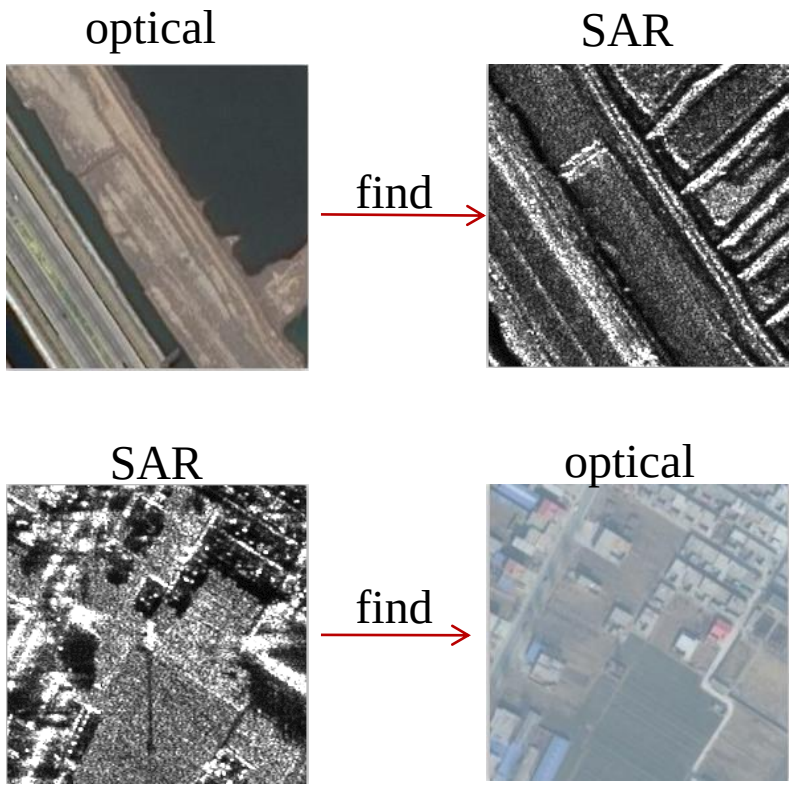
跨模态检索

汇报人：张一帆

问题定义

- 相关的多模态数据：对于同一“事物”，通过不同的方式和角度采集到的数据。跨模态检索：即用一种模态的数据去寻找相关的其他模态的数据。

同种媒体



不同种媒体 image-text

wiki dataset

Martin Luther King's presence in Birmingham was not welcomed by all in the black community. A black attorney was quoted in "Time" magazine as saying, "The new administration should have been given a chance to confer with the various groups interested in change." Black hotel owner A. G. Gaston stated, "I regret the absence of continued communication between white and Negro leadership in our city." A white Jesuit priest assisting in desegregation negotiations attested, "These demonstrations are poorly timed and misdirected." Protest organizers knew they would meet with violence from the Birmingham Police Department but chose a confrontational approach to get the attention of the federal government. Reverend Wyatt Tee Walker, one of the SCLC founders and the executive director from 19601964, planned the tactics of the direct action protests, speci↓cally targeting Bull Connor's tendency to react to demonstrations with violence. "My theory was that if we mounted a strong nonviolent movement, the opposition would surely do something to attract the media, and in turn induce national sympathy and attention to the everyday segregated circumstance of a person living in the Deep South," Walker said. He headed the planning of what he called Project C, which stood for "confrontation". According to this historians Isserman and Kazin, the demands on the city authorities were straightforward: desegregate the economic life of Birmingham its restaurants, hotels, public toilets, and the unwritten policy of hiring blacks for menial jobs only Maurice Isserman and Michael Kazin, America Divided: The Civil War of the 1960s, (Oxford, 2008), p.90. Organizers believed their phones were tapped, so to prevent their plans from being leaked and perhaps in'uencing the mayoral election, they used code words for demonstrations. The plan called for direct nonviolent action to attract media attention to "the biggest and baddest city of the South".Hampton, p. 126. In preparation for the protests, Walker timed the walking distance from the Sixteenth Street Baptist Church, headquarters for the campaign, to the downtown area. He surveyed the segregated lunch counters of department stores, and listed federal buildings as secondary targets should police block the protesters' entrance into primary targets such as stores, libraries, and all-white churches.

Flickr30k dataset

Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

研究意义

- 当今社会是信息技术高速发展的时代，无论在网络还是在新兴的社交媒体（如微信、facebook、快手等），都涌现着大量的多媒体数据，如何高效的利用这些多模态数据，不光是个人，公司的需求，也是国家发展战略的需要。因为仅从单一模式的数据来进行分析，带来的信息还是太少，不能满足日益增长的需要。

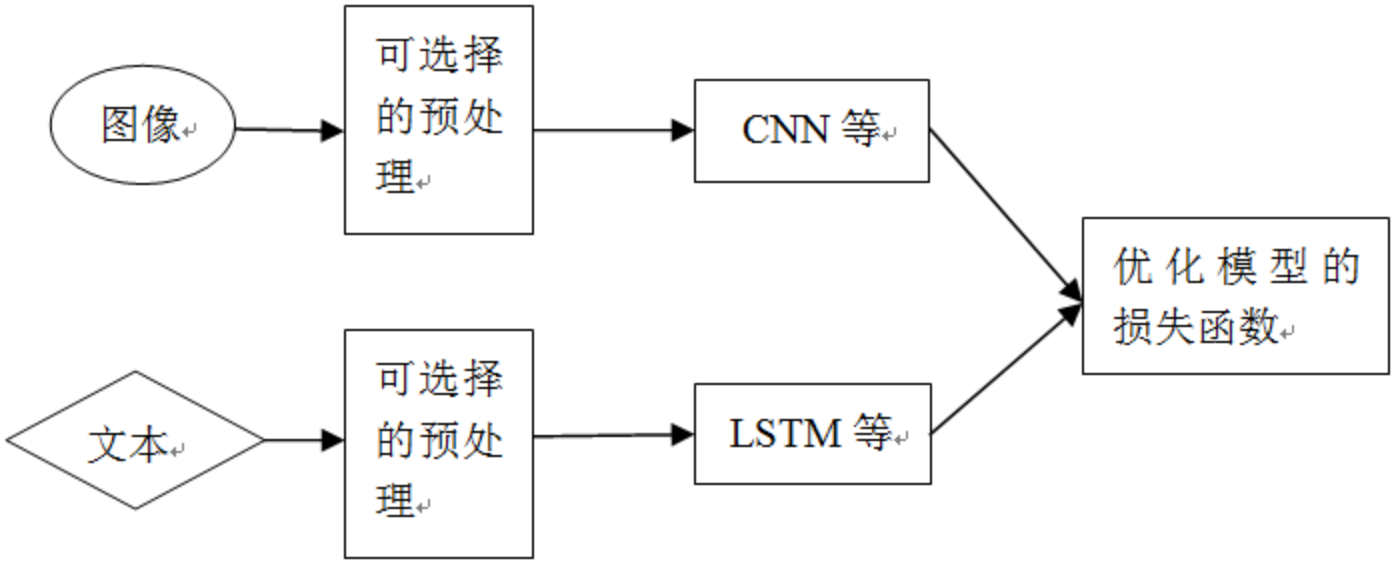
- 1. 多模态信息可以丰富我们的感知。
- 2. 有着丰富的应用场景，如应用图像与文本检索处理视频监控，淘宝物品搜索等问题。
- 3. 有效的跨模态检索有助于改善其他任务性能，如图像配准，目标检测等等。

