Exploring and Distilling Cross-Modal Information for Image Captioning

Fenglin Liu[1], Xuancheng Ren[1]*, Yuanxin Liu[2], Kai Lei[1] and Xu Sun[1]

[1] Peking University, China

[2] Beijing University of Posts and Telecommunications, China

* Equal Contributions

CONTENTS

1 Introduction

Exploring and Distilling Cross-Modal Information for Image Captioning

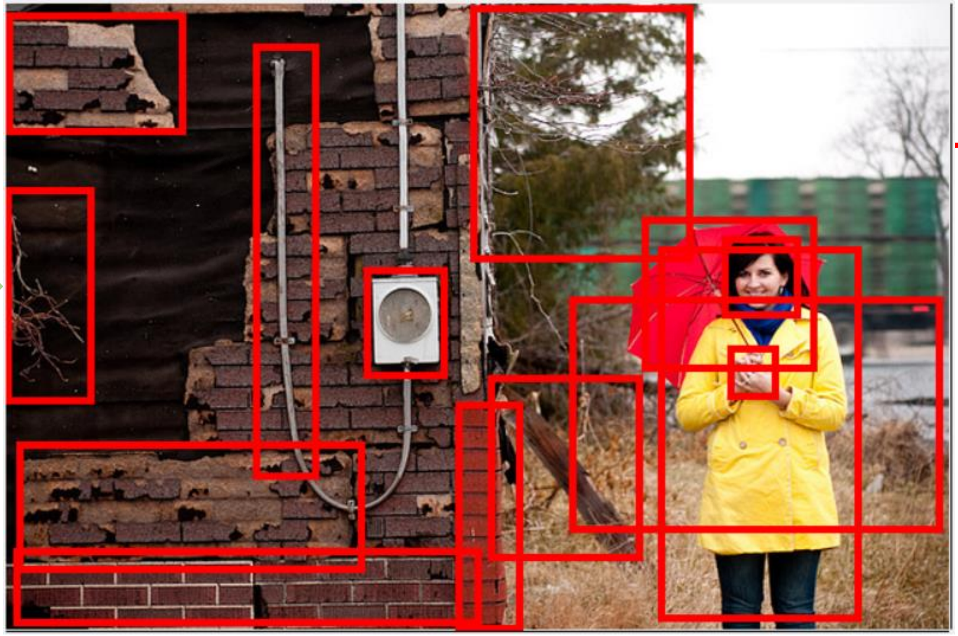
woman, umbrella, holding, yellow, clock, sign, building, standing, street, man, brick, walking, tower, train, her, wall, posing, stop, old, front, girl, bathroom, tennis, person, carrying, young, down, people, wearing, bananas, sitting, tree, toilet, hanging, giraffe, red, cat, white, wooden, field, tall, tracks, surfboard, bench, pole, tie, city, kite, it

(a) (b) (c)

Figure 1: (a) The input image; (b) The extracted object-oriented visual regions; (c) The extracted attribute words.

Introduction: Bottom-Up

(a) The input image;.



(b) The extracted object-oriented visual regions.

All visual region are →unrelated individual parts , and are not guided to comprehend the general correlations of each other.

Bottom-Up: Image $\xrightarrow{\text{encode}}$ Individual Components of Visual Regions $\xrightarrow{\text{decode}}$ Caption

·**Bottom-Up:** Bottom-up and top-down attention for image captioning and VQA . In CVPR 2018

Introduction: ATT-FCN

woman, umbrella, holding, yellow, clock, sign, building, standing, street, man, brick, walking, tower, train, her, wall, posing, stop, old, front, girl, bathroom, tennis, person, carrying, young, down, people, wearing, bananas, sitting, tree, toilet, hanging, giraffe, red, cat, white, wooden, field, tall, tracks, surfboard, bench, pole, tie, city, kite, it

All attribute words are irrelated individual parts.

Do not encode the general correlations of such parts.

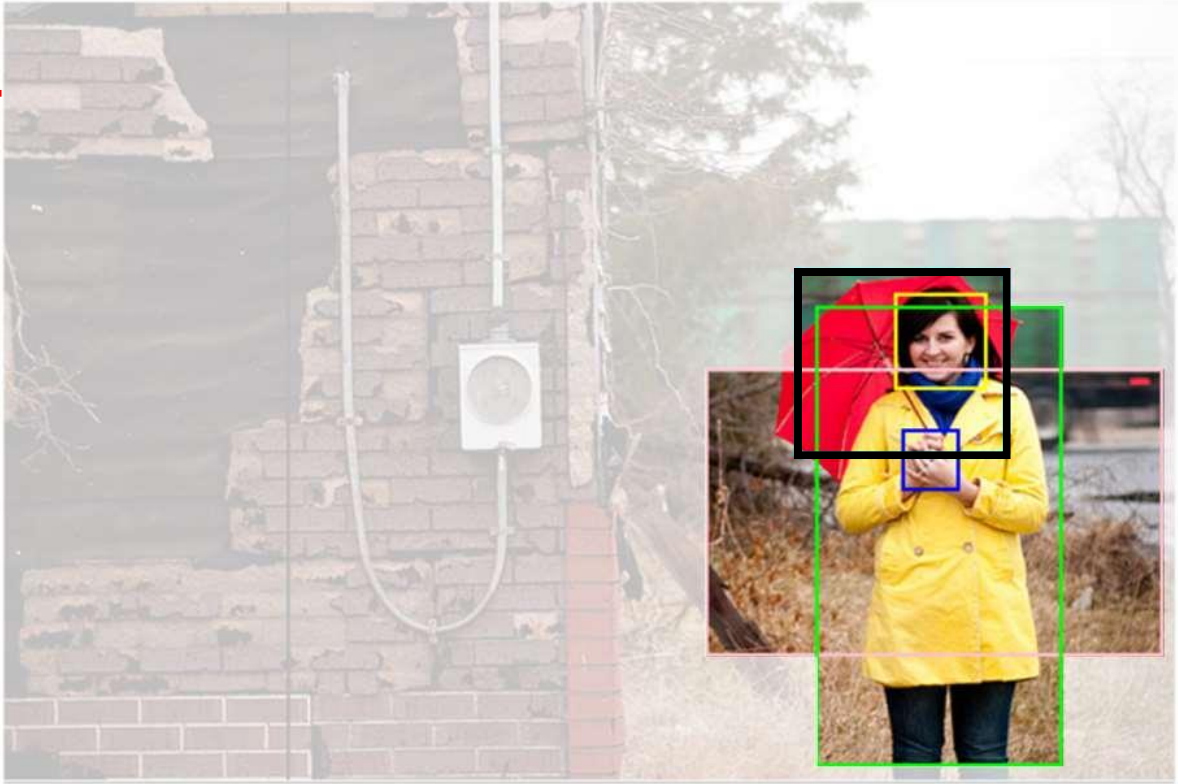
(a) The input image; **(b) The extracted attribute words.**

