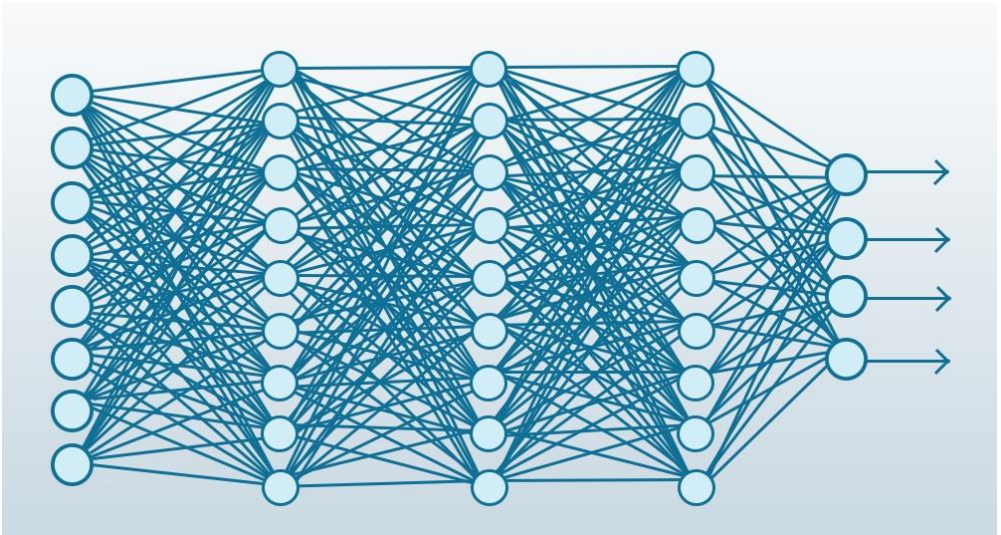
[2] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," arXiv:1705.09406, 2017.

Multimodal Machine Learning

Deep Learning



Multimodal Machine Learning
Deep Learning



Given:

$\varphi : \mathbb{R} \to \mathbb{R}$:= continuous (bounded, non-constant)

I_m := m-dimensional hypercube, $[0,1]^m$ (convexity)

$C(I_m)$:= space of continuous functions on I_m

$\varepsilon > 0$:= arbitrary

$f \in C(I_m)$:= arbitrary sampled function

There exists:

Integer N, real constants v_i, b_i, and real vec$w_i \in \mathbb{R}^m$

Such that we may define:

$$F(x) = \sum_{i=1}^{N} v_i \varphi\left(w_i^T x + b_i\right)$$:= functional approximator

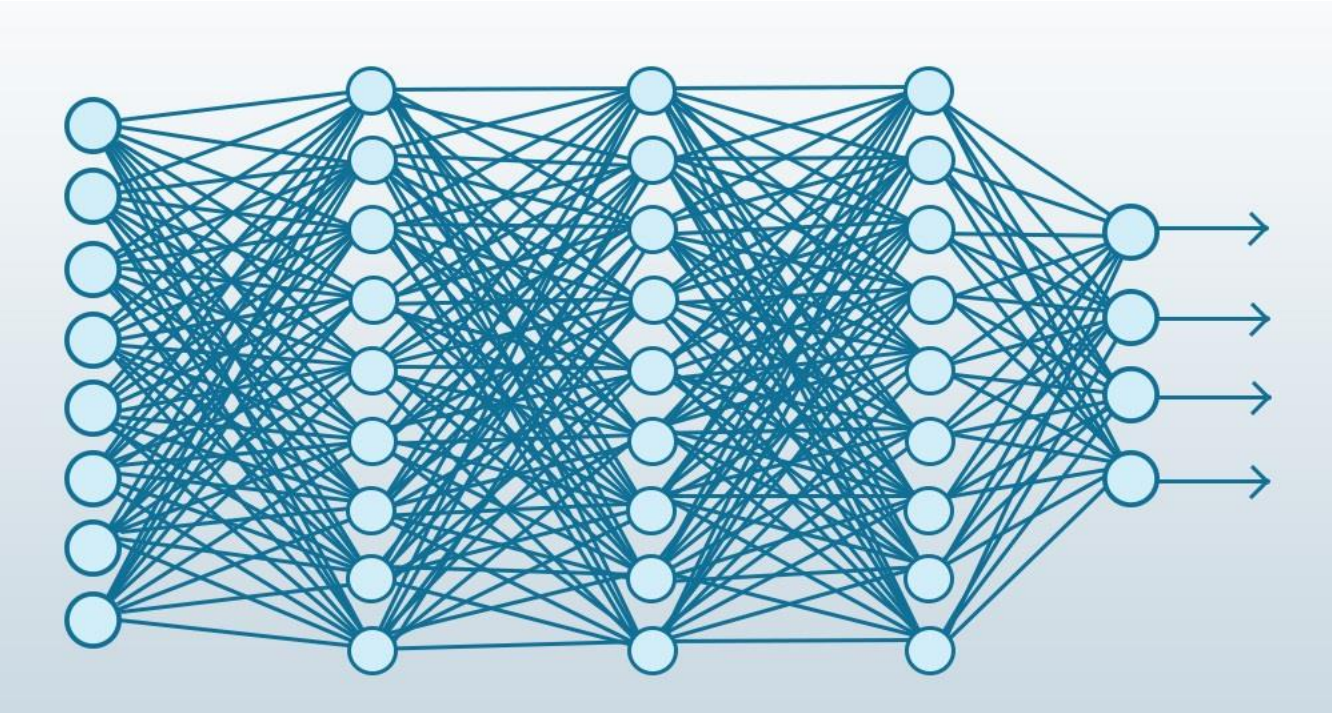
(activation, bias, weight)