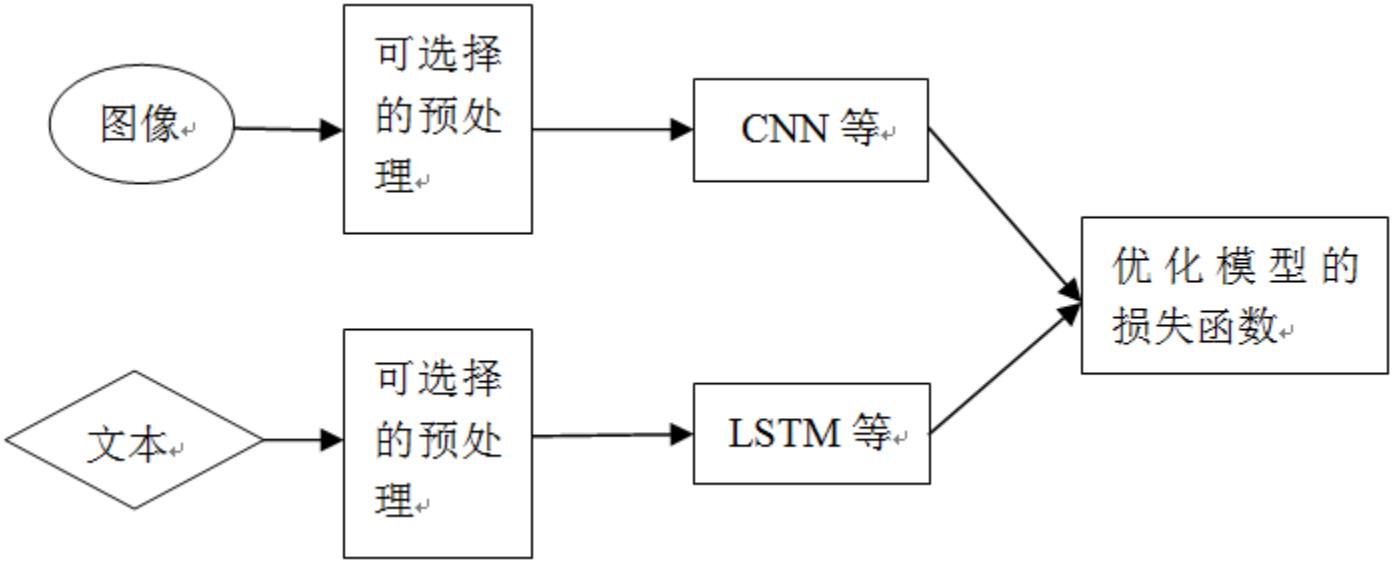
- 因为在实际应用中，跨模态检索是一个重要的问题，所以大量的方法被提出。

- 按是否有语义标签分为：有监督和无监督方法
- 按生成的表达是否二值化分为：基于实值的表达学习和基于二值化的表达学习（跨模态哈希）
- 按是否应用深度网络分为：传统方法和基于深度学习的方法
- ……

