西北工业大学
NORTHWESTERN POLYTECHNICAL UNIVERSITY

Multimodal Machine Learning

汇 报 人：胡艳

汇报时间：2019年4月

Introduction: Multimodal Machine Learning

➢Key words explanation

- ◆**Modality**: A particular mode in which something exists or is experienced or expressed;
- ◆**Multimodal:** A research problem is characterized as multimodal when it includes multiple such modalities.

➢Aims

- ◆Making progress in understanding the world around us for AI;
- ◆Building models that can process and relate information from multiple modalities

➢Challenges

- ◆Representation（表征）;
- ◆Translation（翻译）;
- ◆Alignment（对齐）;
- ◆Fusion（融合）;
- ◆Co-learning/Joint learning（联合学习）.

Introduction: Multimodal Machine Learning

A Summary of Applications Enabled by Multimodal Machine Learning

	Challenges				
Applications	Representation	Translation	Alignment	Fusion	Co-learning
Speech recognition					
Audio-visual speech recognition	✓		✓	✓	✓
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	✓				✓
Multimedia generation					
(Visual) speech and sound synthesis	✓	✓			
Image and scene generation	✓	✓			

Representation

➢Definition

◆We use the term feature and representation interchangeably, with each referring to a vector or tensor representation of an entity, be it an image, audio sample, individual word, or a sentence. A multimodal representation is a representation of data using information from multiple such entities.

➢Challenges

◆ How to combine the data from heterogeneous sources;

◆ How to deal with different levels of noise;

◆ How to handle missing data.

➢What is good representation

◆Smoothness;

◆Temporal and spatial coherence;

◆Sparsity;

◆Natural clustering;

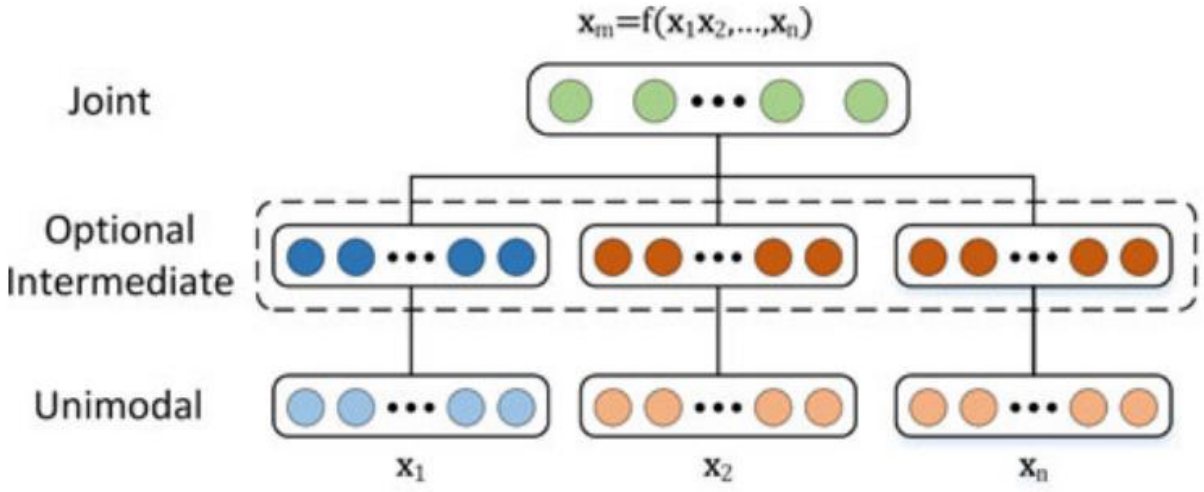
◆Reflect the similarity of the corresponding concepts;

◆Be easy to obtain even in the absence of some modalities;

◆Be possible to fill in missing modalities given the observed ones.

Representation

➢Joint representation

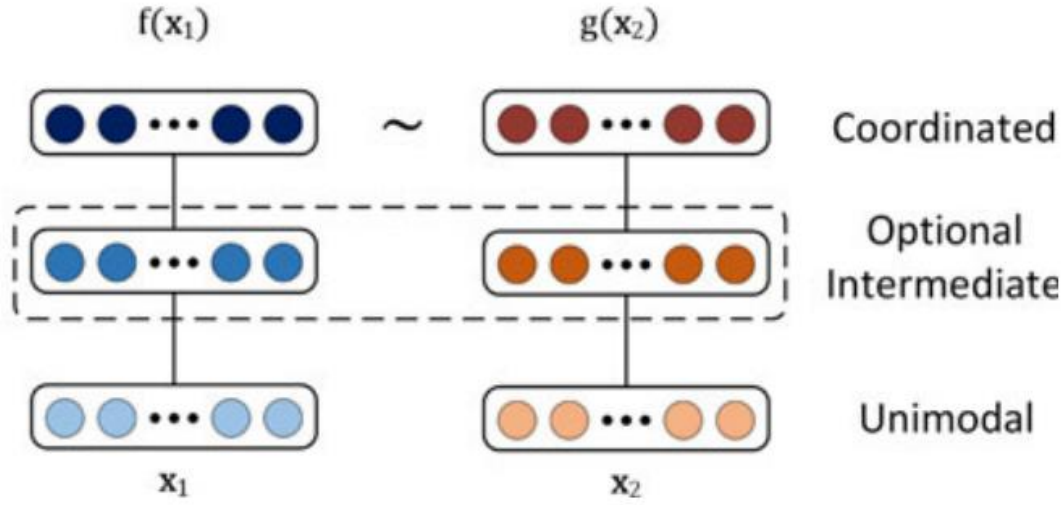


(a) Joint representation

Joint

Neural networks	Images + Audio Images + Text	[150], [157], [235] [191]
Graphical models	Images + Text Images + Audio	[206] [108]
Sequential	Audio + Video Images + Text	[100], [158] [173]

➢Coordinated representation



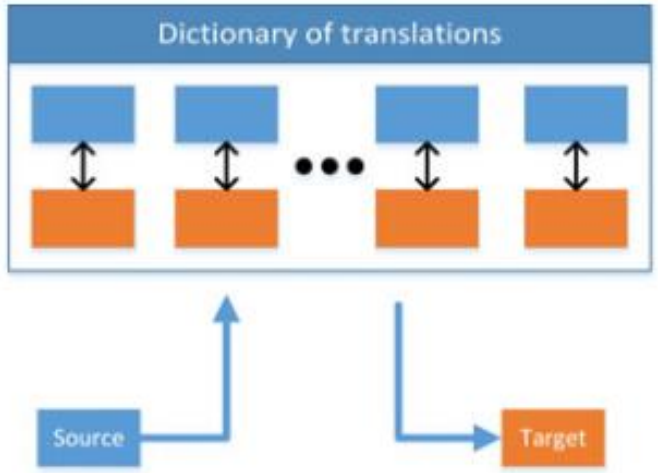
(b) Coordinated representations

Coordinated

Similarity	Images + Text Video + Text	[64], [110] [166], [239]
Structured	Images + Text Audio + Articulatory	[33], [220], [256] [228]