$$|F(x) - f(x)| < \varepsilon$$

BOSCH

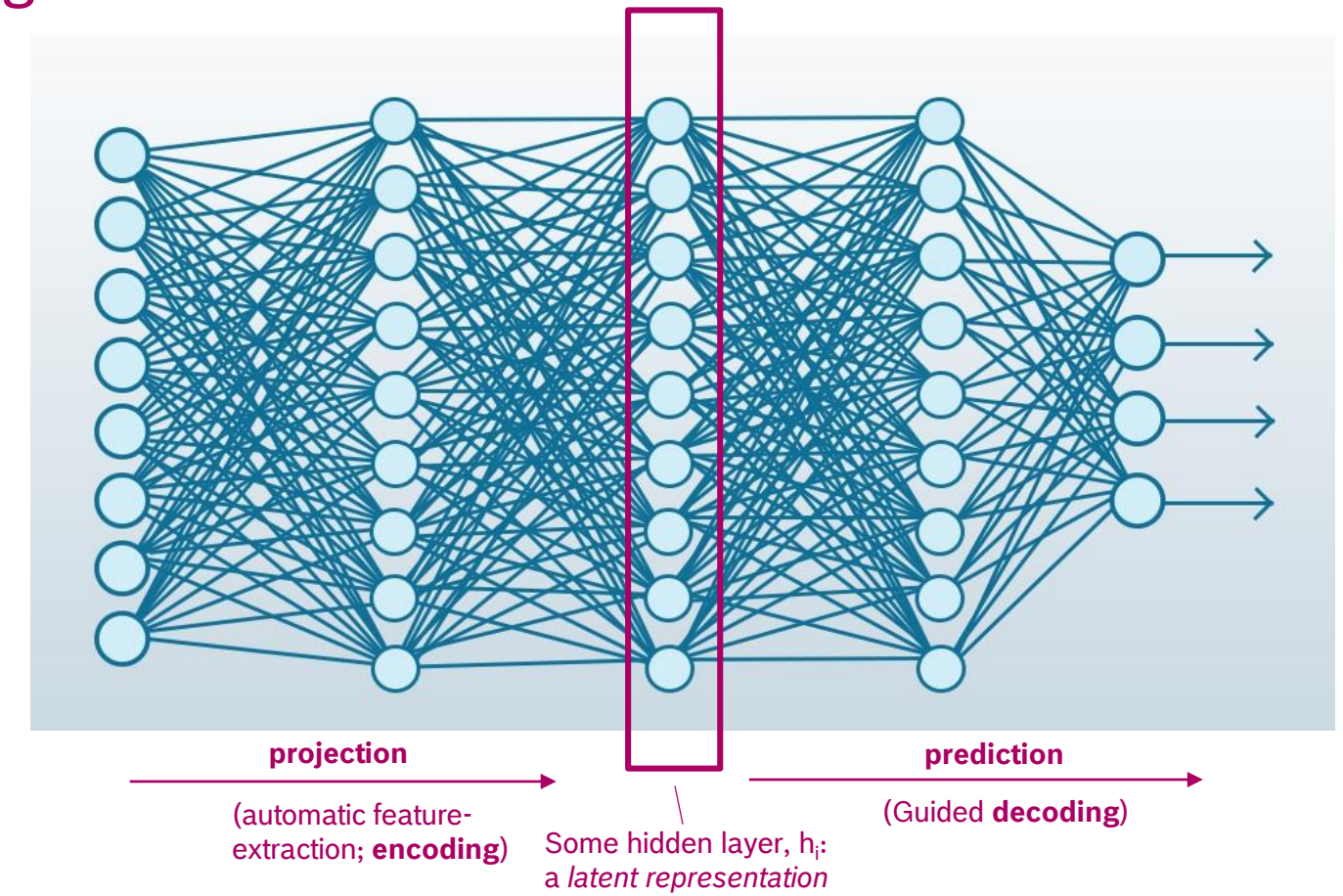
Multimodal Machine Learning

Deep Learning



Multimodal Machine Learning

Deep Learning



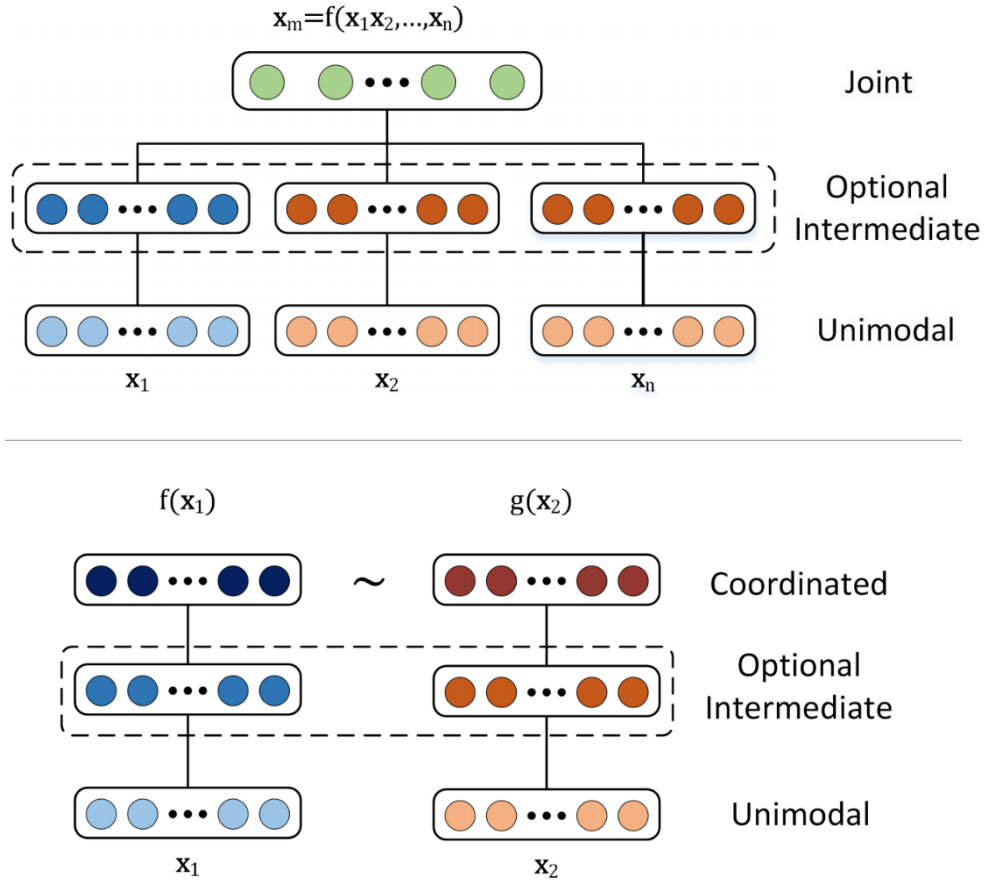
How can we pursue optimal feature extraction, for better performance on downstream tasks?

Multimodal Representation Learning

Multimodal Machine Learning
Representation

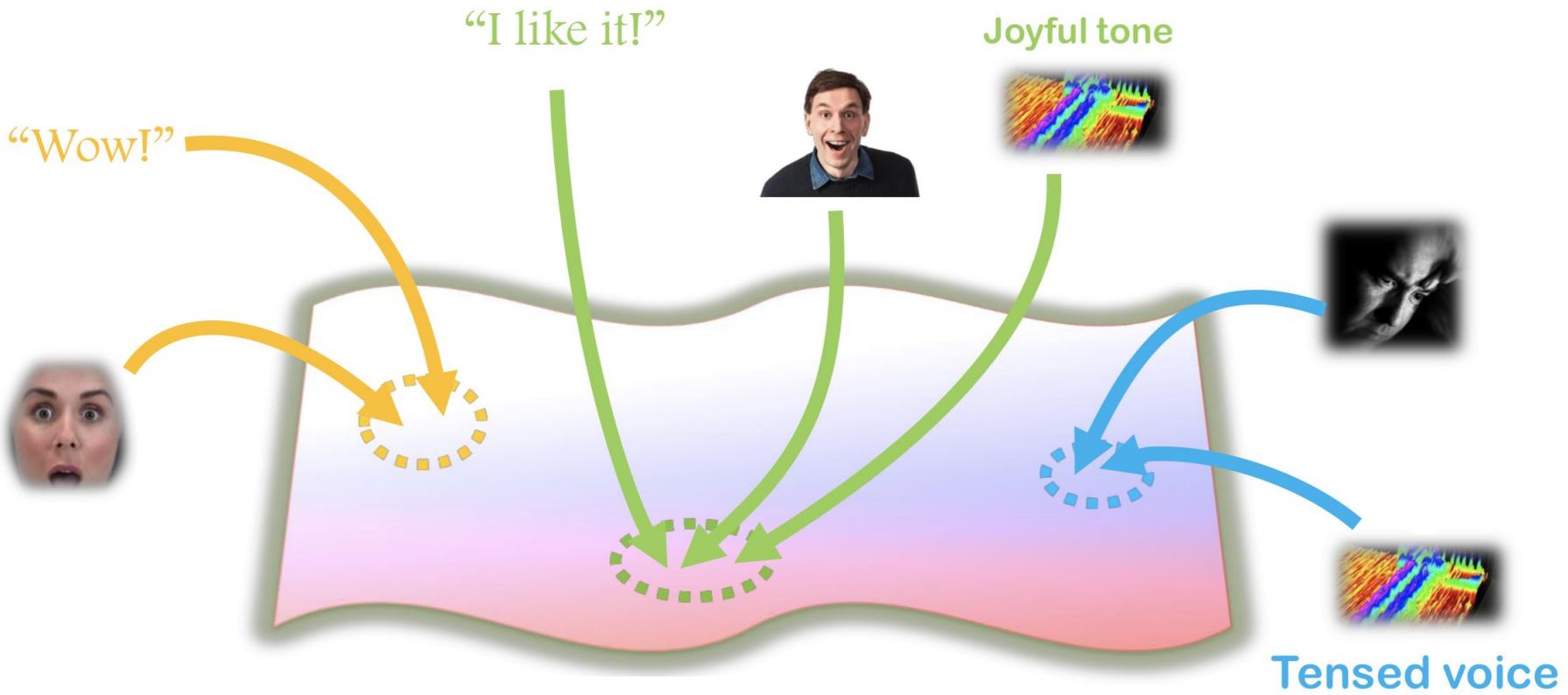
Goal: Learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy.

- *Joint representation:* results in a shared space; geometric regions may translate to semantic concepts
- *Coordinated representation:* each modality is projected into a unimodal space; some subset of dimensions in each spaces are made to be correlated with the others