Common space learning

我们处理跨模态检索问题，是把不同模态的数据映射到high-level的公共特征空间进行比较，完成检索。

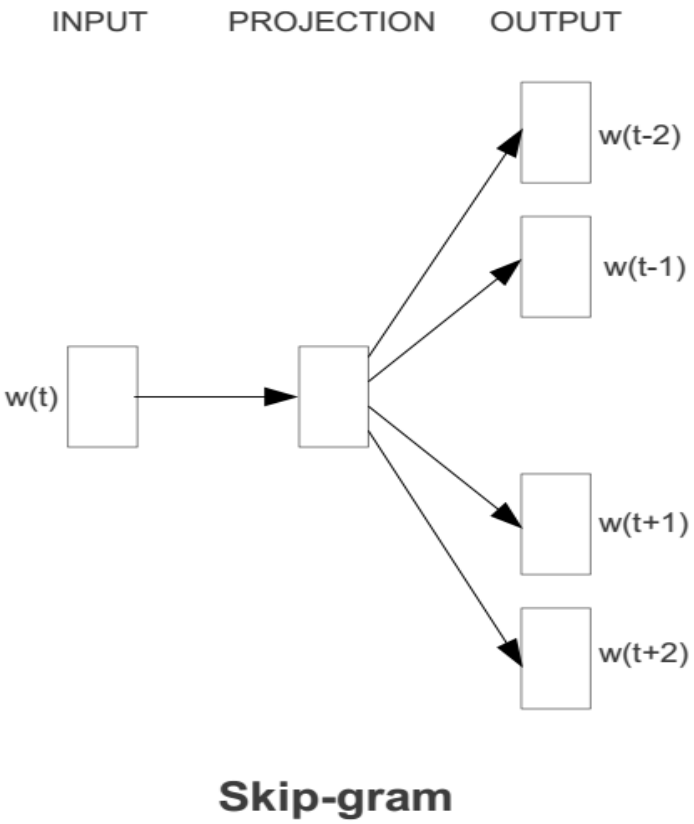


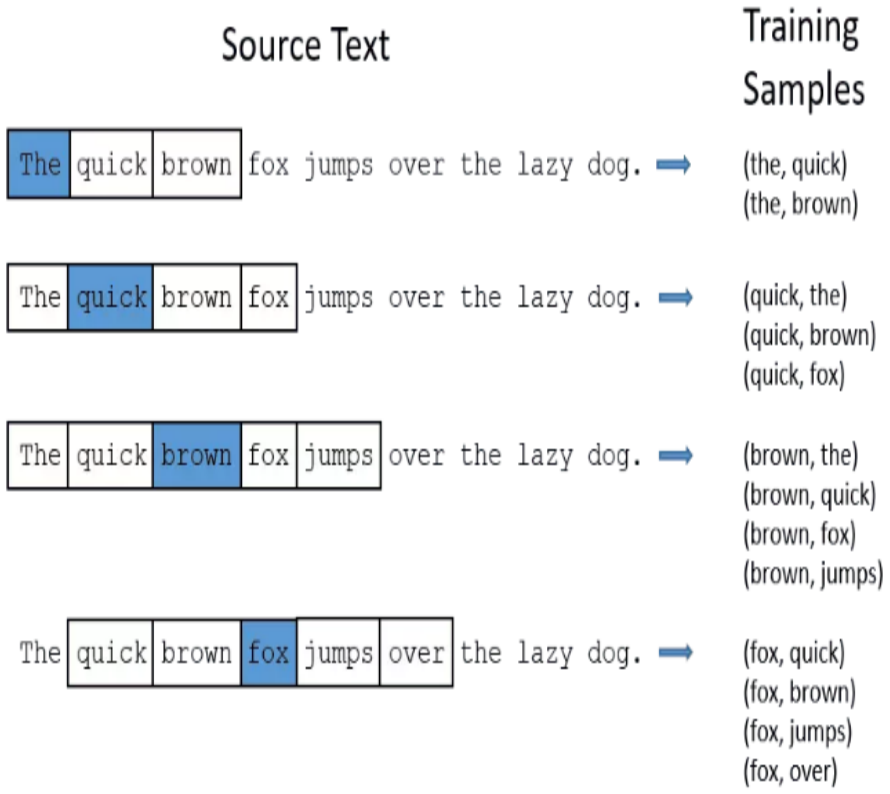


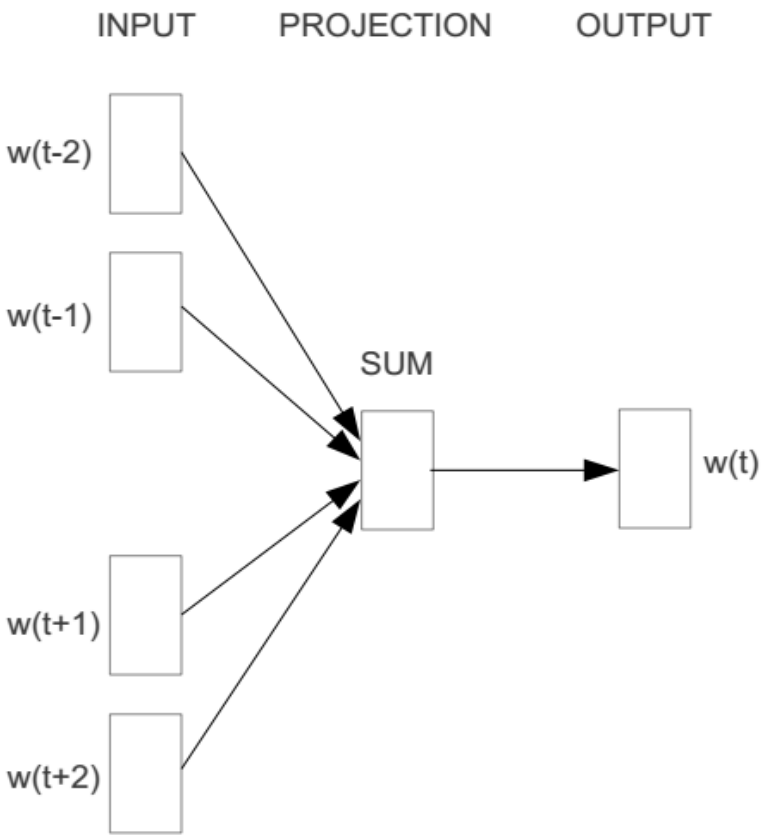
- 关于图像预处理：SIFT特征（BOW），颜色直方图，边缘方向直方图，图像经过预训练的网络的特征等。
- 关于文本预处理：词频特征，潜在狄利克雷分布，word2vec等。

Efficient Estimation of Word Representations in Vector Space

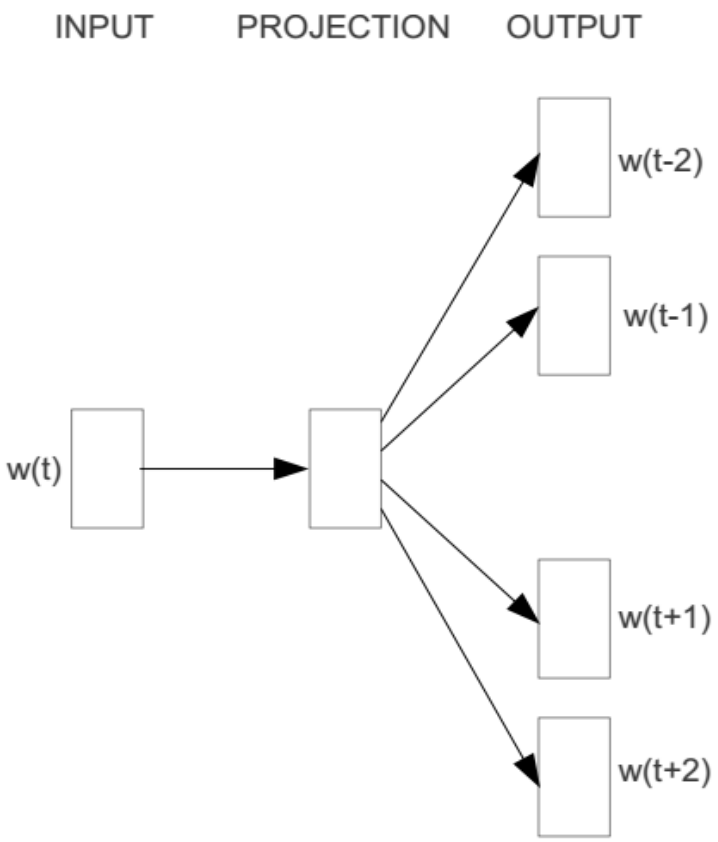
The quick brown fox jumps over the lazy dog.