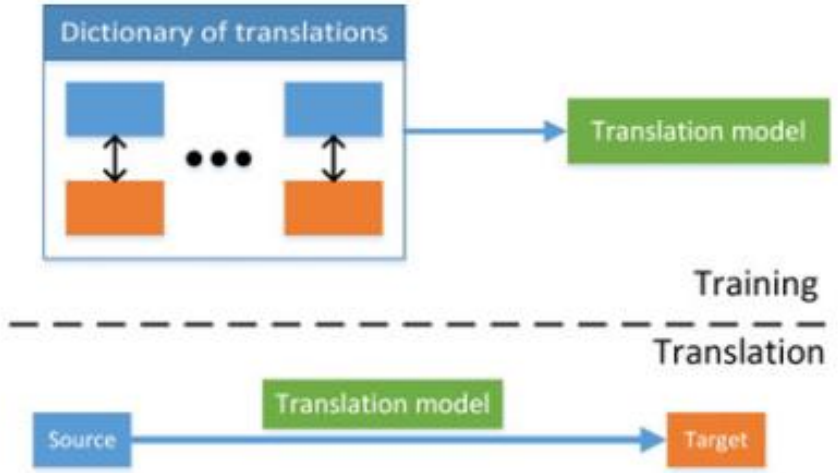
Translation

➢Example-based



(a) Example-based

➢Generative-based



(b) Generative

	Tasks	Dir.	References
Example-based			
Retrieval	Image captioning	⇒	[58], [162]
	Media retrieval	⇔	[199], [239]
	Visual speech	⇒	[27]
	Image captioning	⇔	[102], [103]
Combination	Image captioning	⇒	[77], [119], [124]

Generative			
Grammar based	Video description	⇒	[15], [213]
	Image description	⇒	[53], [126], [147]
Encoder-decoder	Image captioning	⇒	[110], [139]
	Video description	⇒	[222], [249]
	Text to image	⇒	[137], [178]
Continuous	Sounds synthesis	⇒	[161], [164]
	Visual speech	⇒	[5], [49], [212]

Alignment

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

Alignment	Modalities	Reference
Explicit		
Unsupervised	Video + Text Video + Audio	[136], [210], [211] [160], [215], [259]
Supervised	Video + Text Image + Text	[24], [260] [113], [138], [168]
Implicit		
Graphical models	Audio/Text + Text	[194], [224]
Neural networks	Image + Text Video + Text	[102], [236], [238] [244], [249]

- **Challenges**
  - 很少有显式对齐标注的数据集
  - 很难建模不同模态之间相似度计算
  - 存在多个可能的对齐方案并且不是一个模态的所有元素在另一个模态中都存在对应

Fusion

➢Why Fusion

- ◆在观察同一个现象时引入多个模态，可能带来更健壮(robust)的预测
- ◆接触多个模态的信息，可能让我们捕捉到互补的信息（complementary information），尤其是这些信息在单模态下并不“可见”时
- ◆一个多模态系统在缺失某一个模态时依旧能工作

➢Challenge

- ◆在观察同一个现象时引入多个模态，可能带来更健壮(robust)的预测
- ◆接触多个模态的信息，可能让我们捕捉到互补的信息（complementary information），尤其是这些信息在单模态下并不“可见”时
- ◆一个多模态系统在缺失某一个模态时依旧能工作

A Summary of Our Taxonomy of Multimodal Fusion Approaches

Fusion type	Out	Temp	Task	Reference
Model-agnostic				
Early	class	no	Emotion rec.	[35]
Late	reg	yes	Emotion rec.	[175]
Hybrid	class	no	Multimedia event detection	[122]
Model-based				
Kernel-based	class	no	Object class.	[32], [69]
	class	no	Emotion rec.	[94], [189]
Graphical models	class	yes	AVSR	[78]
	reg	yes	Emotion rec.	[14]
	class	no	Media class.	[97]
Neural networks	class	yes	Emotion rec.	[100], [232]
	class	no	AVSR	[157]
	reg	yes	Emotion rec.	[39]

Co-learning/Joint learning

➢Target

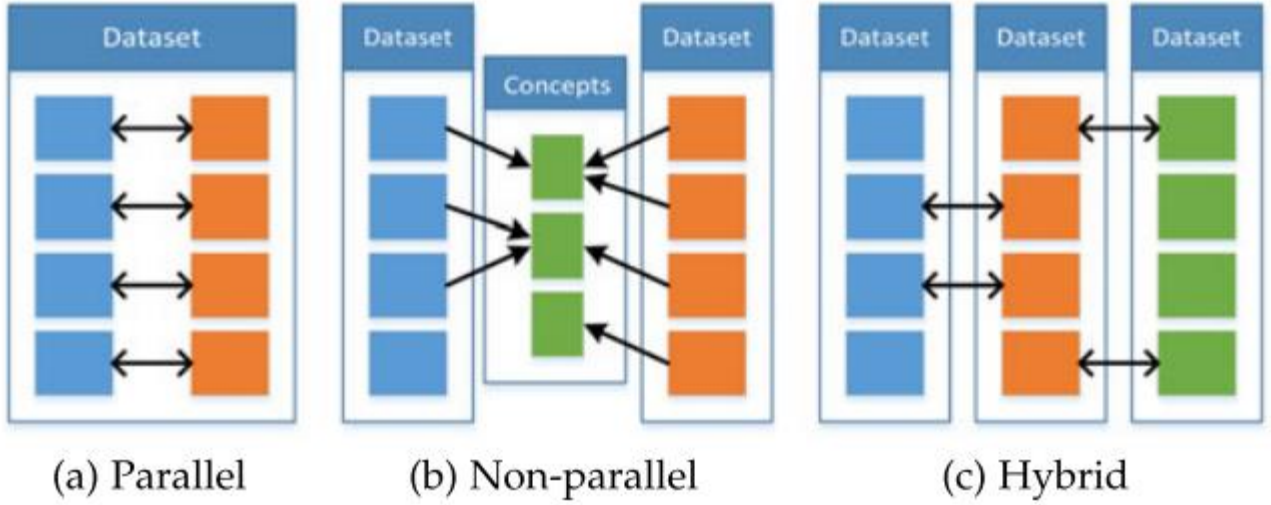
◆Aiding the modeling of a (resource poor) modality, lacking of annotated data, noisy input and unreliable labels by exploiting knowledge from another (resource rich) modality.