Multimodal Machine Learning

Holy Grail: A Consistent Multimodal Semantic Space



Multimodal Machine Learning

Joint Multimodal Representation Learning

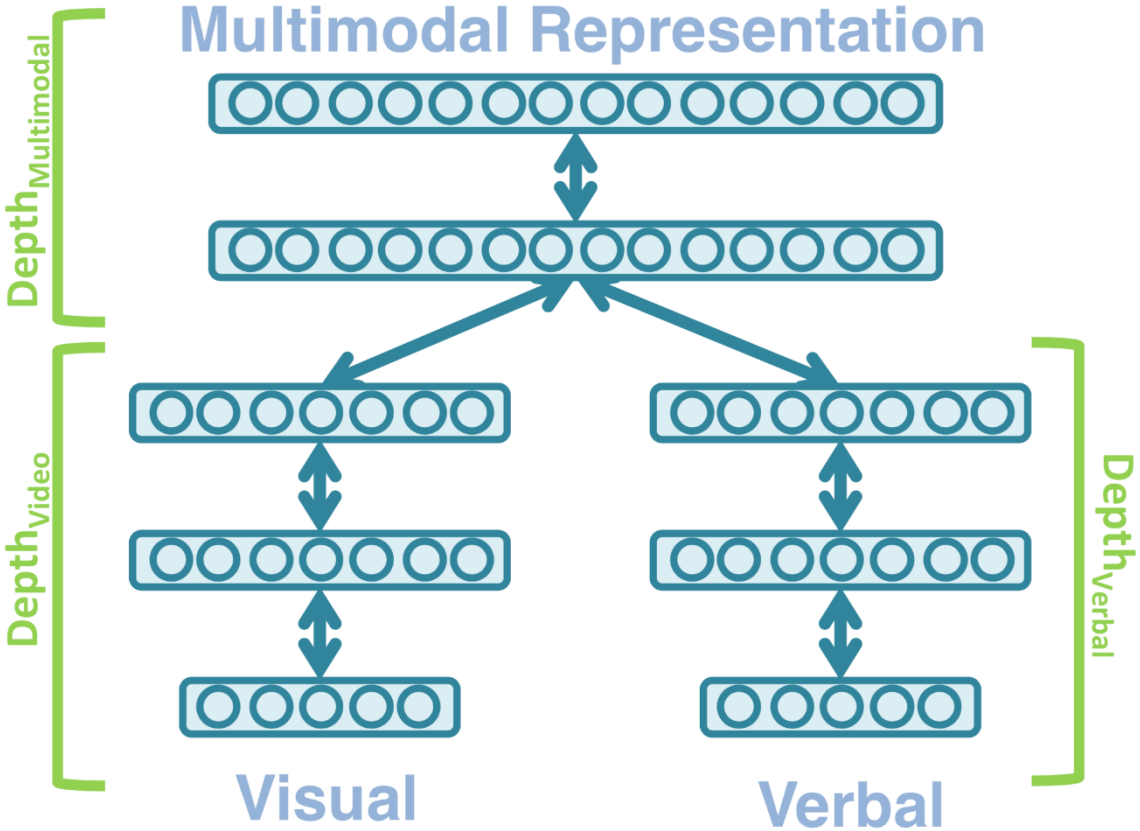
Joint representation: results in a shared space; geometric regions may translate to semantic concepts.

Cross-modal image generation
—> Bimodal latent representation[5]

Audio-visual speech recognition
—> Bimodal Deep Belief Network[8]

Image captioning
—> Multimodal Deep Boltzmann Machines[6]

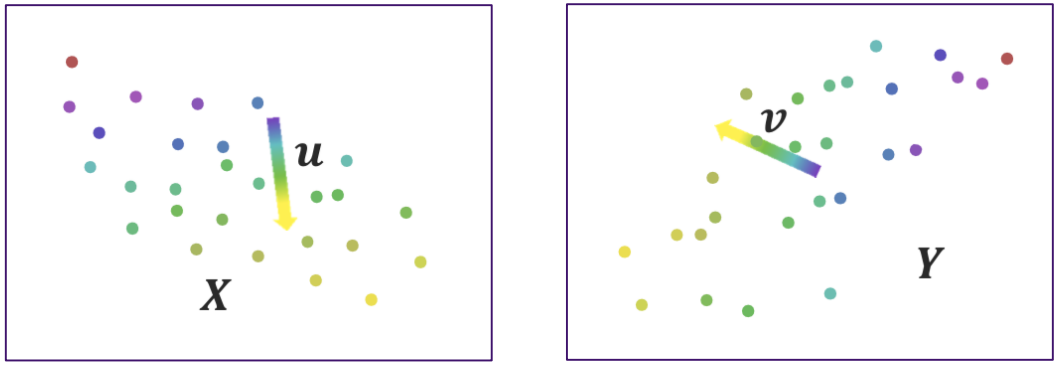
Audio-visual emotion recognition
—> Deep Boltzmann Machine[9]



Multimodal Machine Learning

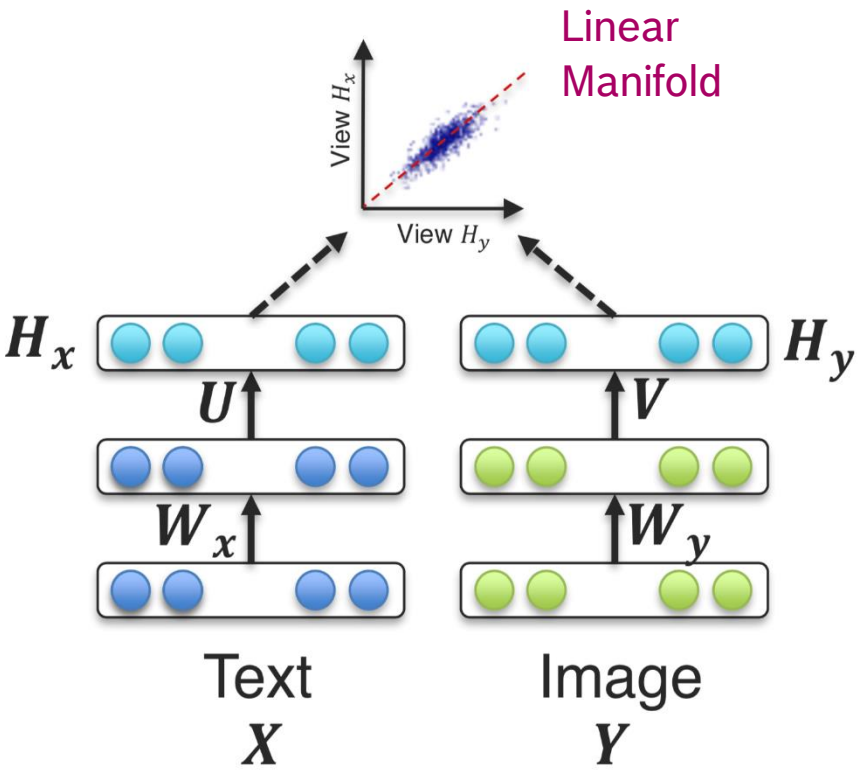
Coordinated Multimodal Representation Learning (Deep CCA)

Coordinated representation: each modality is projected into a unimodal space; (some subset of) the dimensions in each spaces are made to be correlated with the others



Learn a set of projections that are maximally correlated.

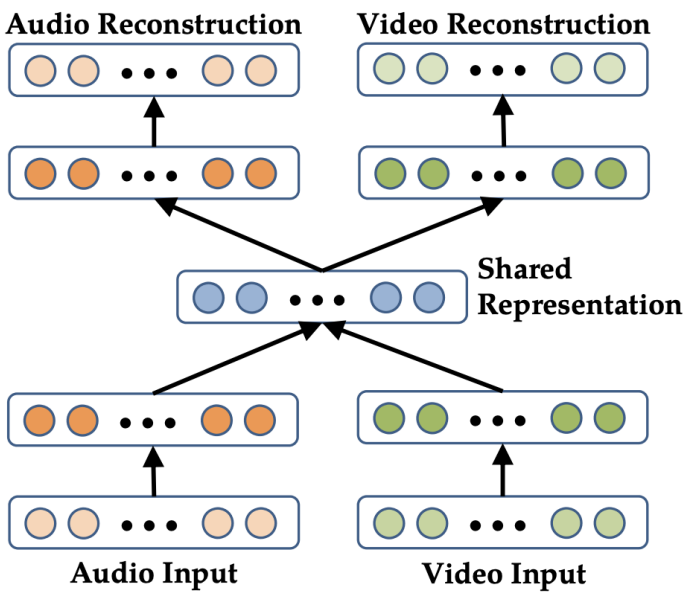
$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr(\boldsymbol{u}^T\boldsymbol{X}, \boldsymbol{v}^T\boldsymbol{Y})$$

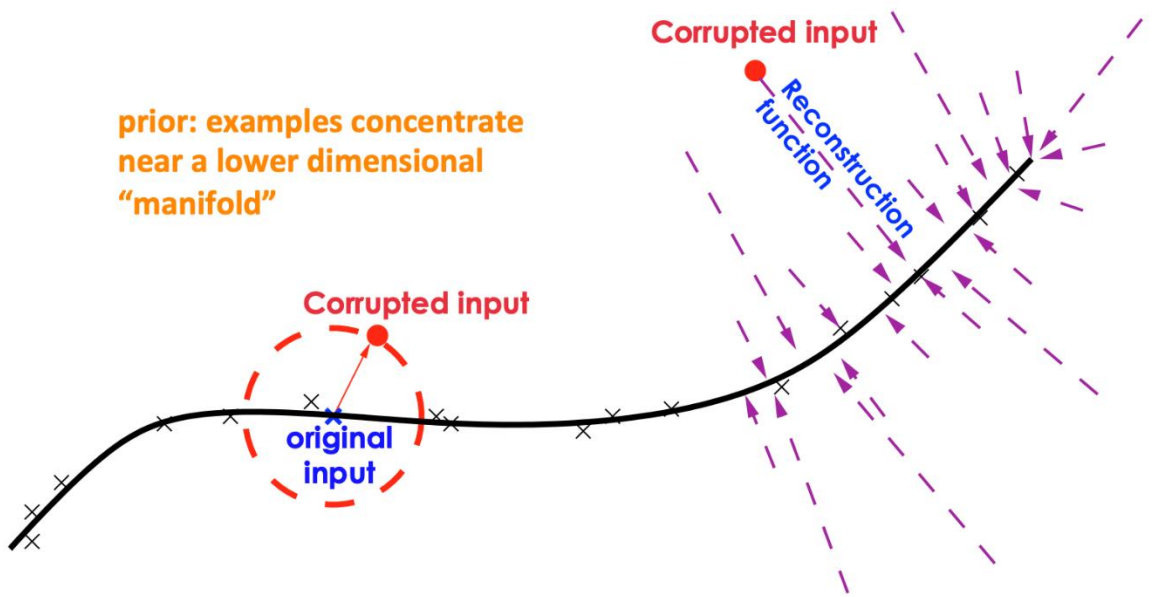


[3] Andrew, G., Arora, R., Bilmes, J. & Livescu, K.. (2013). Deep Canonical Correlation Analysis. Proceedings of the 30th International Conference on Machine Learning, in PMLR 28(3):1247-1255