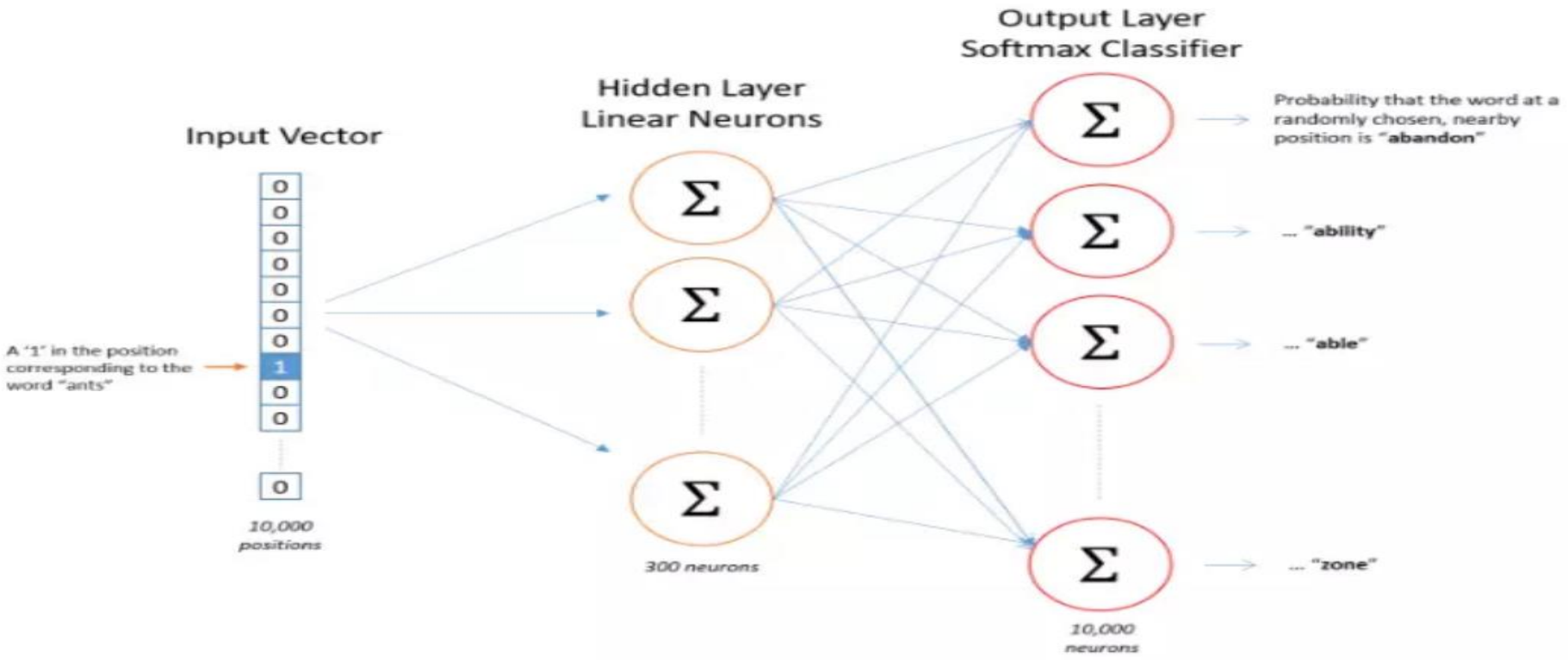


CBOW

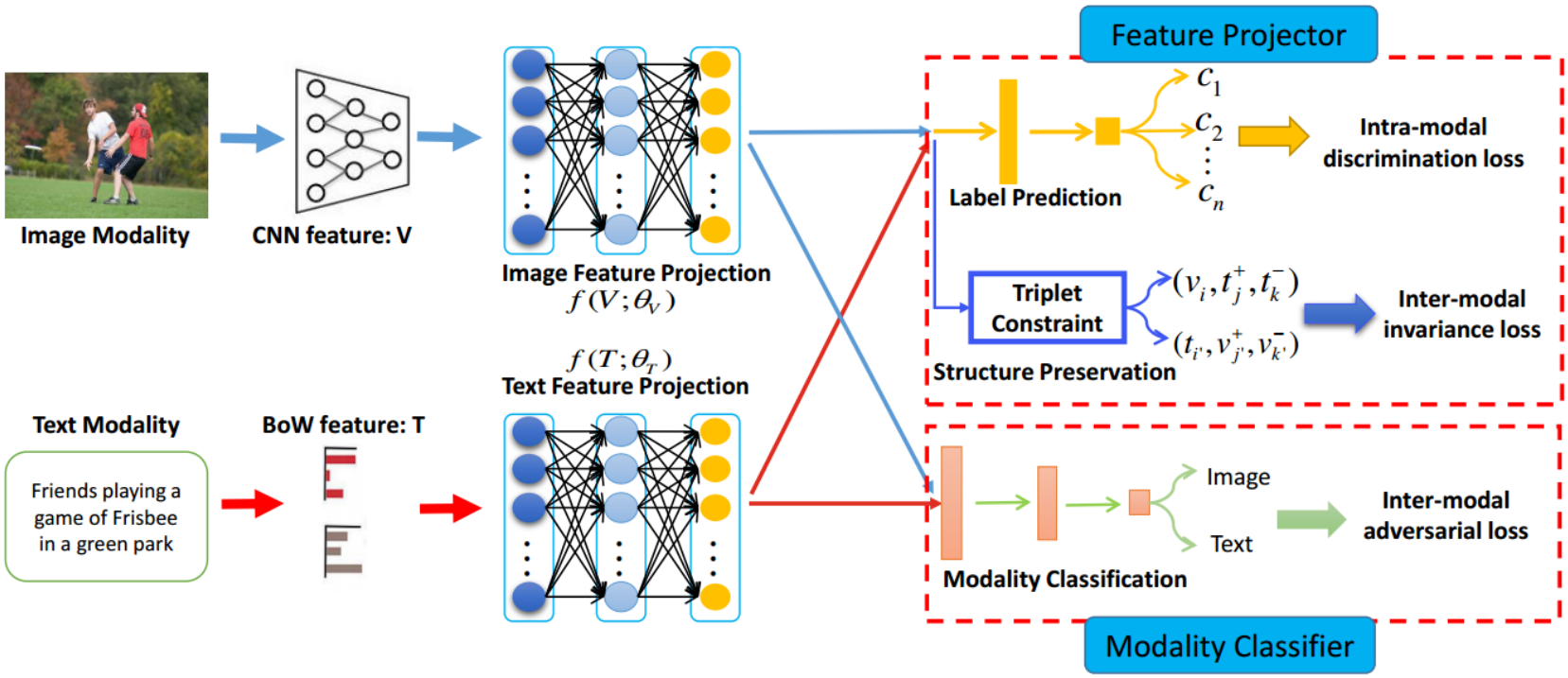


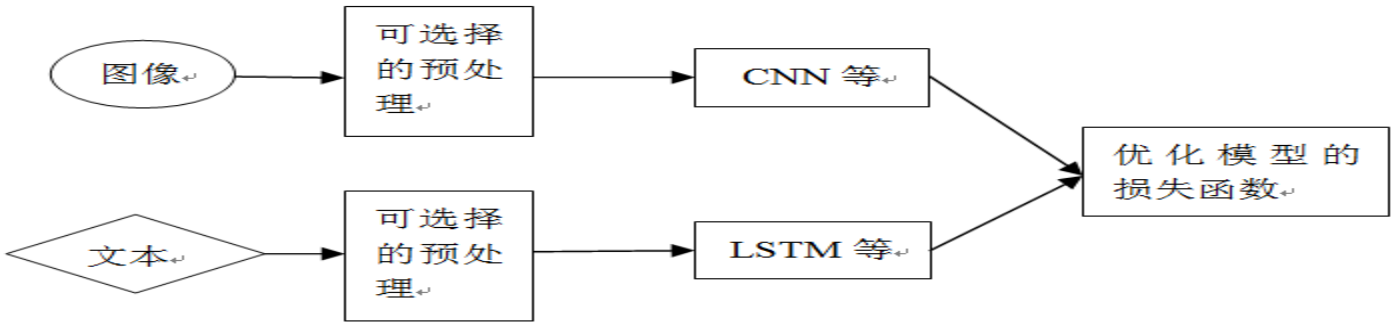
Skip-gram

	The	Quick	Brown	Fox	Jumps	Over	Lazy	Dog
The	1	0	0	0	0	0	0	0
Quick	0	1	0	0	0	0	0	0
Brown	0	0	1	0	0	0	0	0
Fox	0	0	0	1	0	0	0	0
Jumps	0	0	0	0	1	0	0	0
Over	0	0	0	0	0	1	0	0
Lazy	0	0	0	0	0	0	1	0
Dog	0	0	0	0	0	0	0	1