➢Methods: classified by training data

◆**Parallel**: Co-training\Transfer learning

◆**Non-parallel**: Transfer learning\Concept grounding（概念接地）\Zero shot learning

◆**Hybrid**: Bridging



Parallel: modalities are from the **same dataset** and there is a direct correspondence between instances;

Non-parallel: modalities are from **different datasets** and do not have overlapping instances, but overlap in general categories or concepts;

Hybrid: the instances or concepts are bridged by **a third modality or a dataset**.

Parallel Data

➢Co-training

◆Examples:

- ➢Web-page classification (Page text and Hyperlink text)
- ➢Bilmetric recognition systems (Appearance and Voice)

Inputs: An initial collection of labeled documents and one of unlabeled documents.

Loop while there exist documents without class labels:

Build classifier A using the A portion of each document.

Build classifier B using the B portion of each document.

For each class C, pick the unlabeled document about which classifier A is most confident that its class label is C and add it to the collection of labeled documents.

For each class C, pick the unlabeled document about which classifier B is most confident that its class label is C and add it to the collection of labeled documents.

Output: Two classifiers, A and B, that predict class labels for new documents. These predictions can be combined by multiplying together and then renormalizing their class probability scores.

Parallel Data

➢Co-training in deep learning

- ◆**Train two neural networks simultaneously**
- ◆Problems:
  - ➢No guarantee that the views provided by the two networks give different and complementary information about each data point;
  - ➢collapsed neural networks.
- ◆Solution:
  - ➢Co-Training loss function:

$$\mathcal{L}_{\mathrm{cot}}(x) = H\left(\frac{1}{2}\left(p_1(x) + p_2(x)\right)\right) - \frac{1}{2}\left(H\left(p_1(x)\right) + H\left(p_2(x)\right)\right)$$

  - ➢View Difference Constraint:

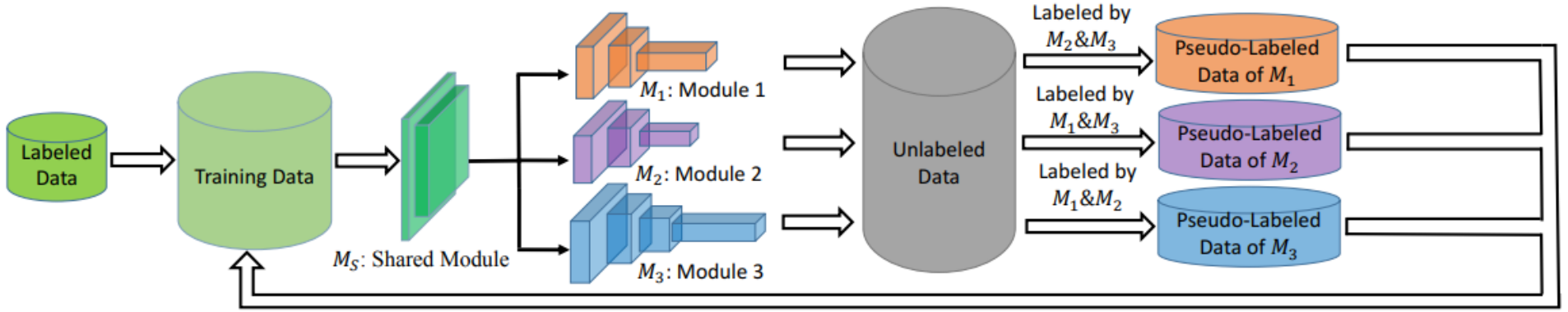
$$\mathcal{L}_{\mathrm{dif}}(x) = H\left(p_1(x), p_2\left(g_1(x)\right)\right) + H\left(p_2(x), p_1\left(g_2(x)\right)\right)$$

Qiao S , Zhang Z . Deep Co-Training for Semi-Supervised Image Recognition[J]. 2018.

Parallel Data

➢Co-training in deep learning

◆Diversity in model initialization



- Output Smearing：通过向有标签数据的添加随机噪声来构造不同的数据集 D1，D2，D3，分别用来初始化对应的高层抽象部分 M1，M2，M3，这样就能使用高层抽象部分多样化；
- Diversity Augmentation：为了解决 collapsed neural networks 问题，在特定的 epoch 继续使用 output smearing 数据集来 fine-tune 网络，以继续增强多样化；
- Pseudo-Label Editing：训练时，需要挑选可靠且稳定的无标签样本加入训练集(通过 Dropout 的随机性，对样本进行多次预测，如果多次预测的结果都几乎一样，则认为是稳定的)。

Chen D, Wang W, Gao W, et al. Tri-net for Semi-Supervised Deep Learning[C]. international joint conference on artificial intelligence, 2018: 2014-2020.