ATT-FCN: Image $\xrightarrow{\text{encode}}$ Individual Components of Attribute Words $\xrightarrow{\text{decode}}$ Caption

·**ATT-FCN :** Image captioning with semantic attention. In CVPR 2016

Example: Umbrella

The focus on the umbrella is naturally extended to the related areas.



umbrella

woman, umbrella, holding, yellow, clock, sign, building, standing, street, man, brick, walking, tower, train, her, wall, posing, stop, old, front, girl, bathroom, tennis, person, carrying, young, down, people, wearing, bananas, sitting, tree, toilet, hanging, giraffe, red, cat, white, wooden, field, tall, tracks, surfboard, bench, pole, tie, city, kite, it

The input word umbrella is associated with common collocations.

GLIED: Image — encode — Visual Regions (Individual) / Visual Attributes (Individual) — GLIED — Visual Regions (Associated) / Visual Attributes (Associated) — decode — Caption

Contributions

- We propose the **G**lobal-and-**L**ocal **I**nformation **E**xploring-and-**D**istilling (**GLIED**) approach that can globally captures the inherent spatial and relational groupings of the individual image regions and attribute words for an aspect-based image representation, and locally it extracts fine-grained source information for precise and accurate word selection.

- The learned region groupings and attribute collocations are in accordance with human intuition.

- The proposed approach outperforms previous works with fewer parameters and faster computation.