Adversarial Cross-Modal Retrieval



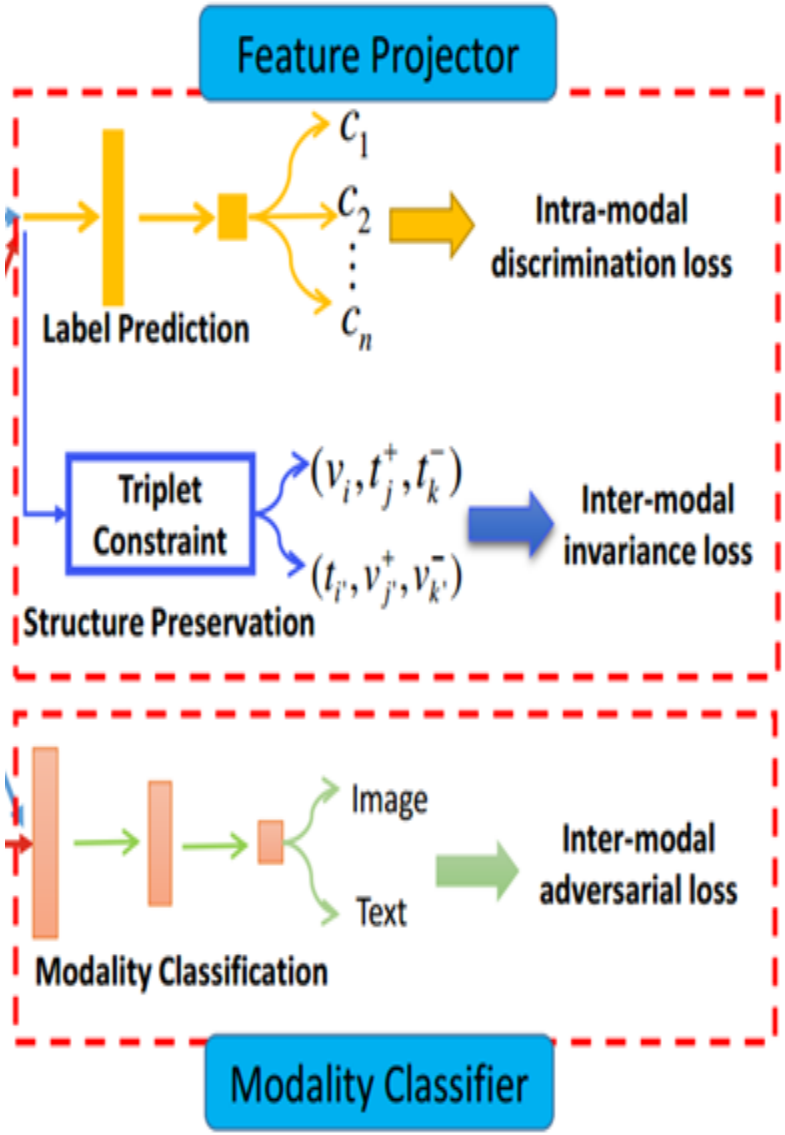


可选择的预处理

Table 1: General statistics of the four datasets used in our experiments, where "*/*" in the "Instances" column stands for the number of training/test image-text pairs.

Dataset	Instances	Labels	Image feature	Text feature
Wikipedia	1,300/1,566	10	128d SIFT 4,096d VGG	10d LDA 3,000d BoW
Pascal Sentence	800/200	20	4,096d VGG	1,000d BoW
NUS-WIDE-10K	8,000/1,000	350	4,096d VGG	1,000d BoW
MSCOCO	66,226/16,557	500	4,096d VGG	3,000d BoW

图像特征来自于VGG的fc7后的特征，文本特征来自于BOW矢量伴随着TF-IDF加权



$$\mathcal{L}_{imd}(\theta_{imd}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot (\log \hat{p}_i(v_i) + \log \hat{p}_i(t_i))).$$

$$\mathcal{L}_{imi,\mathcal{V}}(\theta_{\mathcal{V}}) = \sum_{i,j,k}(\ell_2(v_i, t_j^+) + \lambda \cdot \max(0, \mu - \ell_2(v_i, t_k^-)),$$

$$\mathcal{L}_{imi,\mathcal{T}}(\theta_{\mathcal{T}}) = \sum_{i,j,k}(\ell_2(t_i, v_j^+) + \lambda \cdot \max(0, \mu - \ell_2(t_i, v_k^-)).$$

$$\mathcal{L}_{adv}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{m_i} \cdot (\log D(v_i; \theta_D) + \log(\mathbf{1} - D(t_i; \theta_D))).$$

$$(\hat{\theta}_{\mathcal{V}}, \hat{\theta}_{\mathcal{T}}, \hat{\theta}_{imd}) = \arg\min_{\theta_{\mathcal{V}}, \theta_{\mathcal{T}}, \theta_{imd}} (\mathcal{L}_{emb}(\theta_{\mathcal{V}}, \theta_{\mathcal{T}}, \theta_{imd}) - \mathcal{L}_{adv}(\hat{\theta}_D)),$$

$$\hat{\theta}_D = \arg\max_{\theta_D} (\mathcal{L}_{emb}(\hat{\theta}_{\mathcal{V}}, \hat{\theta}_{\mathcal{T}}, \hat{\theta}_{imd}) - \mathcal{L}_{adv}(\theta_D)).$$

Algorithm 1 Pseudocode of optimizing our ACMR.

Initialization: Image features for current batch: $\mathcal{V} = \{v_1, ..., v_n\}$;
Text features for current batch, $\mathcal{T} = \{t_1, ..., t_n\}$;
Corresponding labels for current batch, $\mathcal{Y} = \{y_1, ..., y_n\}$;
hyperparameters: k, λ, α, β;
m samples in minibatch for each modality;

update until convergence:

1: **for** k steps **do**
2: update parameters $\theta_{\mathcal{V}}$, $\theta_{\mathcal{T}}$ and θ_{imd} by **descending** their stochastic gradients:
3: $\theta_{\mathcal{V}} \leftarrow \theta_{\mathcal{V}} - \mu \cdot \nabla_{\theta_{\mathcal{V}}} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
4: $\theta_{\mathcal{T}} \leftarrow \theta_{\mathcal{T}} - \mu \cdot \nabla_{\theta_{\mathcal{T}}} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
5: $\theta_{imd} \leftarrow \theta_{imd} - \mu \cdot \nabla_{\theta_{imd}} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
6: **end for**
7: update parameters of modality classifier by **ascending** its stochastic gradeints through Gradient Reversal Layer:
8: $\theta_D \leftarrow \theta_D + \mu \cdot \lambda \cdot \nabla_{\theta_D} \frac{1}{m}(\mathcal{L}_{emb} - \mathcal{L}_{adv})$
9: **return** learned representations in common subspace: $f_{\mathcal{V}}(\mathcal{V})$ and $f_{\mathcal{T}}(\mathcal{T})$

Methods	Shallow feature			Deep feature		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA	0.255	0.185	0.220	0.267	0.222	0.245
Multimodal DBN	0.149	0.150	0.150	0.204	0.183	0.194
Bimodal-AE	0.236	0.208	0.222	0.314	0.290	0.302
CCA-3V	0.275	0.224	0.249	0.437	0.383	0.410
LCFS	0.279	0.214	0.246	0.455	0.398	0.427
Corr-AE	0.280	0.242	0.261	0.402	0.395	0.398
JRL	0.344	0.277	0.311	0.453	0.400	0.426
JFSSL	0.306	0.228	0.267	0.428	0.396	0.412
CMDN	-	-	-	0.488	0.427	0.458
ACMR (Proposed)	**0.366**	**0.277**	**0.322**	**0.619**	**0.489**	**0.546**

Table 2: Comparison of the cross-modal retrieval performance on the Wikipedia dataset. Here, "-" denotes that no experimental results with same settings are available.