Parallel Data

➢Challenges of Co-training in deep learning

◆怎么训练具有不同视图信息的分类器？目前看到的方法有二：1）构建不同的数据集；2）使用不同的网络架构。看起来两种方法一起用效果会更好；

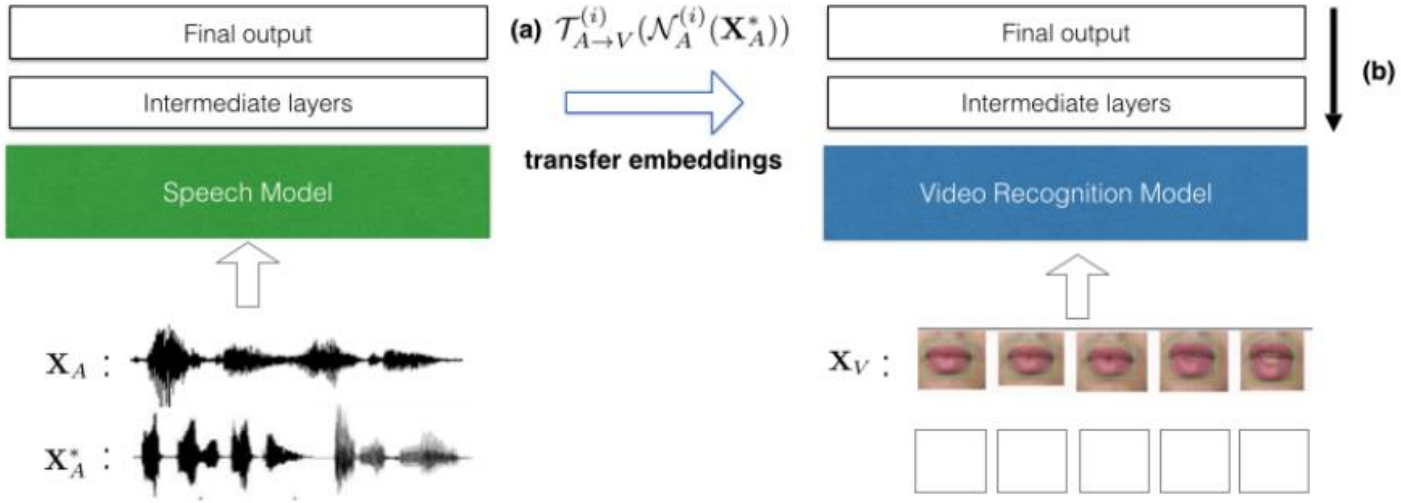
◆如何解决 collapsed neural networks 问题，即如何保持分类器的 diversity；

➢convex clustering

◆如何训练“好”的无标签样本加入训练集？虽然协同训练本身通过一致性原则选择的样本就具有一定的可靠性，但是否有很好的挑选方法？

Parallel Data

➢Transfer Learning



Algorithm 1 Transfer Deep Learning Fine-Tuning (TDLFT)

Input: $\mathcal{N}_A$ trained with $\mathbf{X}_A$, $\mathcal{N}_V$ trained with $\mathbf{X}_V$, an input parameter $i \in \{0, 1, \cdots, l\}$, $\mathcal{T}^{(j)}_{A\to V}$: $\mathbf{H}^{(j)}_A \to \mathbf{H}^{(j)}_V$ learned for $j := i$, a new unparallel data $(\mathbf{X}^*_A, \mathbf{Y}^*)$.
Output: $\mathcal{N}_V$ fine-tuned with $\mathbf{X}^*_A$.
Obtain $\mathbf{H}^{(i)}_V := \mathcal{T}^{(i)}_{A\to V}(\mathcal{N}^{(i)}_A(X^*_A))$, and $\mathbf{H}^{(l+1)}_V := \mathbf{Y}^*$
for $j \in \{i, i+1, \cdots, l\}$ **do**
  $\mathbf{H}^{(j+1)}_V := g(\mathbf{H}^{(j)}_V, \mathcal{W}^{(j\to j+1)}_V)$
end for
Fine-tune $\mathcal{W}^{(j\to j+1)}_V$ for $j \in \{i, i+1, \cdots, l\}$ via a standard backpropagation algorithm.

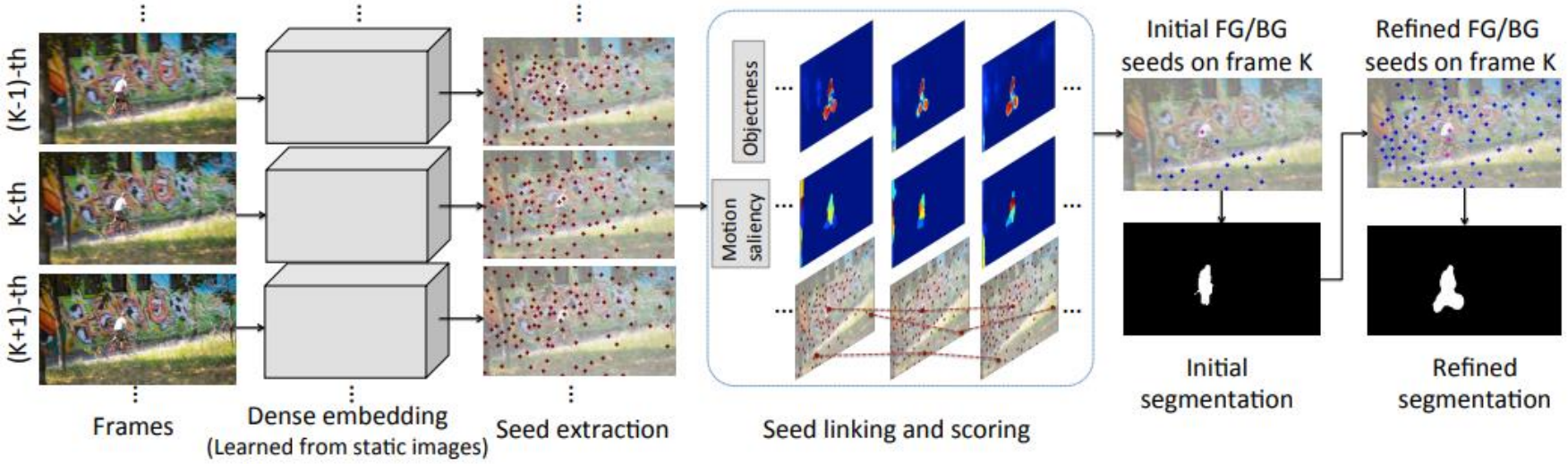
$\boldsymbol{X_A}$和$\boldsymbol{X^*_A}$有同样类型的数据，目标是学习**Embedding function** $T_{A\to V}$.

- Multivariate Support Vector Regression (SVR) Using Nonlinear Kernels;
- KNN-based Non-parametric Mapping;
- Normalized Canonical Correlation Analysis (NCCA)

Moon S, Kim S, Wang H, et al. Multimodal Transfer Deep Learning with Applications in Audio-Visual Recognition[J]. arXiv: Neural and Evolutionary Computing, 2014.