Multimodal Machine Learning

Manifold Learning (e.g., Denoising Autoencoder)

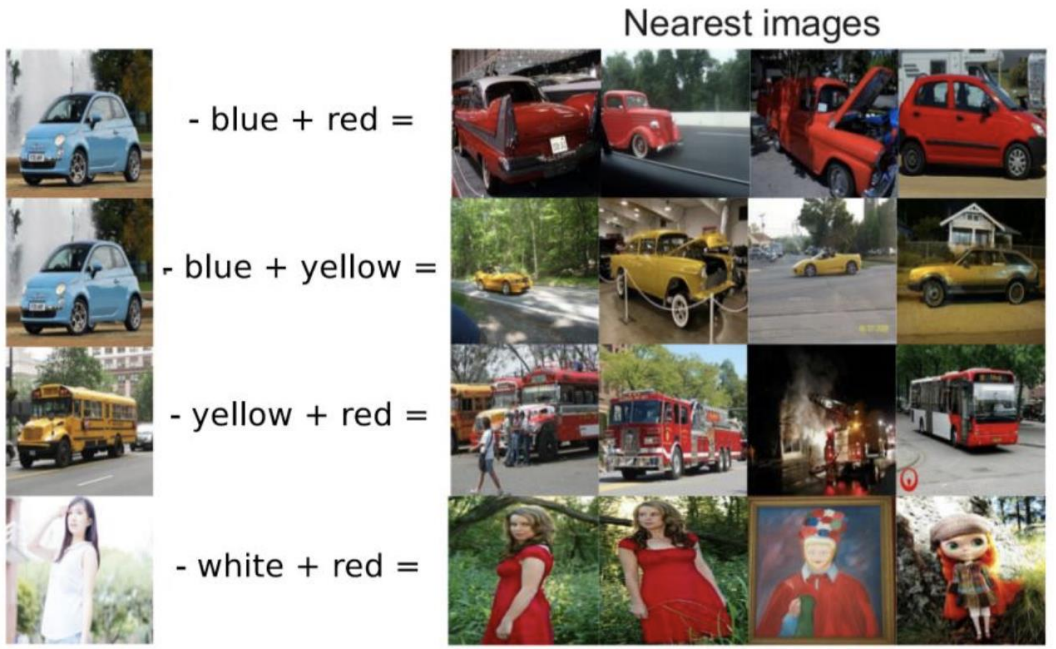


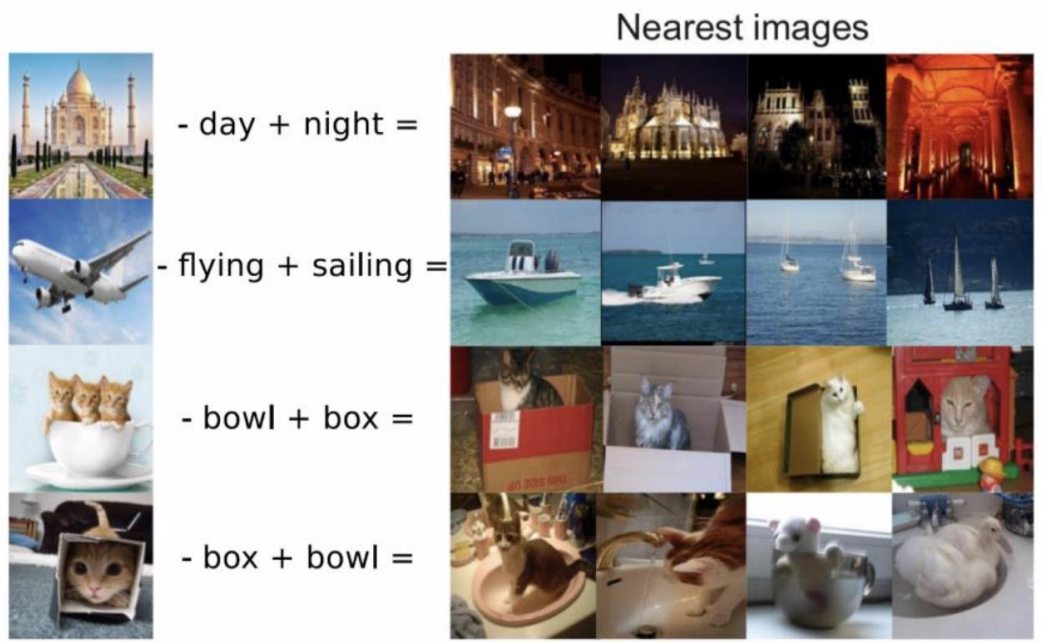


[10] Bengio, Yoshua and Courville, Aaron and Vincent, Pascal. "Representation Learning: A Review and New Perspectives," IEEE Trans. Pattern Anal. Mach. Intell., Washington, DC, 2013, pp. 1798-1828. 10.1109/TPAMI.2013.50