2 Approach

Overview

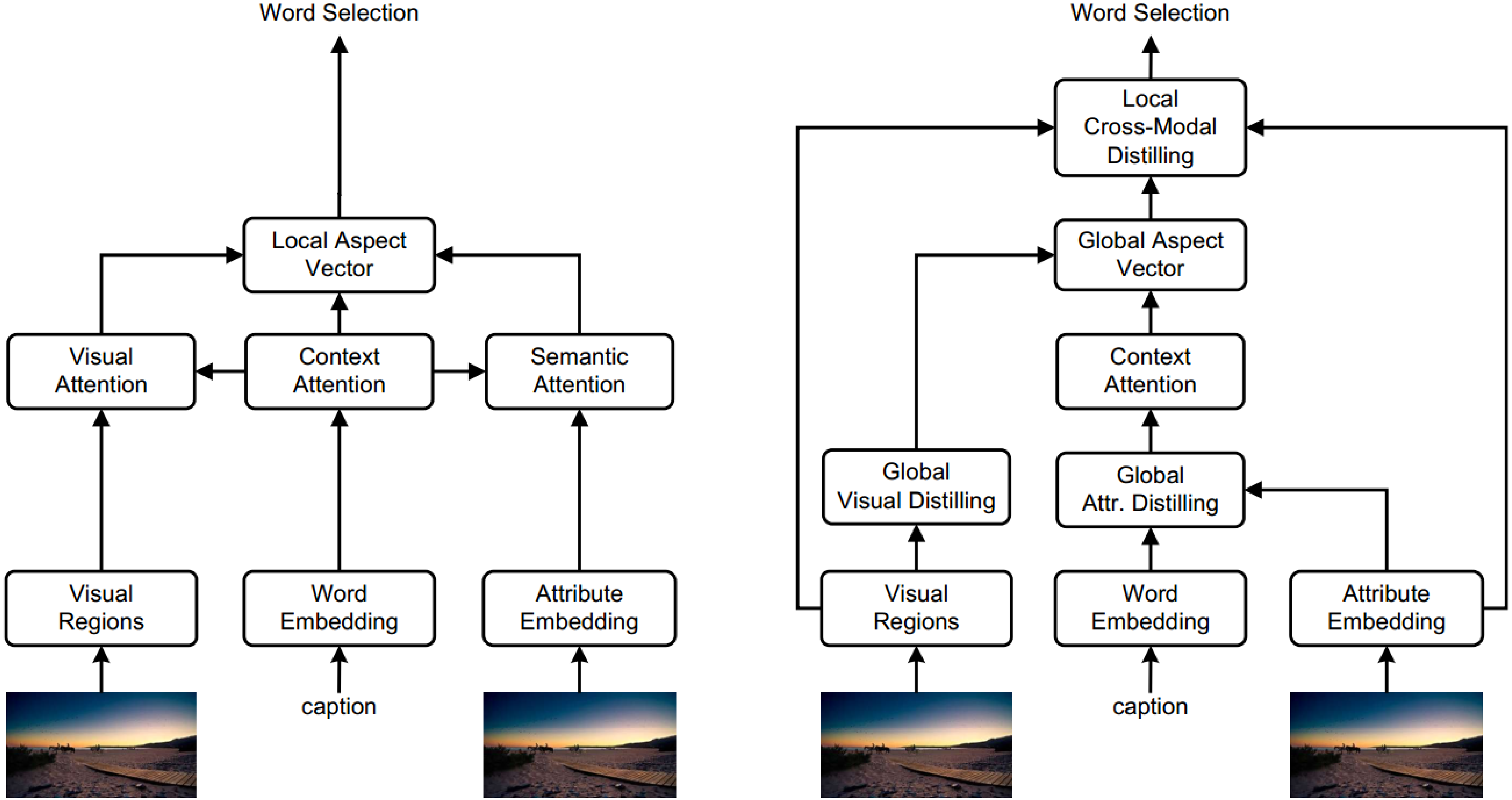
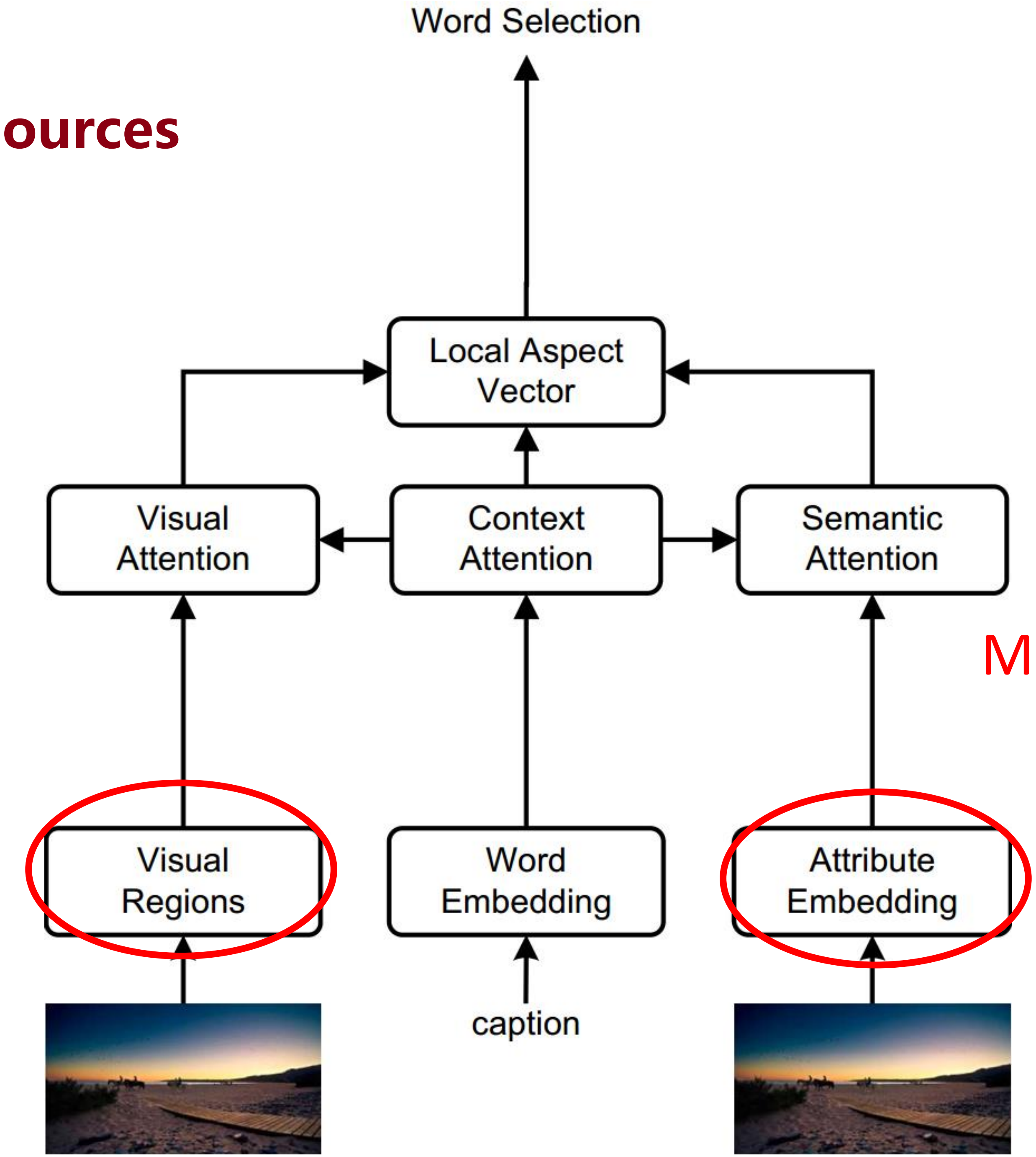


Figure 2: Illustration of the difference between our cross-modal fully-attentive base model (Left) and the proposed model that distills the source information both globally and locally (Right).

Information Sources



Visual Region Extractor:

Faster-RCNN

Attribute Word Extractor:

Multiple Instance Learning

Ren et al., 2015: Faster R-CNN: towards real-time object detection with region proposal networks . In NeuralPS 2015.

Anderson et al., 2018: Bottom-up and top-down attention for image captioning and VQA . In CVPR 2018.

Zhang et al., 2006: Multiple instance boosting for object detection. In NIPS 2006.