Parallel Data

➢Transfer Learning



◆Given the video sequences, the dense embeddings are obtained by applying an instance segmentation network trained on static images;

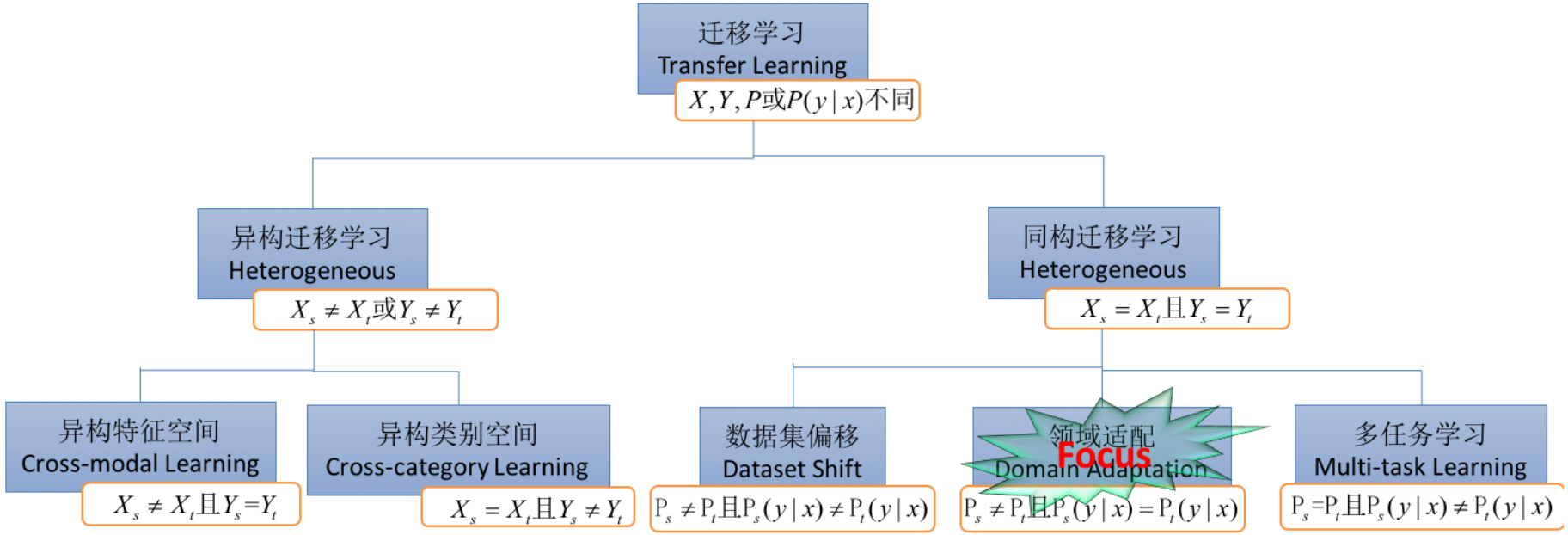
➢Challenges

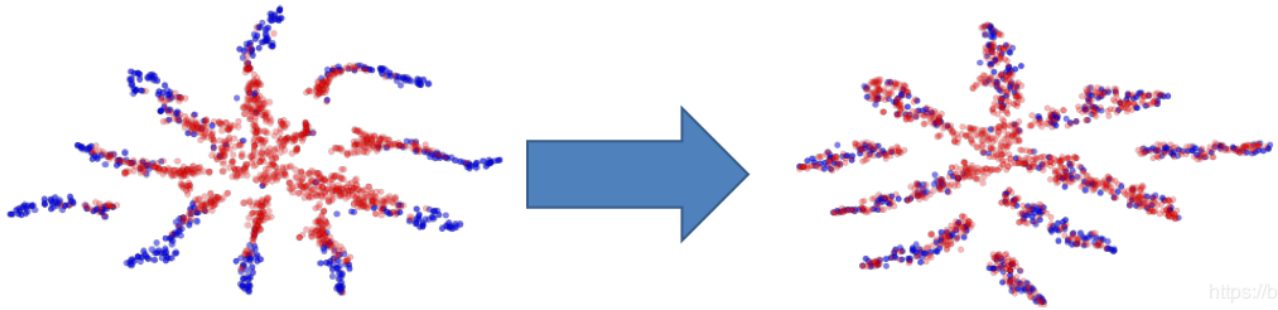
◆**Embedding function ????**

Li S , Seybold B , Vorobyov A , et al. Instance Embedding Transfer to Unsupervised Video Object Segmentation[J]. 2018.

Non-parallel Data

➢Transfer Learning





Non-parallel Data

➢Transfer Learning

◆基于实例的迁移

- ➢如何从源领域中挑选出，对目标领域的训练有用的实例，比如对源领域的有标记数据实例进行有效的权重分配，让源域实例分布接近目标域的实例分布，从而在目标领域中建立一个分类精度较高的、可靠地学习模型。

◆基于特征的迁移

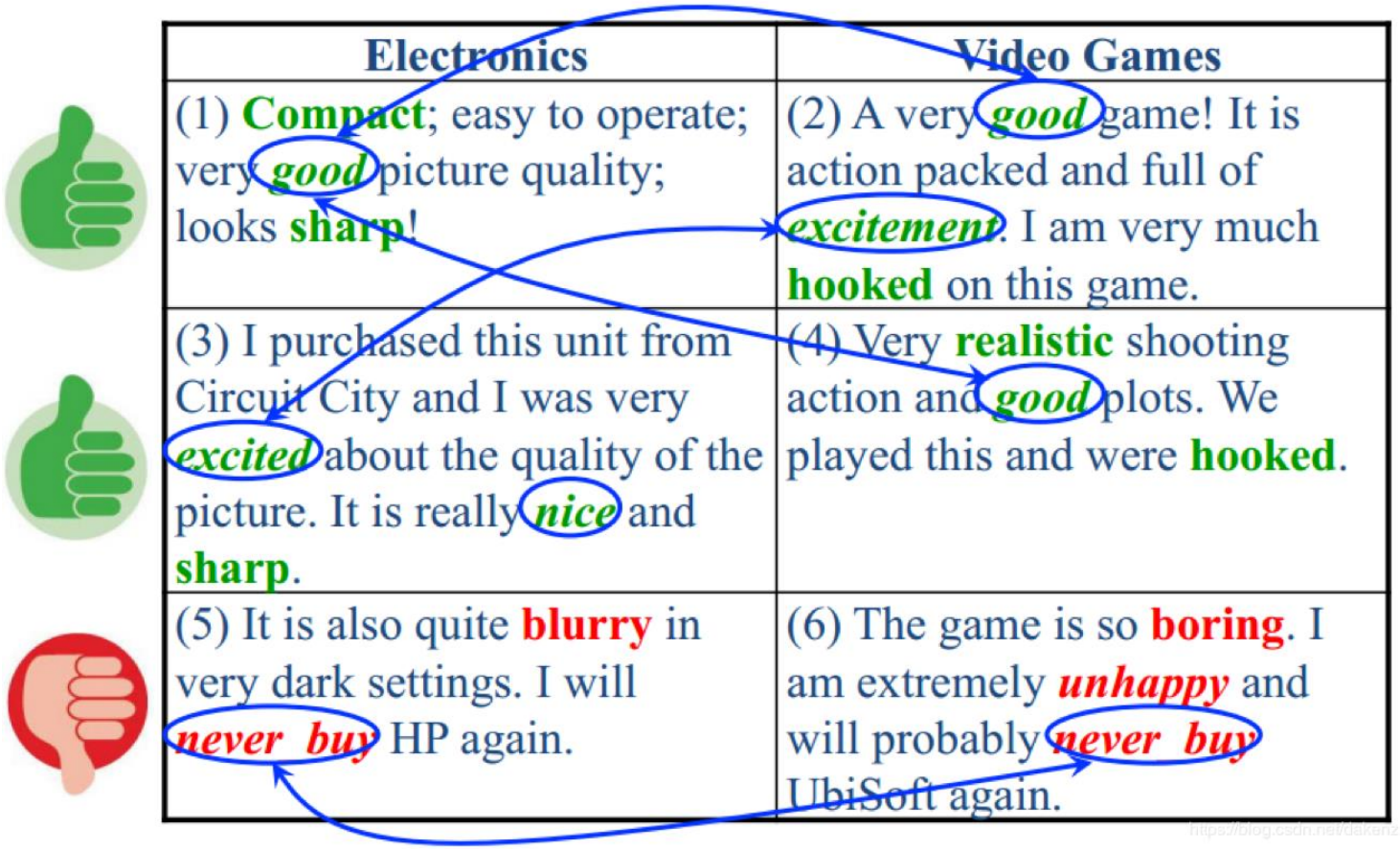
- ➢基于特征选择的迁移学习算法，关注的是如何找出源领域与目标领域之间共同的特征表示，然后利用这些特征进行知识迁移。