Multimodal Machine Learning

Representation: “Semantic” Vector Arithmetic

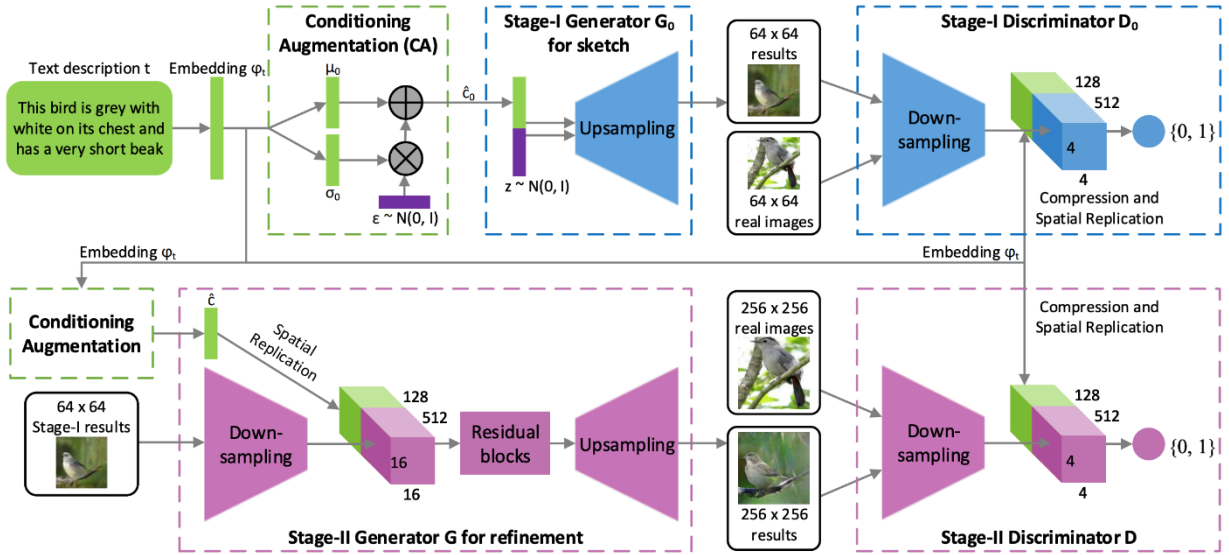


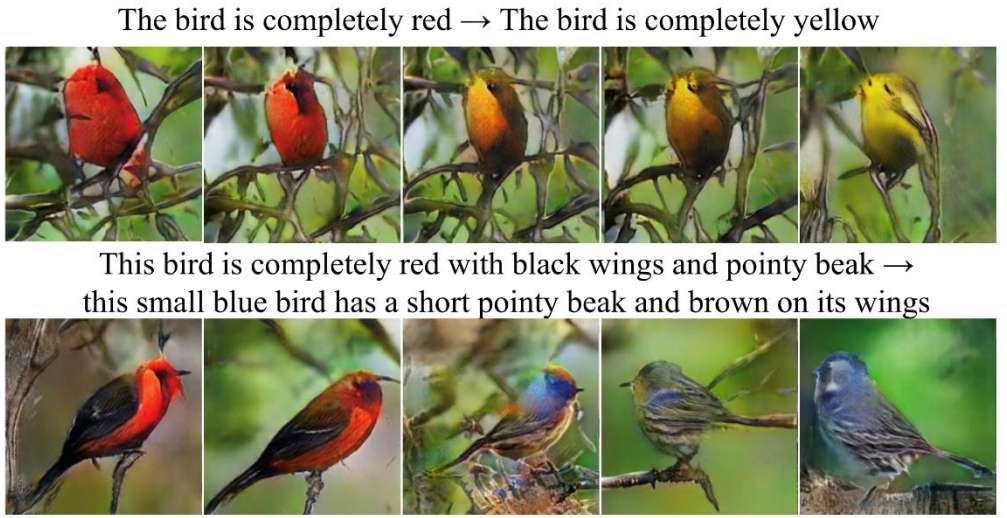


[4] R. Kiros, R. Salakhutdinov, R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models", arXiv preprint arXiv:1411.2539, 2014

Multimodal Machine Learning

Representation: Linear Interpolation





[5] Han Zhang and Tao Xu and Hongsheng Li and Shaoting Zhang and Xiaogang Wang and Xiaolei Huang and Dimitris Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," ICCV, 2017. https://arxiv.org/abs/1612.03242

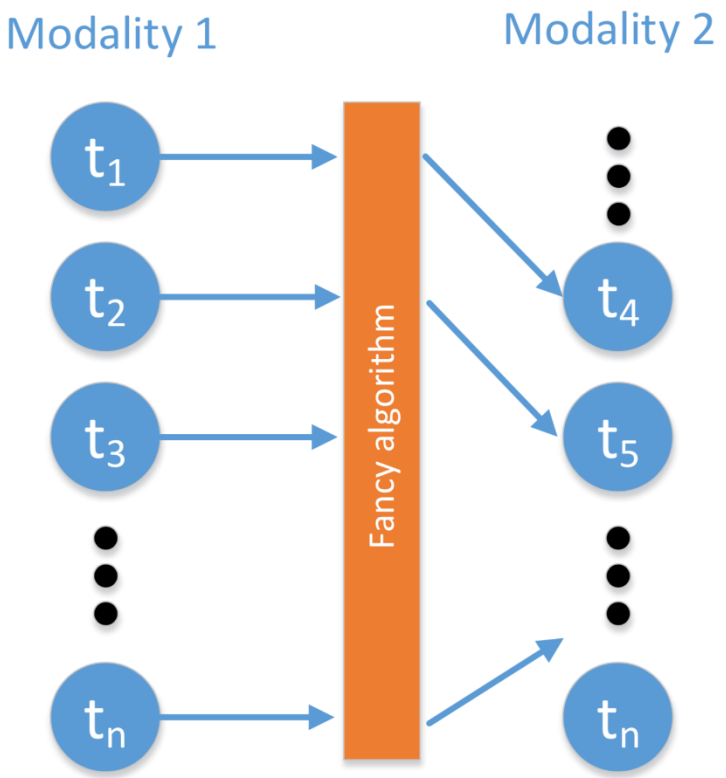
Multimodal Alignment

Multimodal Machine Learning
Alignment

Goal: Identify the direct relations between (sub-) elements from two or more different modalities.

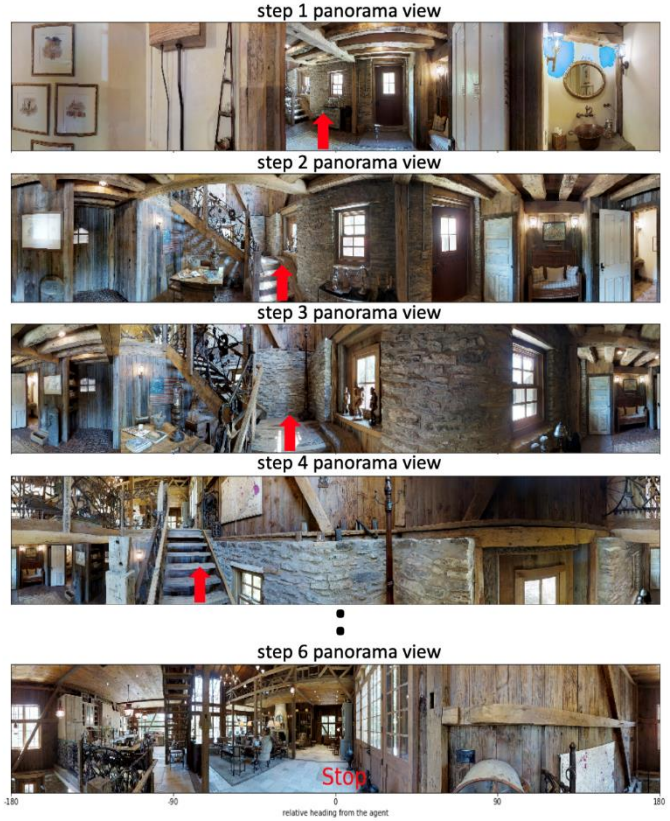
Explicit alignment: directly find correspondences between elements of different modalities (usually based on heuristics – e.g., external domain knowledge)

Implicit alignment: uses internally-learned latent alignment of modalities (e.g., attention mechanisms) in order to solve a problem



Multimodal Machine Learning

Navigation: Temporal Sequence Alignment

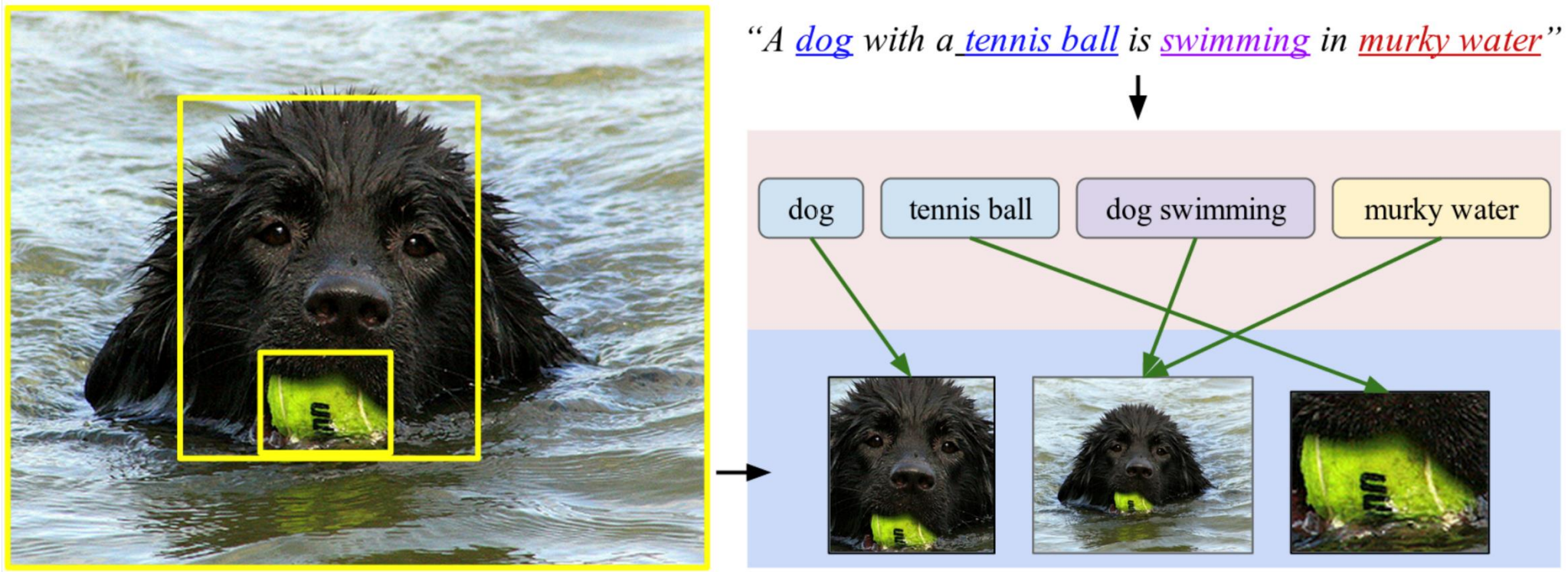


Instruction: Exit the door and turn left towards the staircase. Walk all the way up the stairs, and stop at the top of the stairs.

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

[12] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, Lei Zhang. "Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation," CVPR 2019.