Fang et al., 2015: From captions to visual concepts and back. In CVPR2015

Background: Multi-Head Attention

Scaled Dot-Product Attention:

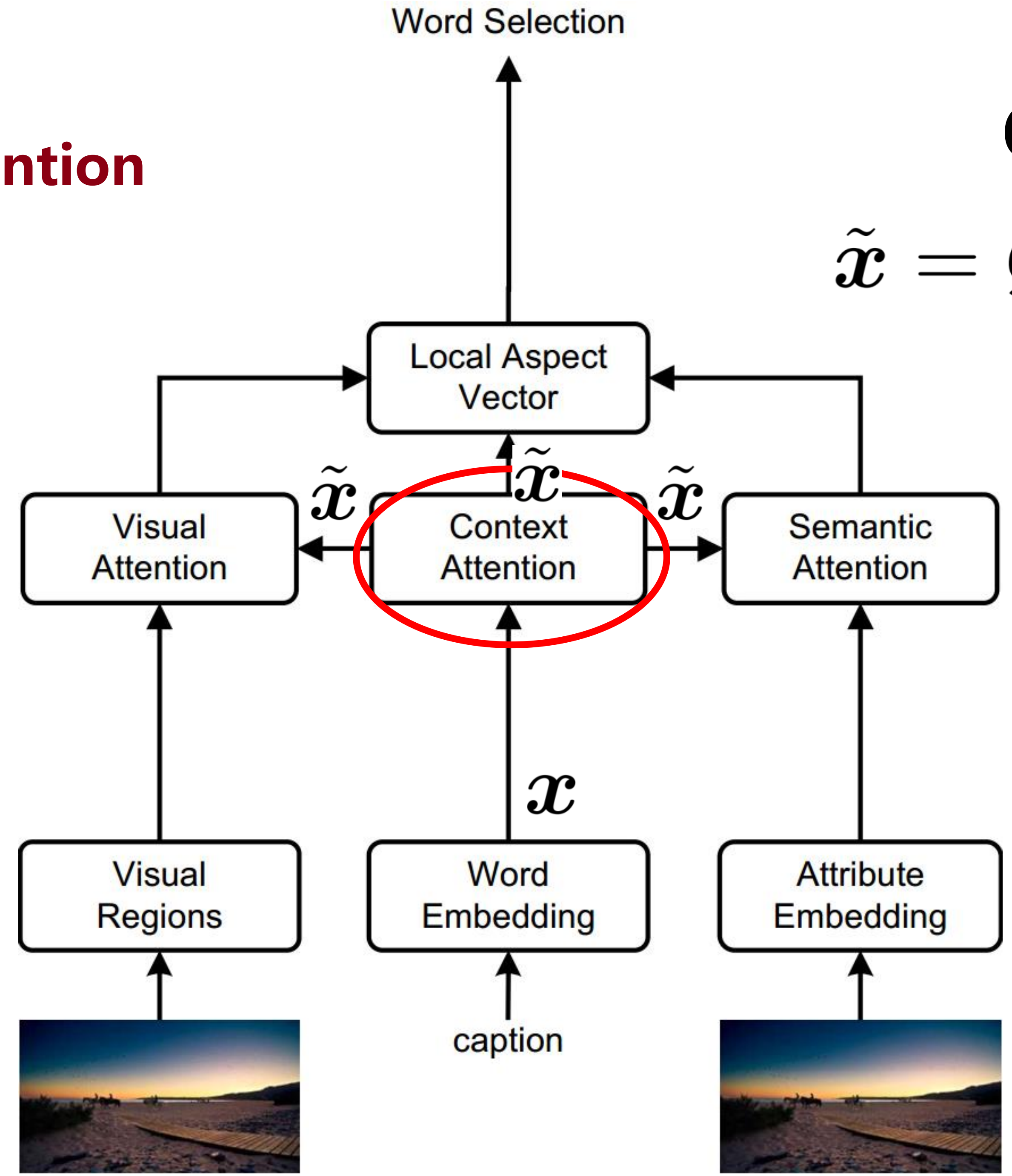
$$\mathcal{A}(Q, K, V)_i = \mathrm{softmax}\left(\frac{QW_i^Q (KW_i^K)^\top)}{\sqrt{d_k}}\right) VW_i^V$$

Multi-Head Attention:

$$\mathcal{H}(Q, K, V) = [\mathcal{A}_1; \mathcal{A}_2; \ldots; \mathcal{A}_k] W_k$$

The multi-head attention is followed by a series of operations of shortcut connection, dropout, and layer normalization, which we denote as function **G(·; ∗)**, where $*$ is the input.

Base Model: Context Attention



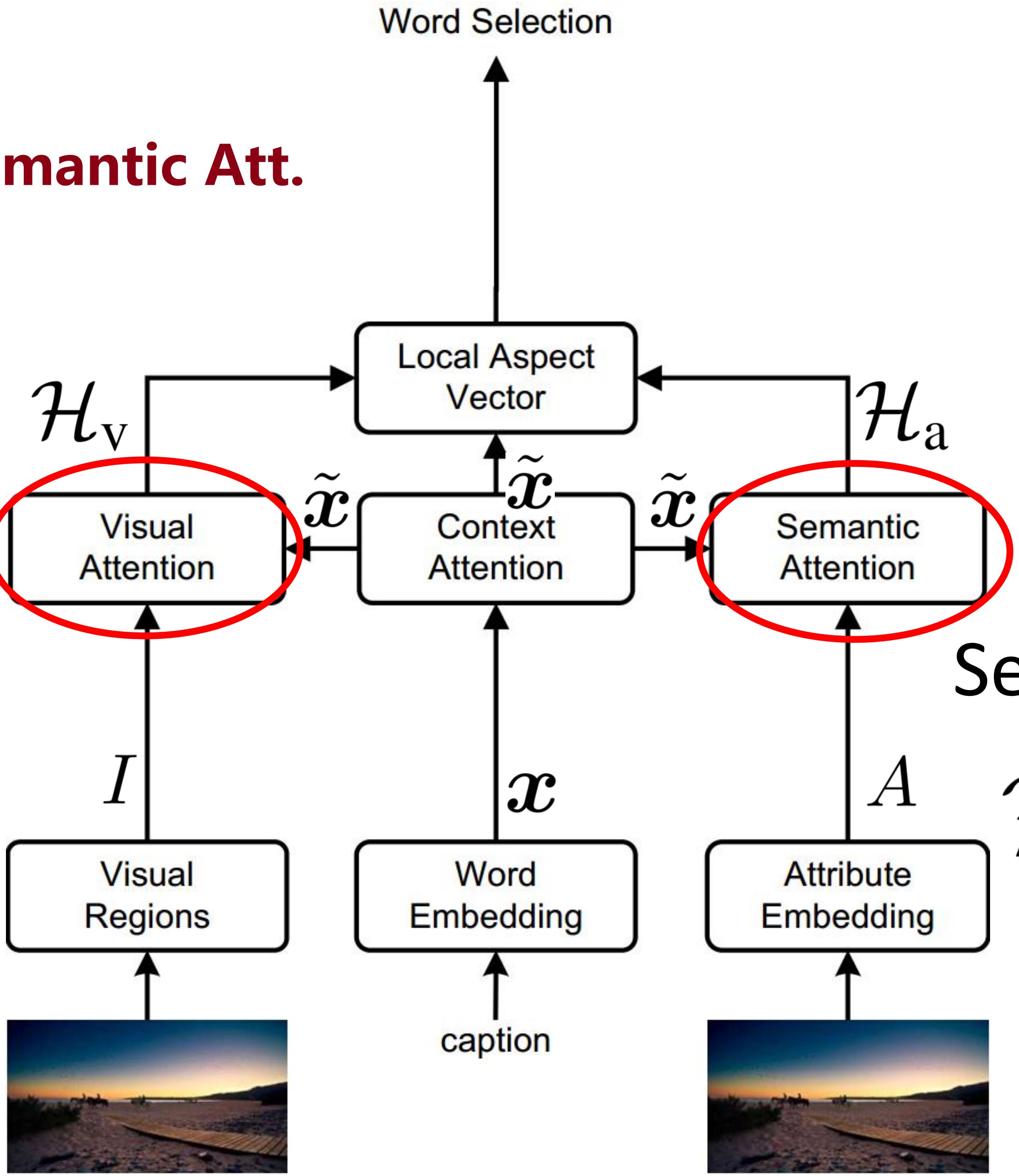
Context Attention:

$$\tilde{x} = \mathcal{G}(\mathcal{H}_{\mathrm{X}}(x, X, X), x)$$

x : the current input caption word.

X : the previously generated words.

Base Model:
Visual Att. and Semantic Att.



Visual Attention:

$$\mathcal{H}_\text{v}(\tilde{x}, \underline{I}, \underline{I})$$

Semantic Attention

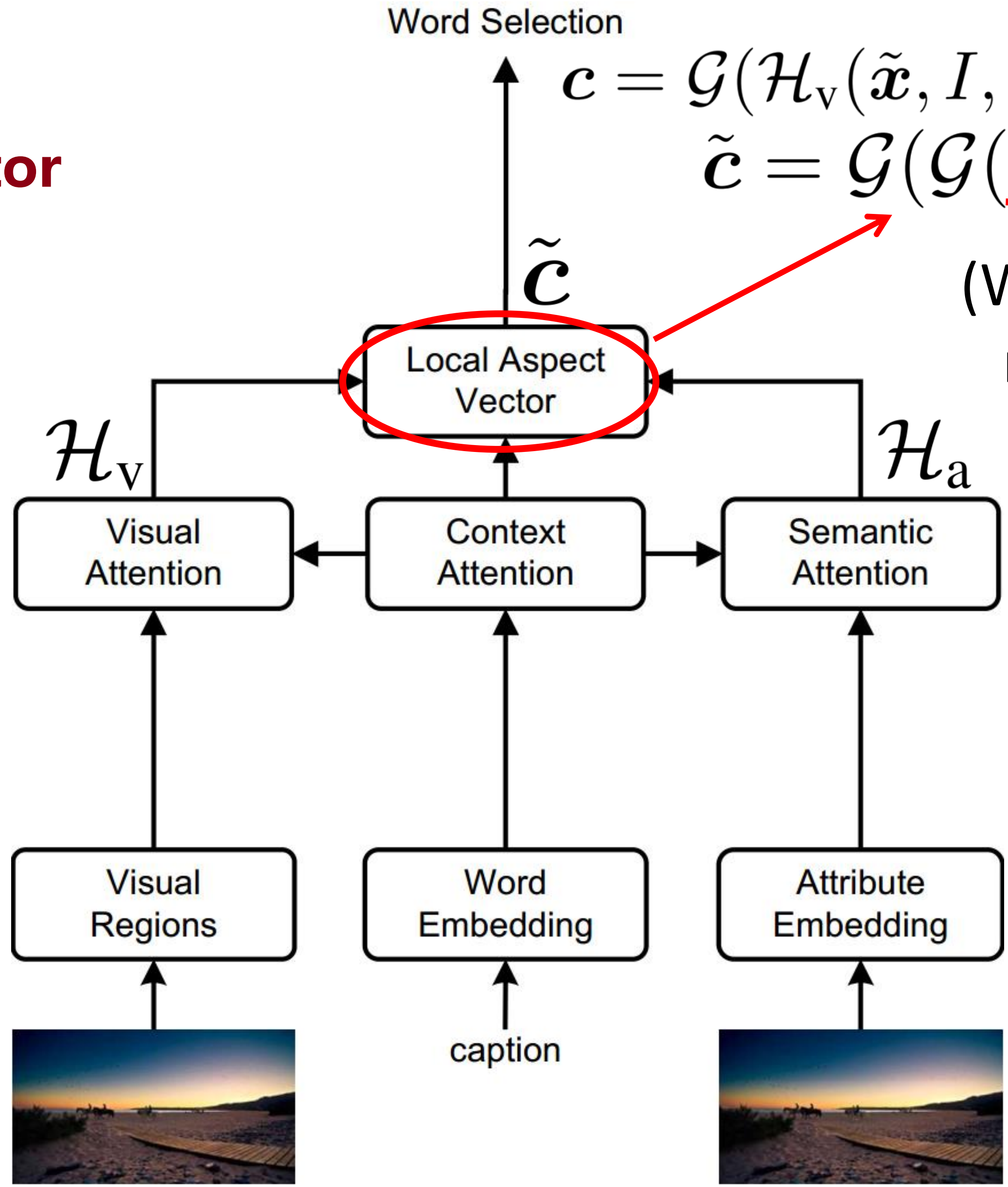
$$\mathcal{H}_\text{a}(\tilde{x}, \underline{A}, \underline{A})$$