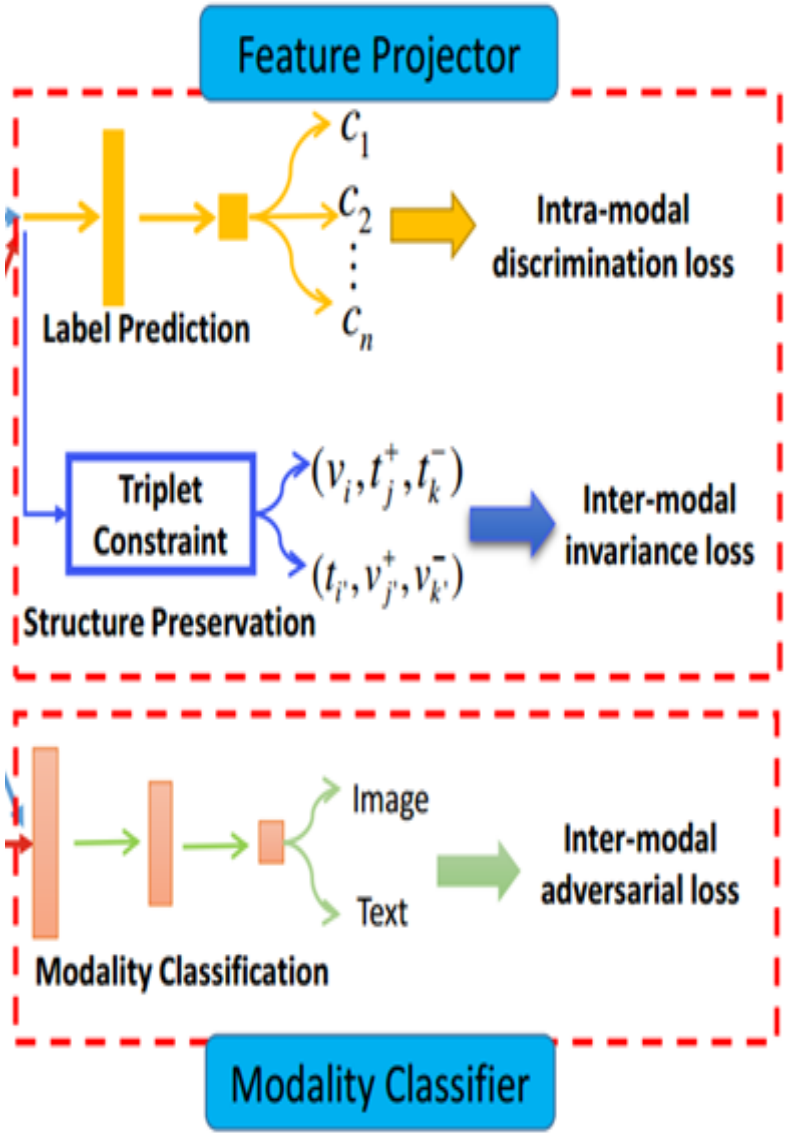
评价指标：mAP，PR曲线

Methods	Pascal Sentences			NUSWIDE-10k		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA	0.363	0.219	0.291	0.189	0.188	0.189
Multimodal DBN	0.477	0.424	0.451	0.201	0.259	0.230
Bimodal-AE	0.456	0.470	0.458	0.327	0.369	0.348
LCFS	0.442	0.357	0.400	0.383	0.346	0.365
Corr-AE	0.489	0.444	0.467	0.366	0.417	0.392
JRL	0.504	0.489	0.496	0.426	0.376	0.401
CMDN	0.534	0.534	0.534	0.492	0.515	0.504
ACMR (Proposed)	**0.535**	**0.543**	**0.539**	**0.544**	**0.538**	**0.541**

Table 3: Cross-modal retrieval comparison in terms of mAP on Pascal Sentences and NUSWIDE-10k dataset.

Methods	Img2Txt	Txt2Img	Avg.
CCA (FV HGLMM) [14]	0.791	0.765	0.778
CCA (FV GMM+HGLM) [14]	0.809	0.766	0.788
DVSA [12]	0.805	0.748	0.777
m-RNN [19]	0.835	0.770	0.803
m-CNN [18]	0.841	0.828	0.835
DSPE [33]	0.892	0.869	0.881
ACMR (Proposed)	**0.932**	**0.871**	**0.902**

Table 4: Cross-modal retrieval comparison in terms of mAP on MSCOCO dataset.



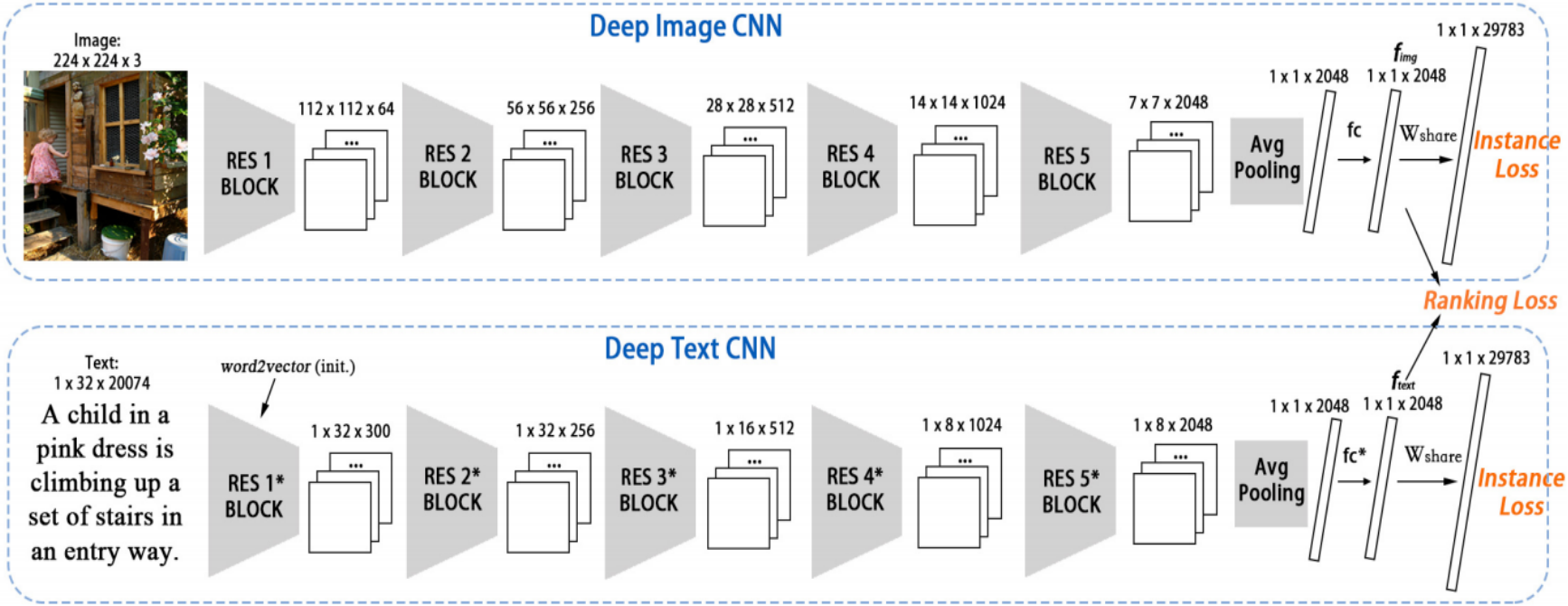
Methods	Wikipedia			Pascal Sentences		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
ACMR (with $\mathcal{L}_{imi}$ only)	0.352	0.430	0.391	0.289	0.274	0.282
ACMR (with $\mathcal{L}_{imd}$ only)	0.425	0.413	0.419	0.533	0.453	0.493
Full ACMR	**0.509**	**0.431**	**0.470**	**0.535**	**0.486**	**0.511**