◆基于共享参数的迁移

- ➢基于共享参数的迁移研究的是如何找到源数据和目标数据的空间模型之间的共同参数或者先验分布，从而可以通过进一步处理，达到知识迁移的目的，假设前提是，学习任务中的的每个相关模型会共享一些相同的参数或者先验分布。

Non-parallel Data

➢基于特征映射的迁移学习

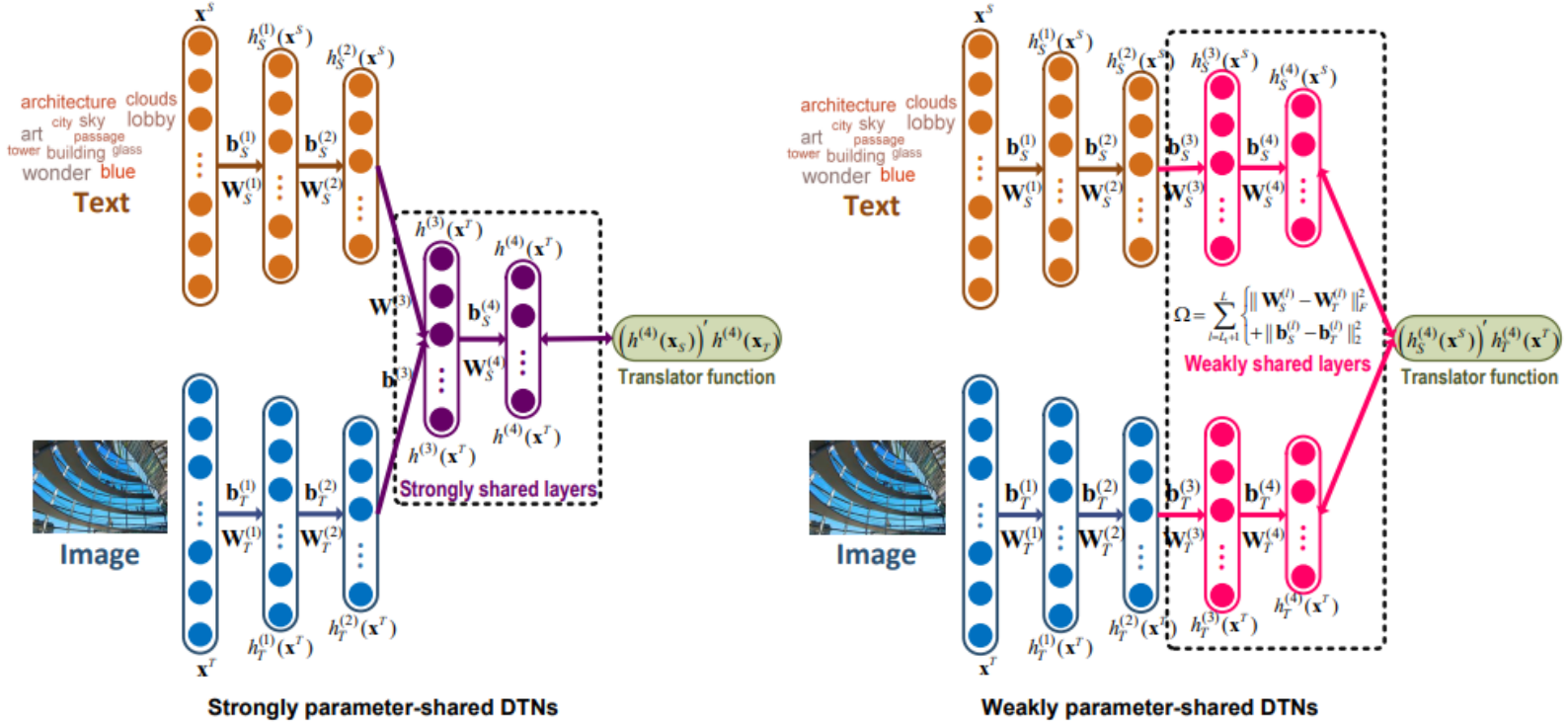


基于特征映射的迁移学习算法，关注的是如何将源领域和目标领域的数据从原始特征空间映射到新的特征空间中去。

这样，在该空间中，源领域数据与的目标领域的数据分布相同，从而可以在新的空间中，更好地利用源领域已有的有标记数据样本进行分类训练，最终对目标领域的数据进行分类测试。