Base Model: Local Aspect Vector

$$\boldsymbol{c} = \mathcal{G}(\mathcal{H}_{\mathrm{v}}(\tilde{\boldsymbol{x}}, I, I) + \mathcal{H}_{\mathrm{a}}(\tilde{\boldsymbol{x}}, A, A), \tilde{\boldsymbol{x}})$$

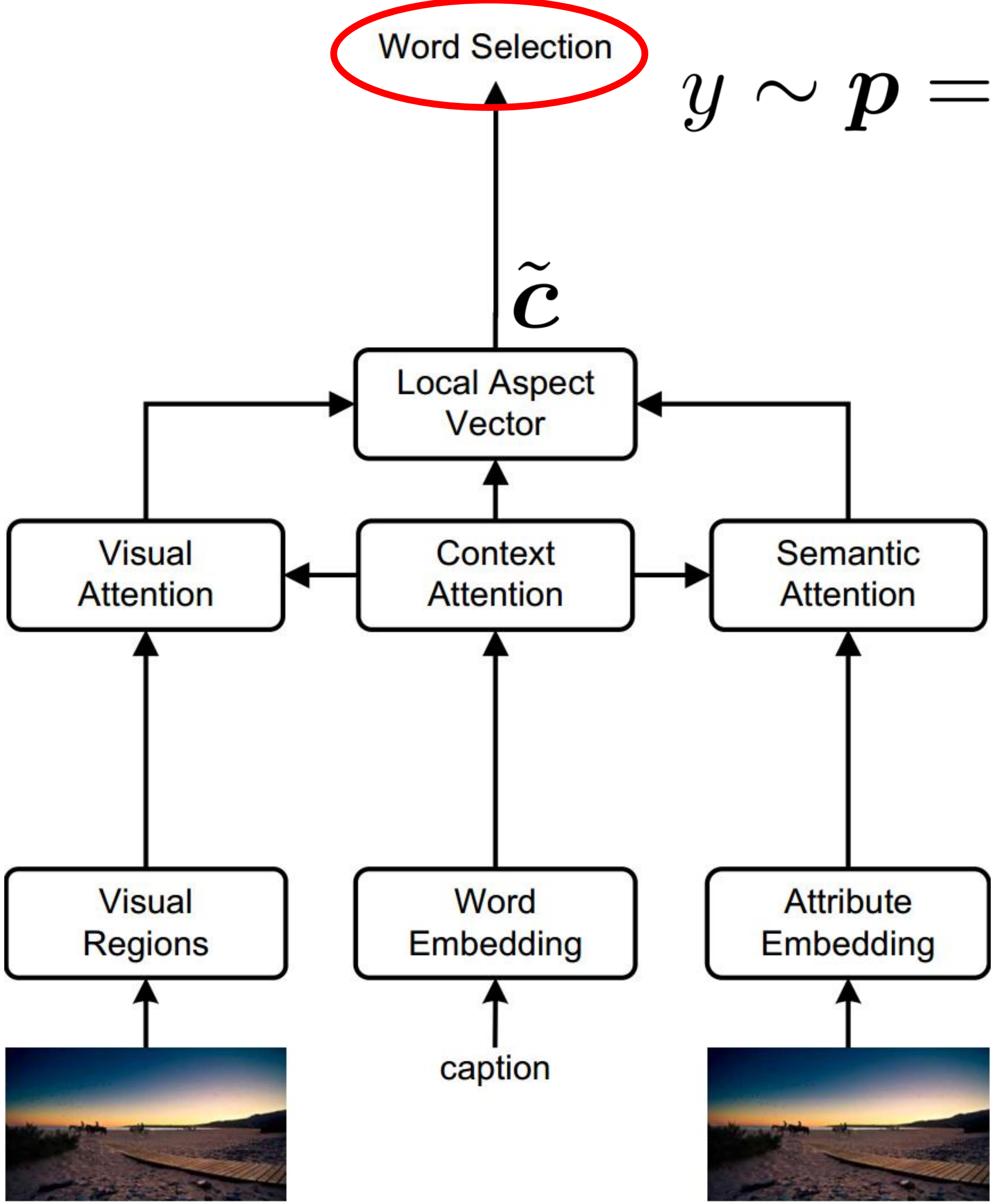
$$\tilde{\boldsymbol{c}} = \mathcal{G}(\mathcal{G}(\mathcal{F}(\boldsymbol{c}), \boldsymbol{c}), \boldsymbol{x})$$

(Where *F* is a two-layer rectified linear unit)