Multimodal Machine Learning
Visual Attention: Grounding (Implicit)

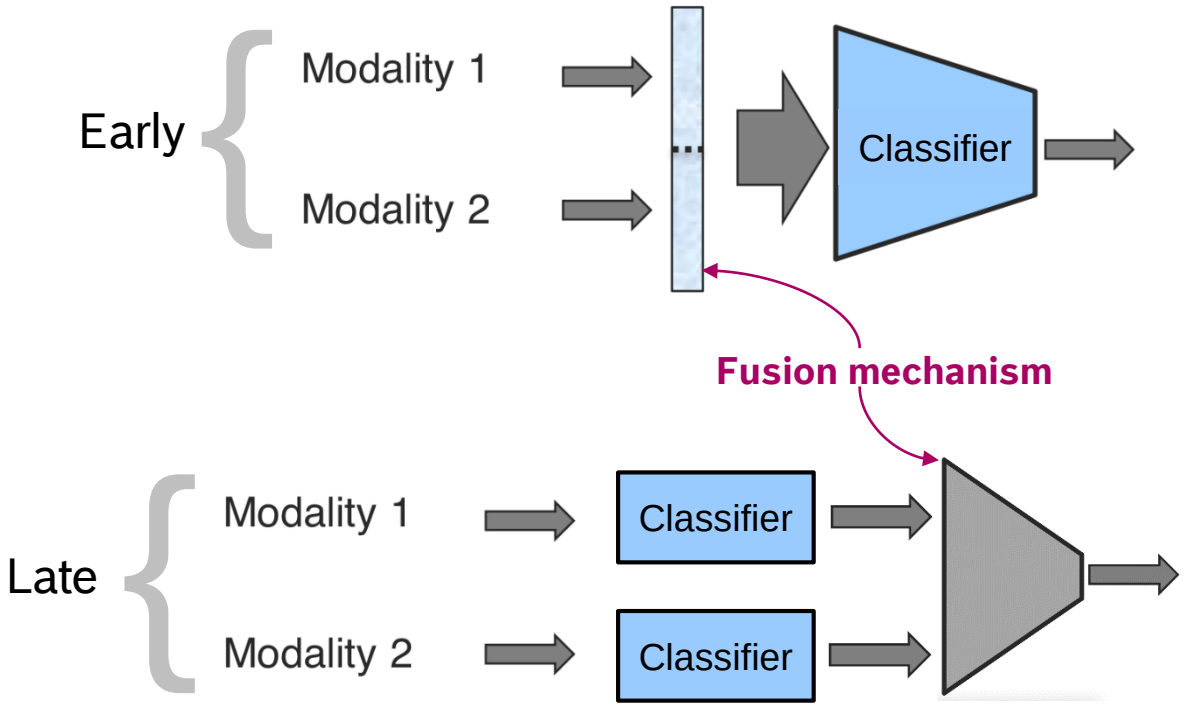


[11] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Vol. 2. MIT Press, Cambridge, MA, USA, 1889-1897.

Data Fusion

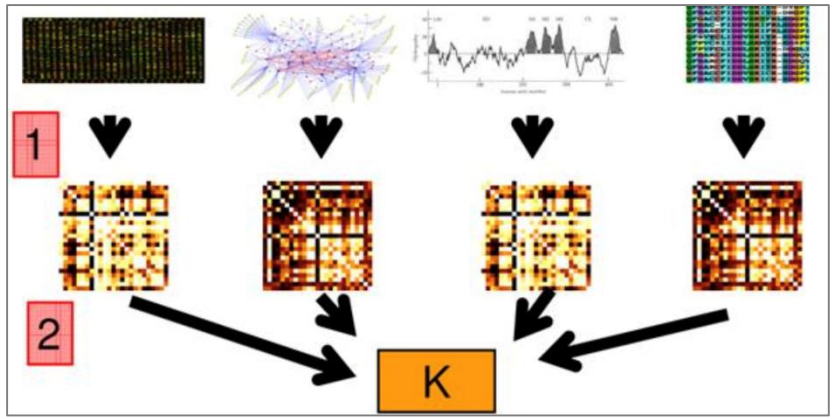
Multimodal Machine Learning
Fusion: Model-agnostic

Goal: Join information from two or more modalities to perform a prediction task.
Approach: *model-agnostic* – explicit fusion mechanisms

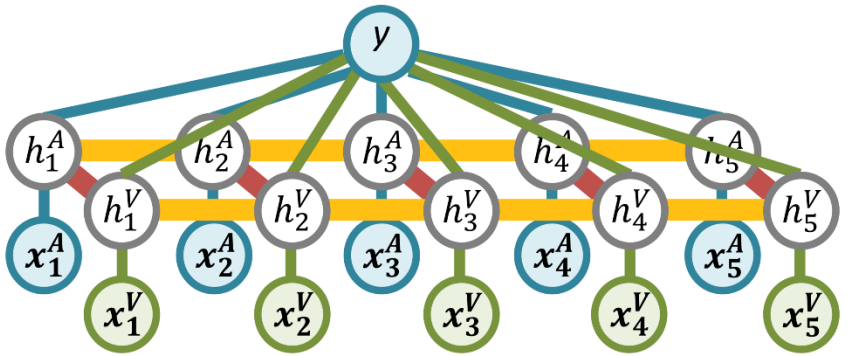


Multimodal Machine Learning
Fusion: Model-based

Goal: Join information from two or more modalities to perform a prediction task.
Approach: *model-based* – implicit fusion mechanism, handled by the model



Multiple Kernel Learning

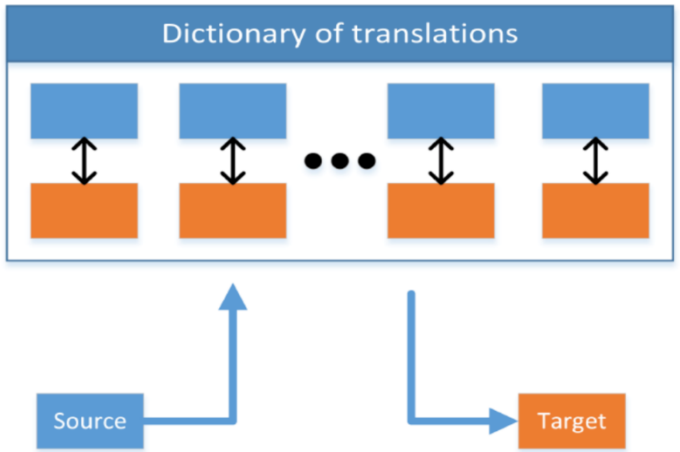


Multimodal Hidden CRF

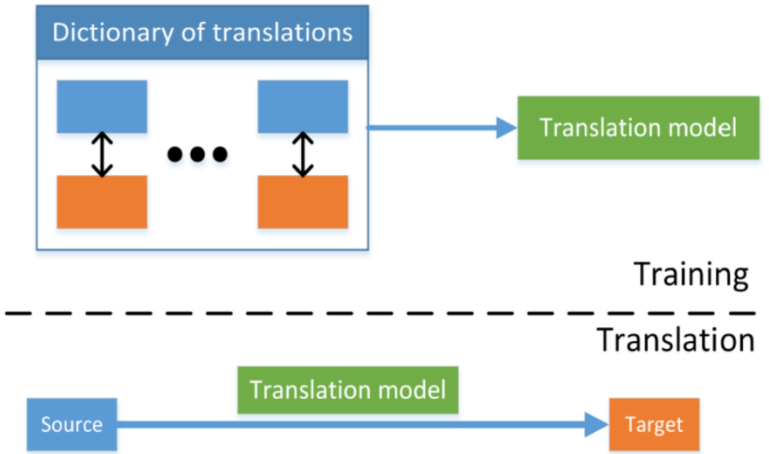
Data/Knowledge Translation

Multimodal Machine Learning
Translation

Goal: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.