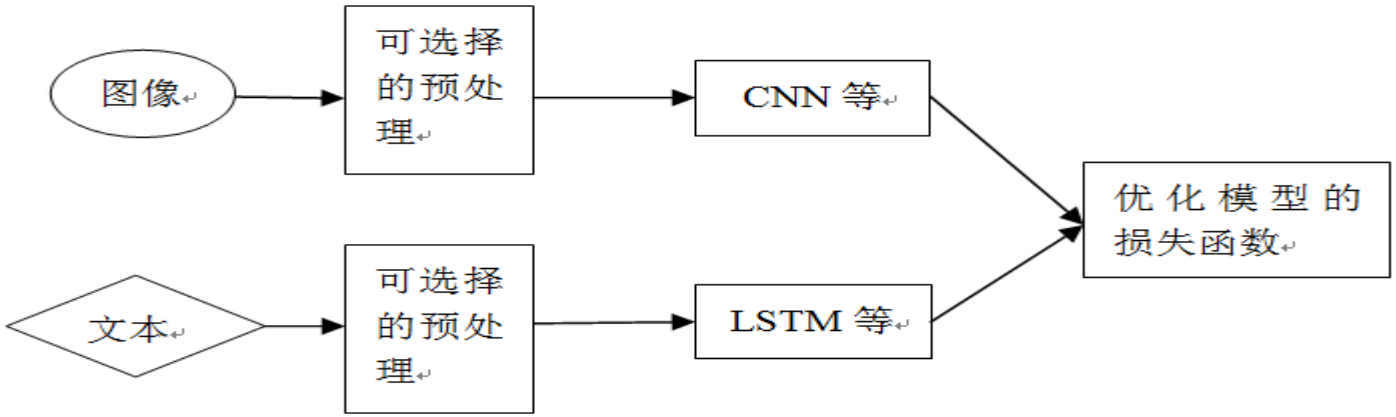
Table 5: Performance of cross-modal retrieval using the proposed ACMR method, ACMR method with $\mathcal{L}_{imi}$ only, and ACMR method with $\mathcal{L}_{imd}$ only.

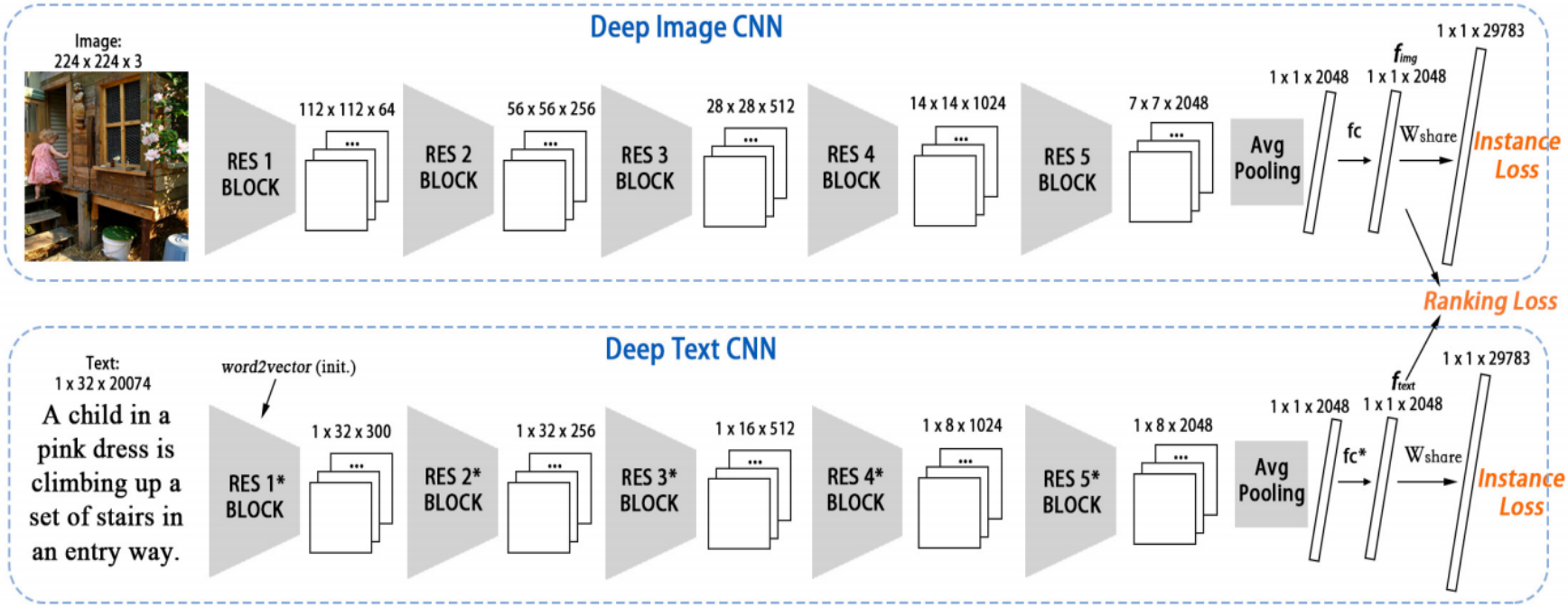
L_{imd}表示intra-modal，L_{imi}表示inter-modal