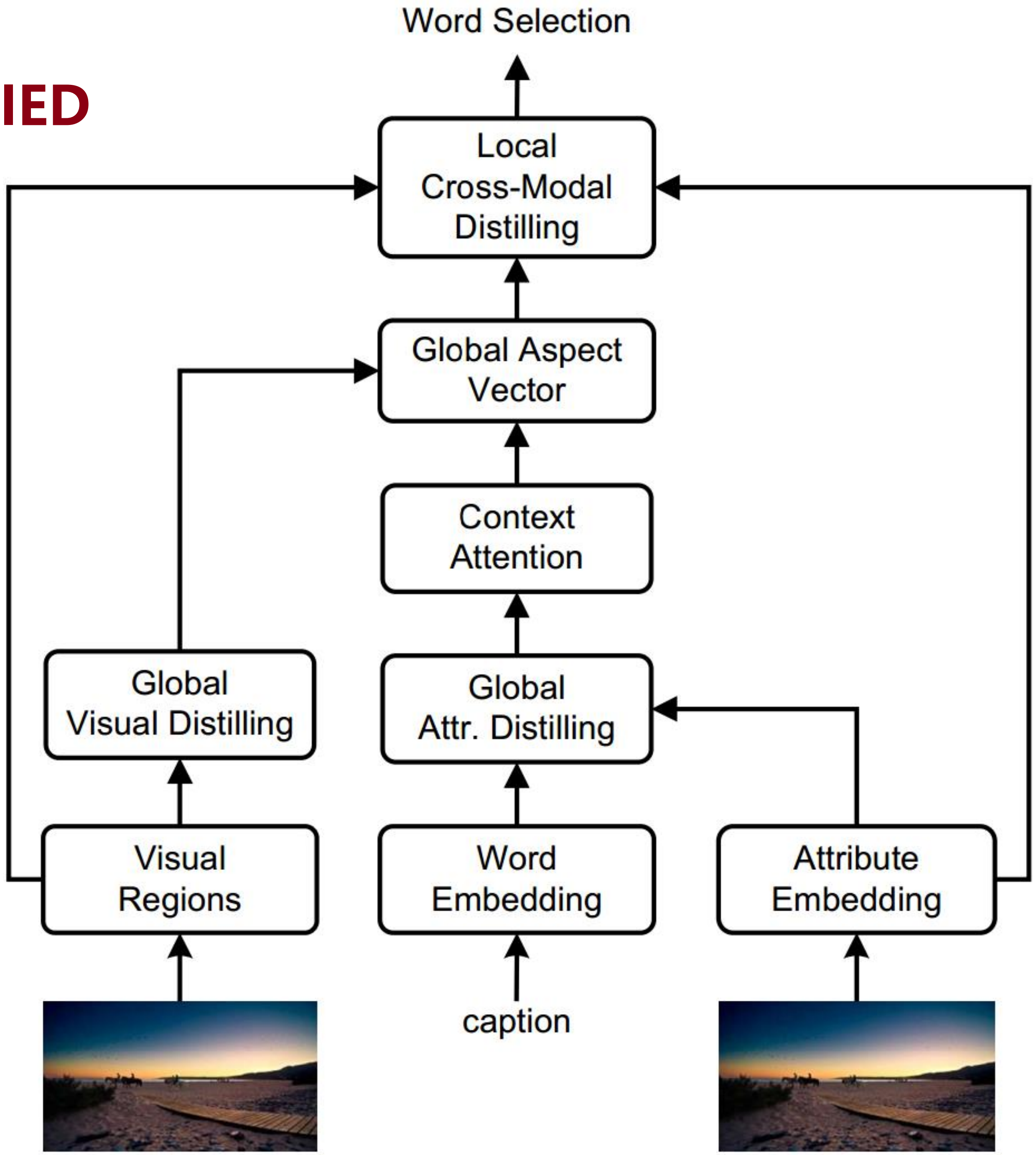


Base Model: Word Selection

$$y \sim \boldsymbol{p} = \mathrm{softmax}(W^{\mathrm{C}}\tilde{\boldsymbol{c}})$$