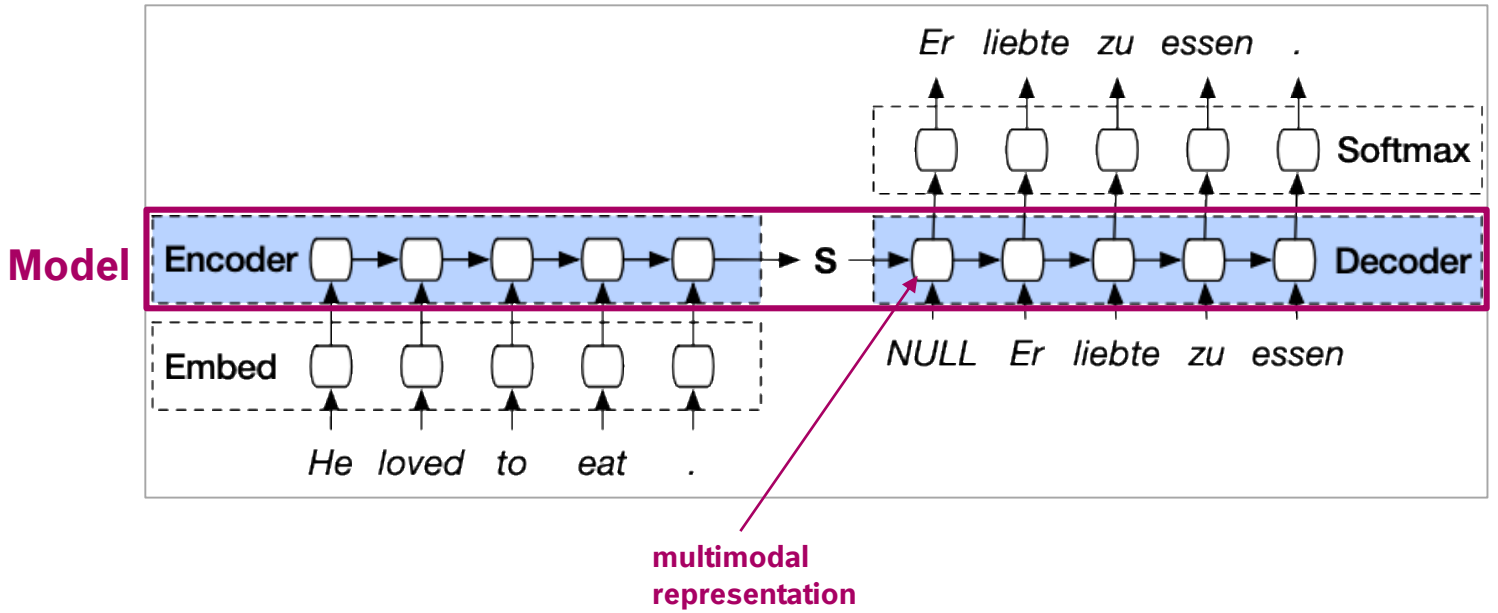
Instance-based



Model-based

Multimodal Machine Learning
Translation

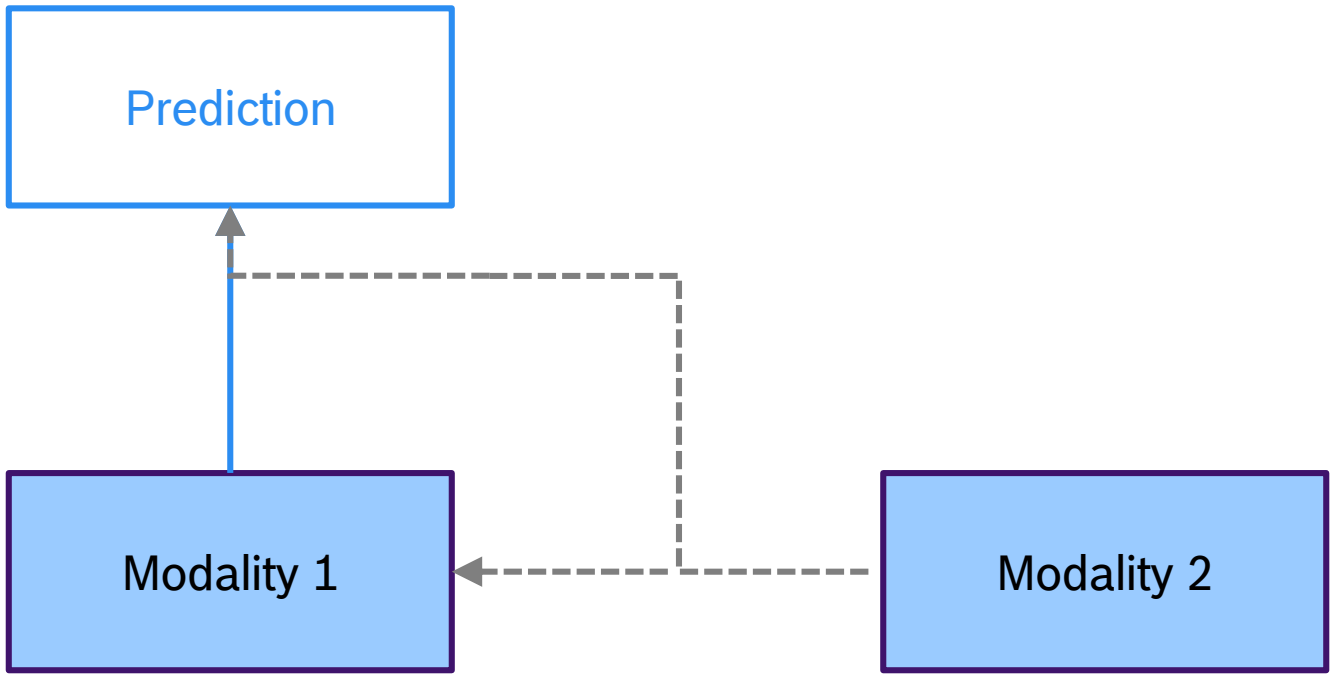
Goal: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.



Modality Co-learning

Multimodal Machine Learning
Co-learning

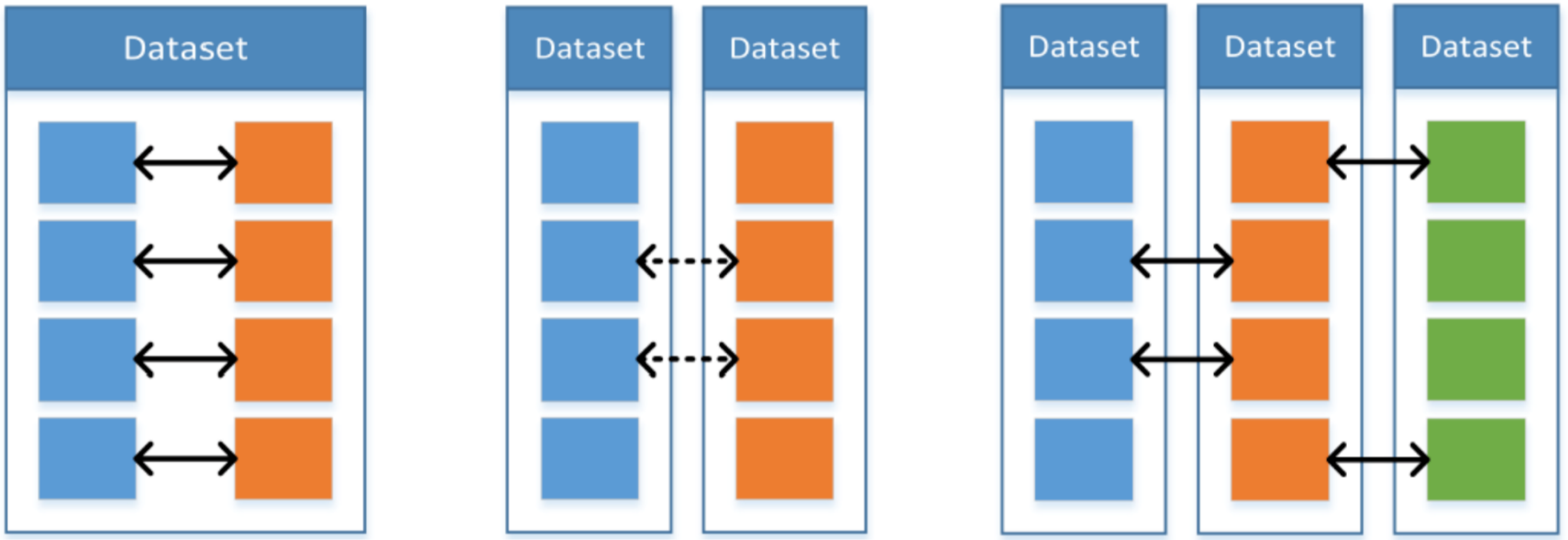
Goal: Transfer “knowledge” between modalities, including their representations and predictive models.



Multimodal Machine Learning

Co-learning

Goal: Transfer knowledge between modalities, including their representations and predictive models.



Multimodal Machine Learning

Taxonomy of Core Challenges

#1. Representation

- *Joint*
  - Neural networks
  - Graphical models
  - Sequential
- *Coordinated*
  - Similarity
  - Structured

#2. Alignment

- *Explicit*
  - Unsupervised
  - Supervised
- *Implicit*
  - Graphical models
  - Neural Networks

#3. Fusion

- *Model Agnostic*
  - Early
  - Late
  - Hybrid
- *Model-based*
  - *Kernel*
  - *Graphical Model*
  - *Neural Networks*

#4. Translation

- *Example-based*
  - Retrieval
  - Combination
- *Model-based*
  - Grammar
  - Encode-decoder
  - Online prediction

#5. Co-Learning

- *Parallel data*
  - Co-training
  - Transfer learning
- *Non-parallel data*
  - Zero-shot
  - Grounding
- *Hybrid Data*
  - *Bridging*

T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," arXiv:1705.09406, 2017.