Non-parallel Data

➢Heterogeneous-domain: Text-Image



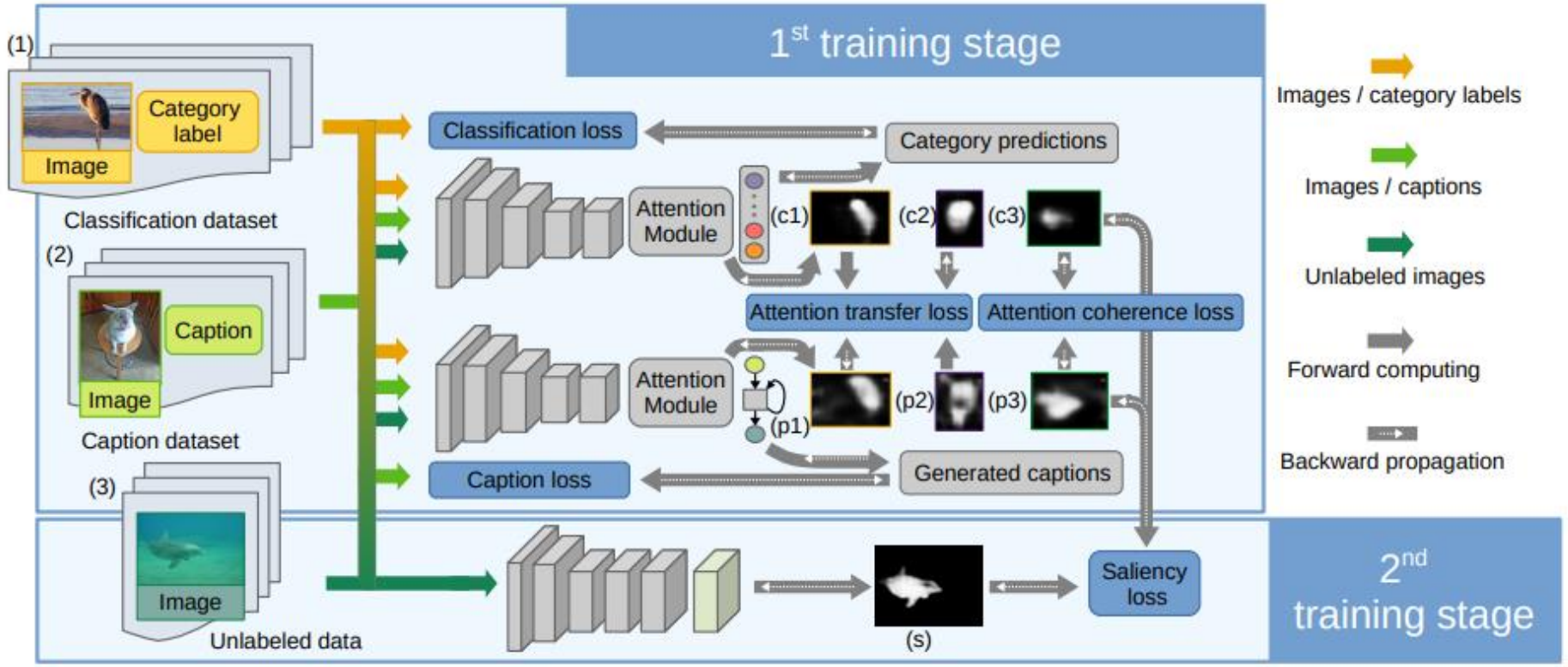
◆把图片和文本先输入到一个各自独有的网络中，但是这两个独有的网络并不是独立的，需要满足如下面的约束。然后再输入到一个共享的网络中，希望提取出公共的特征。

$$\Omega = \sum_{l=1}^{L} (\|W_S^l - W_T^l\|_F^2 + \|b_S^l - b_T^l\|_F^2)$$

Shu X, Qi G, Tang J, et al. Weakly-Shared Deep Transfer Networks for Heterogeneous-Domain Knowledge Propagation[C]. acm multimedia, 2015: 35-44.