跨模态检索的趋势：由粗粒度到细粒度的一种转变

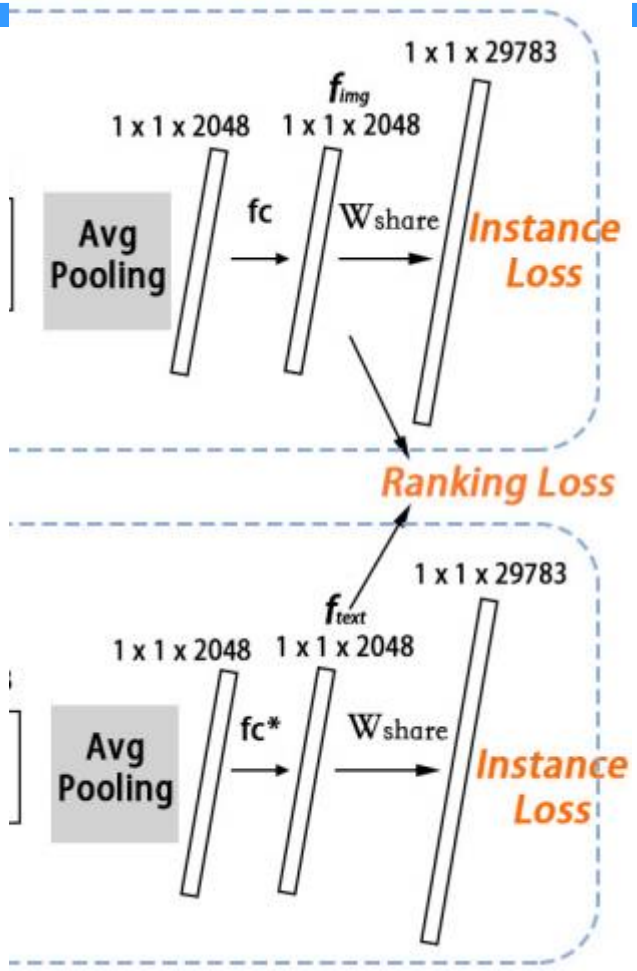
Dual-Path Convolutional Image-Text Embeddings with Instance Loss







处理图像是一个在ImageNet上预训练的ResNet-50, 处理文本是一个类似ResNet-50的残差块。



$$L_{rank} = \overbrace{max[0, \alpha - (D(f_{I_a}, f_{T_a}) - D(f_{I_a}, f_{T_n}))]}^{image\ anchor} + \underbrace{max[0, \alpha - (D(f_{T_a}, f_{I_a}) - D(f_{T_a}, f_{I_n}))]}_{text\ anchor},$$

$$P_{visual} = softmax(W_{share}^T f_{img}),$$
$$L_{visual} = -\log(P_{visual}(c)),$$
$$P_{textual} = softmax(W_{share}^T f_{text}),$$
$$L_{textual} = -\log(P_{text}(c)),$$

- 训练：
- 1. 固定图像支路CNN部分，用instance loss训练剩余部分，能够提供更好的初始化，学到fine-grained的特征。
- 2. 端到端的训练整个网络。

- 在Flickr30k, MSCOCO, CUHL-PEDES上进行了评测

Method	Visual	Textual	Image Query				Text Query			
			R@1	R@5	R@10	Med	R@1	R@5	R@10	Med r
DeVise [5]	ft AlexNet	ft skip-gram	4.5	18.1	29.2	26	6.7	21.9	32.7	25
Deep Fragment [6]	ft RCNN	fixed word vector from [58]	16.4	40.2	54.7	8	10.3	31.4	44.5	13
DCCA [59]	ft AlexNet	TF-IDF	16.7	39.3	52.9	8	12.6	31.0	43.0	15
DVSA [32]	ft RCNN (init. on Detection)	w2v + ft RNN	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
LRCN [60]	ft VGG-16	ft RNN	23.6	46.6	58.3	7	17.5	40.3	50.8	9
m-CNN [7]	ft VGG-19	4 × ft CNN	33.6	64.1	74.9	3	26.2	56.3	69.6	4
VQA-A [18]	fixed VGG-19	ft RNN	33.9	62.5	74.5	-	24.9	52.6	64.8	-
GMM-FV [17]	fixed VGG-16	w2v + GMM + HGLMM	35.0	62.0	73.8	3	25.0	52.7	66.0	5
m-RNN [16]	fixed VGG-16	ft RNN	35.4	63.8	73.7	3	22.8	50.7	63.1	5
RNN-FV [19]	fixed VGG-19	feature from [17]	35.6	62.5	74.2	3	27.4	55.9	70.0	4
HM-LSTM [21]	fixed RCNN from [32]	w2v + ft RNN	38.1	-	76.5	3	27.7	-	68.8	4
SPE [8]	fixed VGG-19	w2v + HGLMM	40.3	68.9	79.9	-	29.7	60.1	72.1	-
sm-LSTM [20]	fixed VGG-19	ft RNN	42.5	71.9	81.5	2	30.2	60.4	72.3	3
RRF-Net [61]	fixed ResNet-152	w2v + HGLMM	47.6	77.4	87.1	-	35.4	68.3	79.9	-
2WayNet [49]	fixed VGG-16	feature from [17]	49.8	67.5	-	-	36.0	55.6	-	-
DAN (VGG-19) [9]	fixed VGG-19	ft RNN	41.4	73.5	82.5	2	31.8	61.7	72.5	3
DAN (ResNet-152) [9]	fixed ResNet-152	ft RNN	55.0	81.8	89.0	1	**39.4**	69.2	79.1	2
Ours (VGG-19) Stage I	fixed VGG-19	ft ResNet-50† (w2v init.)	37.5	66.0	75.6	3	27.2	55.4	67.6	4
Ours (VGG-19) Stage II	ft VGG-19	ft ResNet-50† (w2v init.)	47.6	77.3	87.1	2	35.3	66.6	78.2	3
Ours (ResNet-50) Stage I	fixed ResNet-50	ft ResNet-50† (w2v init.)	41.2	69.7	78.9	2	28.6	56.2	67.8	4
Ours (ResNet-50) Stage II	ft ResNet-50	ft ResNet-50† (w2v init.)	53.9	80.9	**89.9**	**1**	39.2	**69.8**	80.8	**2**
Ours (ResNet-152) Stage I	fixed ResNet-152	ft ResNet-152† (w2v init.)	44.2	70.2	79.7	2	30.7	59.2	70.8	4
Ours (ResNet-152) Stage II	ft ResNet-152	ft ResNet-152† (w2v init.)	**55.6**	**81.9**	89.5	**1**	39.1	69.2	**80.9**	**2**

训练阶段loss

Method	Stage	Image Query		Text Query	
		R@1	R@10	R@1	R@10
Only Ranking Loss	I	6.1	27.3	4.9	27.8
Only Instance Loss	I	39.9	79.1	28.2	67.9
Only Instance Loss	II	50.5	86.0	34.9	75.7
Only Ranking Loss	II	47.5	85.4	29.0	68.7
Full model	II	55.4	89.3	39.7	80.8

类别数目

Methods	Image-Query R@1	Text-Query R@1
3000 categories (StageI)	38.0	26.1
10000 categories (StageI)	44.7	31.3
Our (StageI)	52.2	37.2

Position shift

Method	Image Query		Text Query	
	R@1	R@10	R@1	R@10
Left alignment	34.1	73.1	23.6	61.4
Position shift	39.9	79.1	28.2	67.9

是否word2vec初始化

Method	Image Query		Text Query	
	R@1	R@10	R@1	R@10
Random initialization [52]	38.0	78.7	26.6	66.6
Word2vec initialization	39.9	79.1	28.2	67.9

VSE++: Improving Visual-Semantic Embeddings with Hard Negatives

$$\ell_{MH}(i,c) = \max_{c'} \left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'} \left[\alpha + s(i',c) - s(i,c)\right]_+ .$$

Thanks!