Multimodal Machine Learning

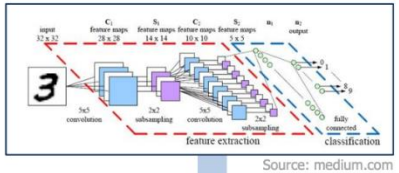
Applications

	Challenges				
Applications	Representation	Translation	Fusion	Alignment	Co-learning
Speech Recognition and Synthesis					
Audio-visual Speech Recognition	✓		✓	✓	✓
(Visual) Speech Synthesis	✓	✓			
Event Detection					
Action Classification	✓		✓		✓
Multimedia Event Detection	✓		✓		✓
Emotion and Affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media Description					
Image Description	✓	✓		✓	✓
Video Description	✓	✓	✓	✓	✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval					
Cross Modal retrieval	✓	✓		✓	✓
Cross Modal hashing	✓				✓

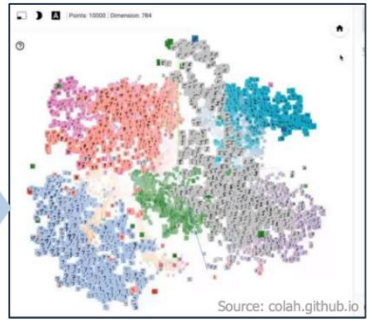
Multimodal Machine Learning

How do we get to *General* AI? Through Hybrid AI.

Learning Grounded Representations



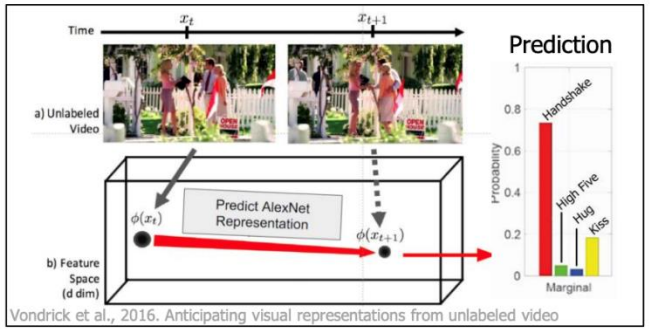
Source: medium.com



Source: colah.github.io

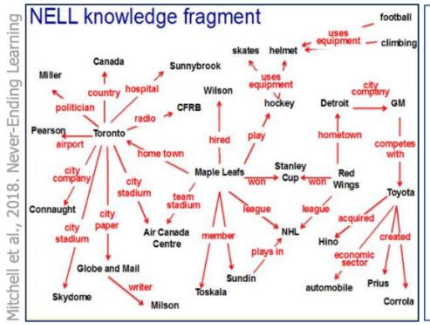
Vector-based "embeddings" extracted from hidden layers

Learning Predictive Models from Experience

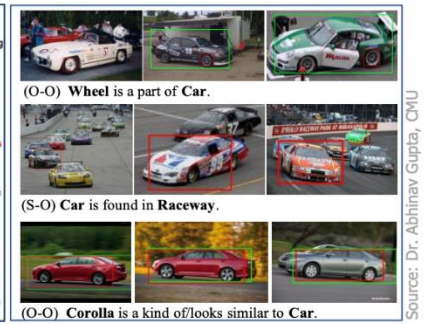


Vondrick et al., 2016. Anticipating visual representations from unlabeled video

Learning Commonsense Knowledge from the Web



Never Ending Language Learning (NELL)



Never Ending Image Learning (NEIL)

Understanding & Modeling Childhood Cognition

Source: scholar.harvard.edu

Elizabeth Spelke (Harvard)

- Objects
- Agents
- Places
- Number
- Forms
- Social Beings

Core Domains of Child Cognition

https://www.darpa.mil/program/machine-common-sense
https://www.darpa.mil/attachments/Final-MCS-PD-2018-10-18-prog-page.pdf

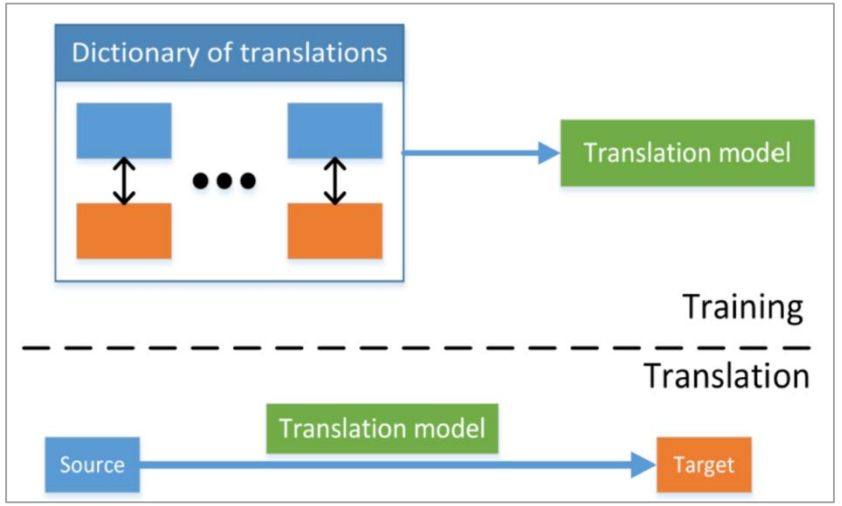
BOSCH

Multimodal Machine Learning

Hybrid AI: Symbolic + Connectionist Systems

Hybrid artificial intelligence

- Prominent components:
  - *Symbolic systems:* can characterise the world/context at the level of abstraction most consumable for humans; discontinuous and cannot be directly used with connectionist systems (e.g., neural networks, Boltzmann machines)
  - *Connectionist systems:* can easily approximate high-dimensional, non-linear functional mappings; input needs to be first projected into a continuous vector space
- Goal: fuse symbolic systems with connectionist systems
  - Knowledge can endow connectionist perception with context
  - Connectionist systems can endow symbolic systems with a statistical characterisation of the world
  - Embrace knowledge as a *modality*
    - Representation: symbol token embeddings (distr. semantics)
    - Co-learning: Knowledge-guided domain-adaptation
    - Alignment: grounding saliency
    - Translation: using, e.g., physical models to regularise encoders, policies in RL, etc.
    - Knowledge graph trajectory encodings



BOSCH

Multimodal Machine Learning

References

[1] R. Velik, R. Lang, D. Bruckner and T. Deutsch, "Emulating the perceptual system of the brain for the purpose of sensor fusion," 2008 Conference on Human System Interactions, Krakow, 2008, pp. 657-662. doi: 10.1109/HSI.2008.4581518

[2] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," arXiv:1705.09406, 2017.

[3] Andrew, G., Arora, R., Bilmes, J. & Livescu, K.. (2013). "Deep Canonical Correlation Analysis," in Proceedings of the 30th International Conference on Machine Learning, in PMLR 28(3):1247-1255

[4] R. Kiros, R. Salakhutdinov, R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models", arXiv preprint arXiv:1411.2539, 2014

[5] Han Zhang and Tao Xu and Hongsheng Li and Shaoting Zhang and Xiaogang Wang and Xiaolei Huang and Dimitris Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," ICCV, 2017. https://arxiv.org/abs/1612.03242

[6] N. Srivastava and R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," Advances in Neural Information Processing Systems 25 (NIPS), 2012. http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf

[7] Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013.

[8] Ngiam, Jiquan and Khosla, Aditya and Kim, Mingyu and Nam, Juhan and Lee, Honglak and Ng, Andrew Y. "Multimodal Deep Learning," in Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML 2011), 2011. http://dl.acm.org/citation.cfm?id=3104482.3104569

[9] Y. Kim, H. Lee and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 3687-3691. doi: 10.1109/ICASSP.2013.6638346

[10] Bengio, Yoshua and Courville, Aaron and Vincent, Pascal. "Representation Learning: A Review and New Perspectives," IEEE Trans. Pattern Anal. Mach. Intell., Washington, DC, 2013, pp. 1798-1828. 10.1109/TPAMI.2013.50

[11] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Vol. 2. MIT Press, Cambridge, MA, USA, 1889-1897.

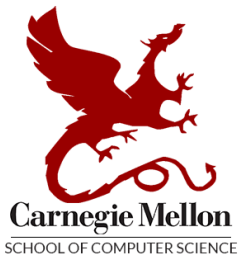
[12] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, Lei Zhang. "Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation," CVPR 2019.

Some lecture material borrowed from LP Morency @ CMU

Thank you!


BOSCH


Carnegie Mellon University
Language
Technologies
Institute


Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE


CARNEGIE MELLON UNIVERSITY
PITTSBURGH PENNSYLVANIA 1900