Non-parallel Data

➢Heterogeneous-domain: Text and Image



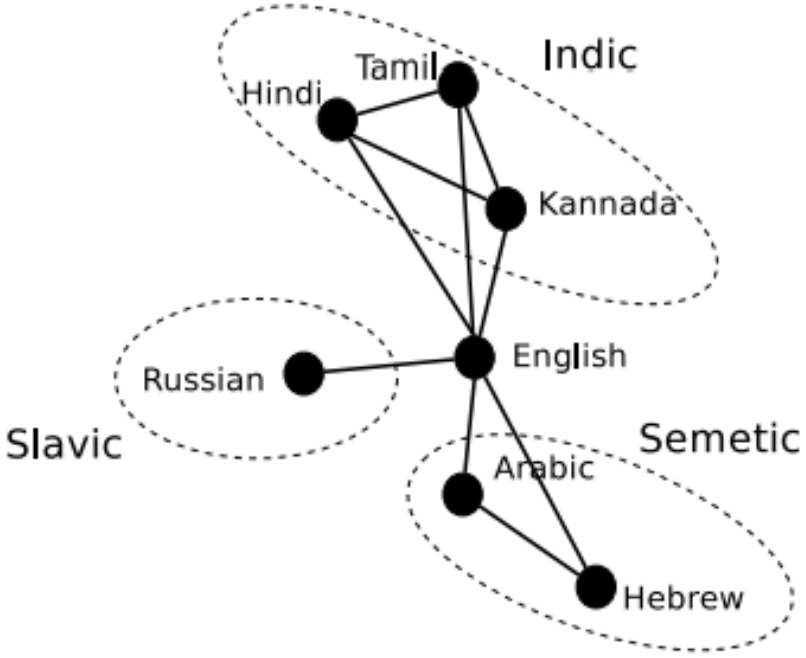
◆每一个域都是弱监督，Joint Learning实现以往需要强监督实现的任务

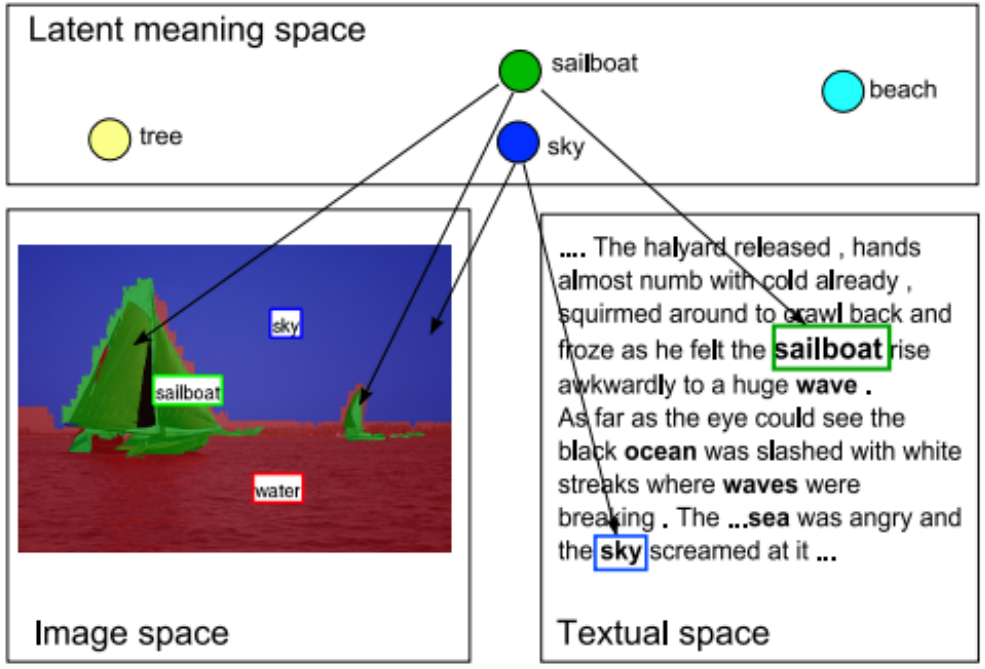
◆Attention，共有特征

Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, Yizhou Yu. Multi-source weak supervision for saliency detection. Accepted by CVPR 2019

Hybrid Data

- Two non-parallel modalities are bridged by a shared modality or a dataset.

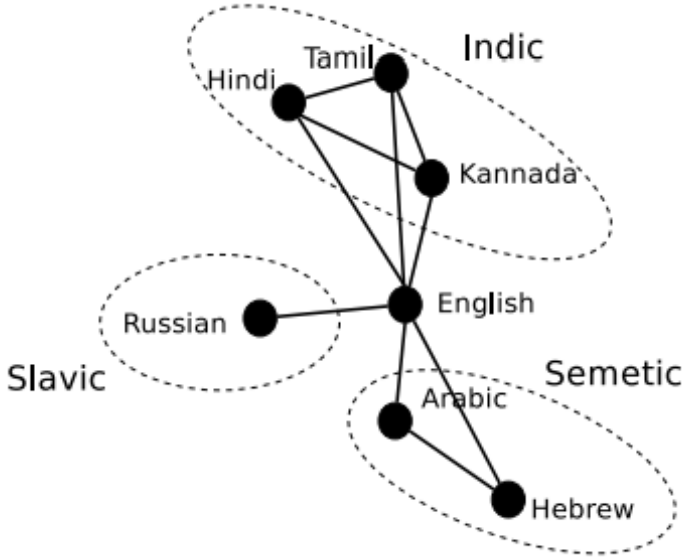


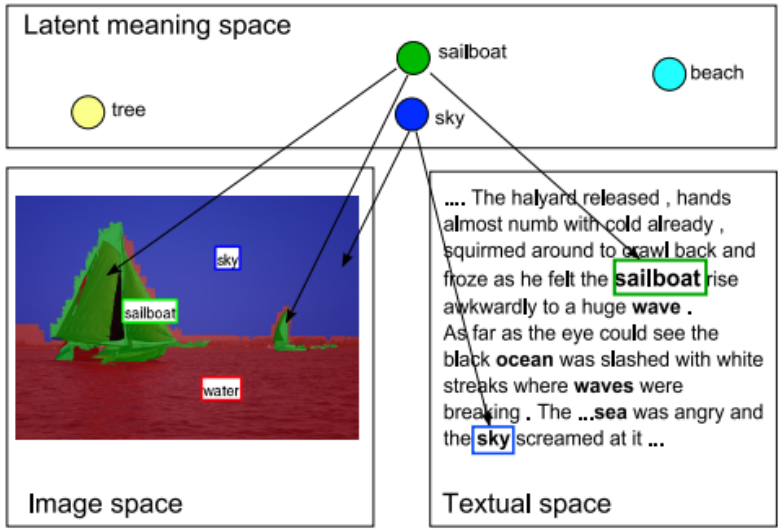


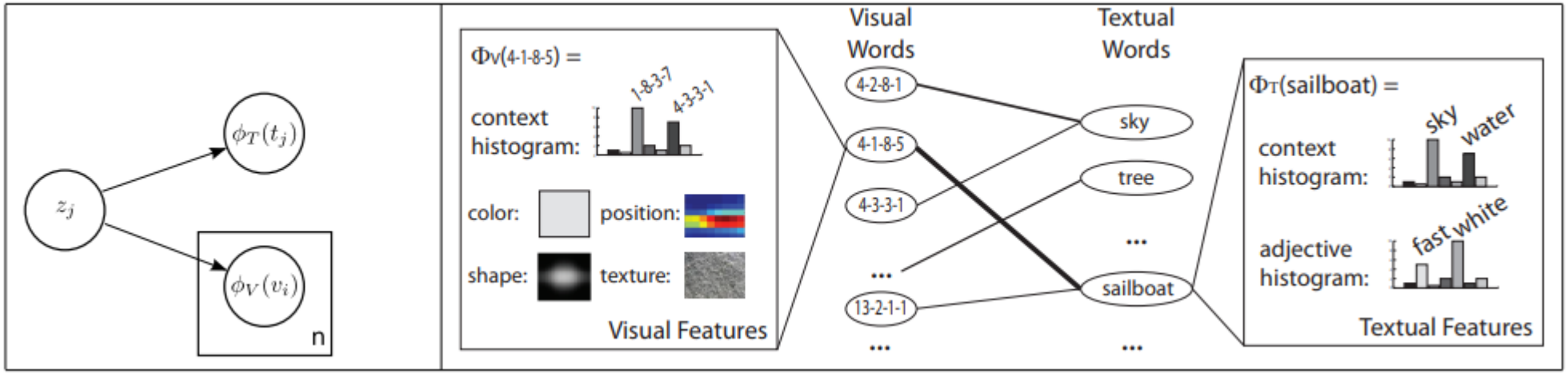
Khapra M M, Kumaran A, Bhattacharyya P, et al. Everybody loves a rich cousin: An empirical study of transliteration through bridge languages[C]. north american chapter of the association for computational linguistics, 2010: 420-428.

R. Socher and L. Fei-Fei, “Connecting modalities: Semi-supervised segmentation and annotation of images using unalignedtext corpora,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,2010, pp. 966–973

Hybrid Data







Khapra M M, Kumaran A, Bhattacharyya P, et al. Everybody loves a rich cousin: An empirical study of transliteration through bridge languages[C]. north american chapter of the association for computational linguistics, 2010: 420-428.

R. Socher and L. Fei-Fei, “Connecting modalities: Semi-supervised segmentation and annotation of images using unalignedtext corpora,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,2010, pp. 966–973

谢谢

THE END