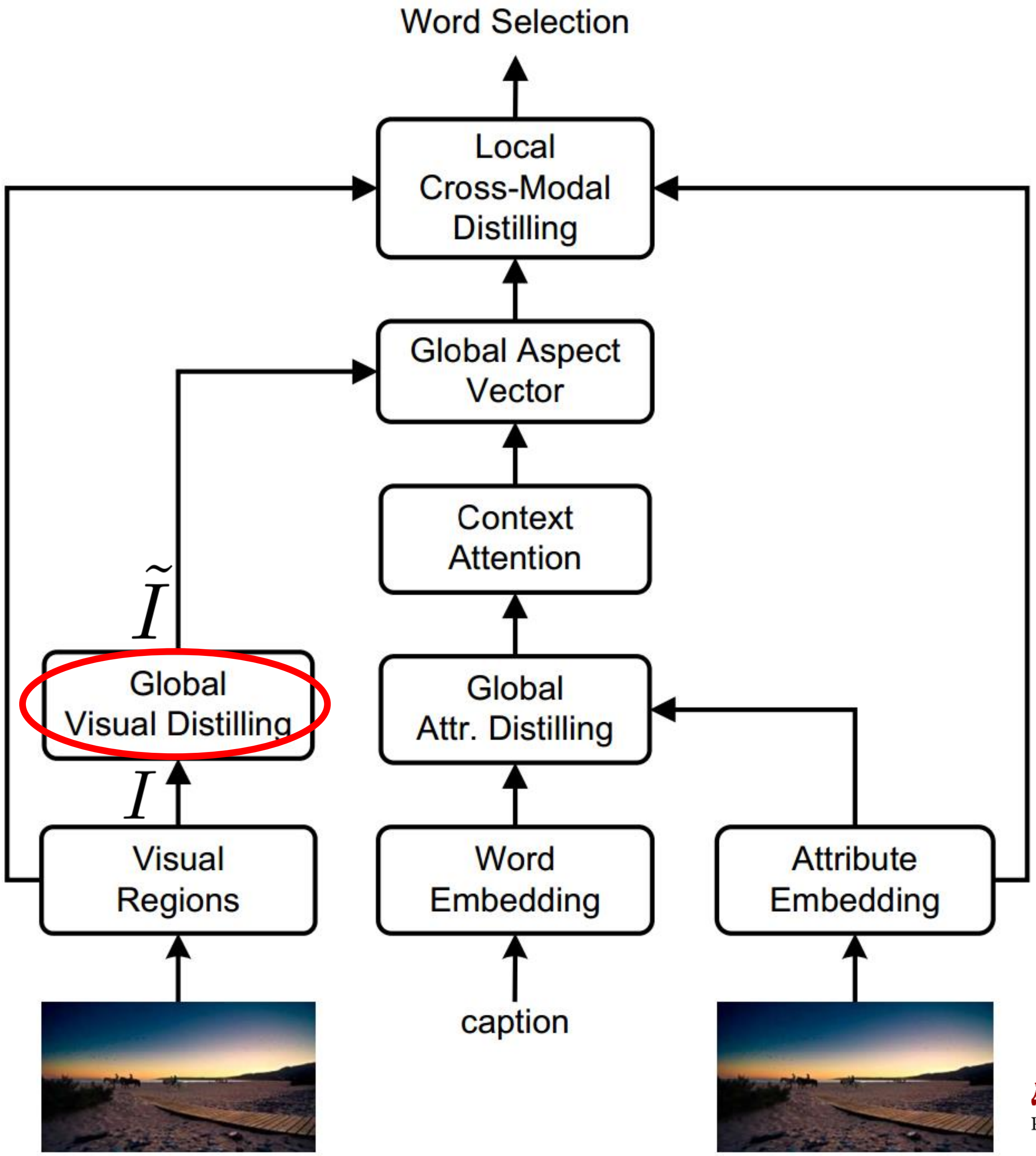
GLIED



- The **Global Visual Distilling** learns salient region groupings and distills naturally related image regions for a higher-level representation of the image in the vision domain.

- The **Global Attribute Distilling** learns attribute collocations and have the ability of thinking in association and using collocations when phrasing sentences.

GLIED:
Global Visual Distilling

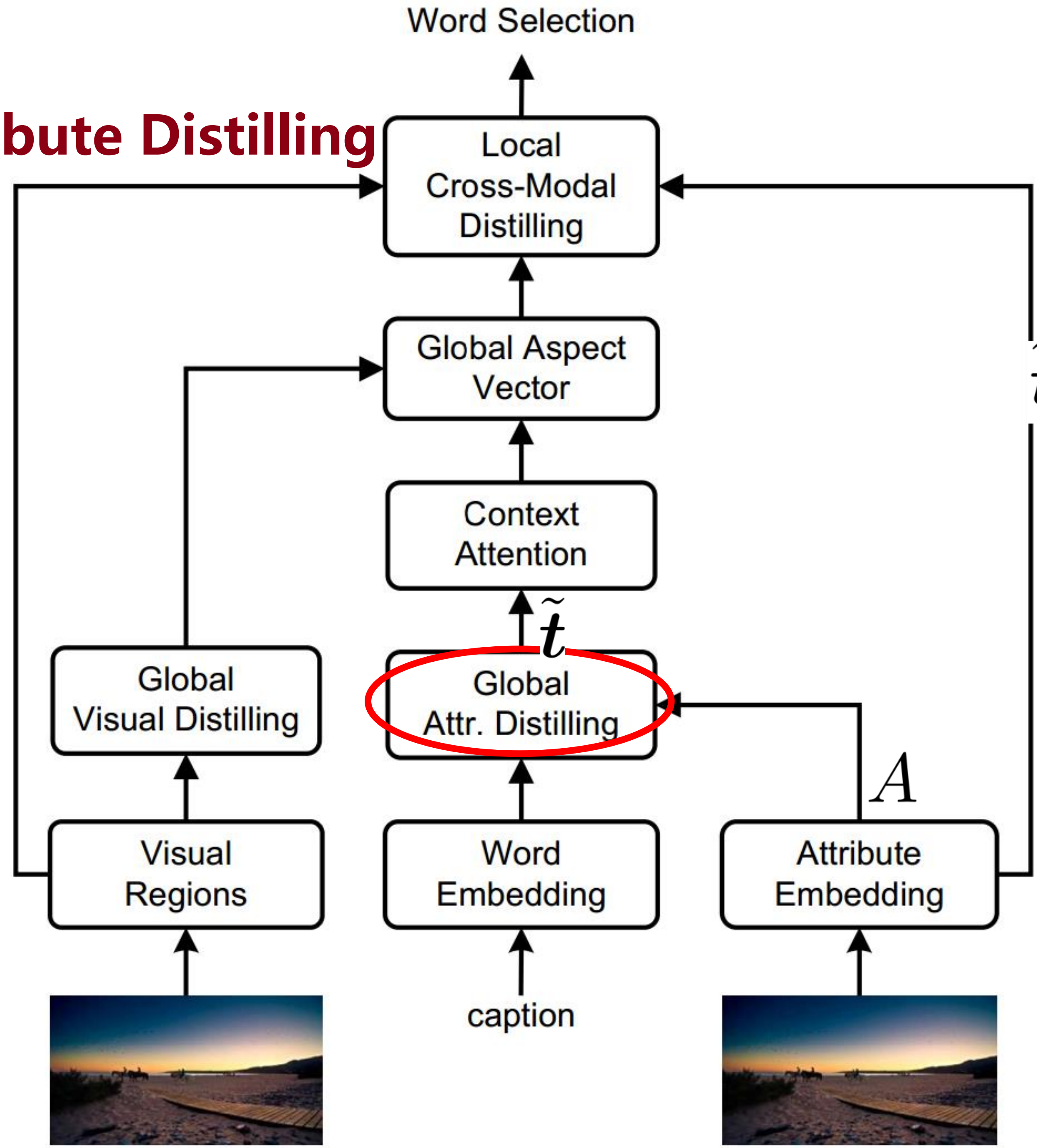


Global Visual Distilling:

$$\tilde{I} = \mathcal{G}(\mathcal{H}_{\mathrm{vd}}(\underline{I}, I, I), \underline{I}))$$

Extending focus on one specific object to its surrounding areas and seek for other objects that often appears together with the object. Those spatially or semantically related objects form an inherent group we attend to.

GLIED:
Global Attribute Distilling

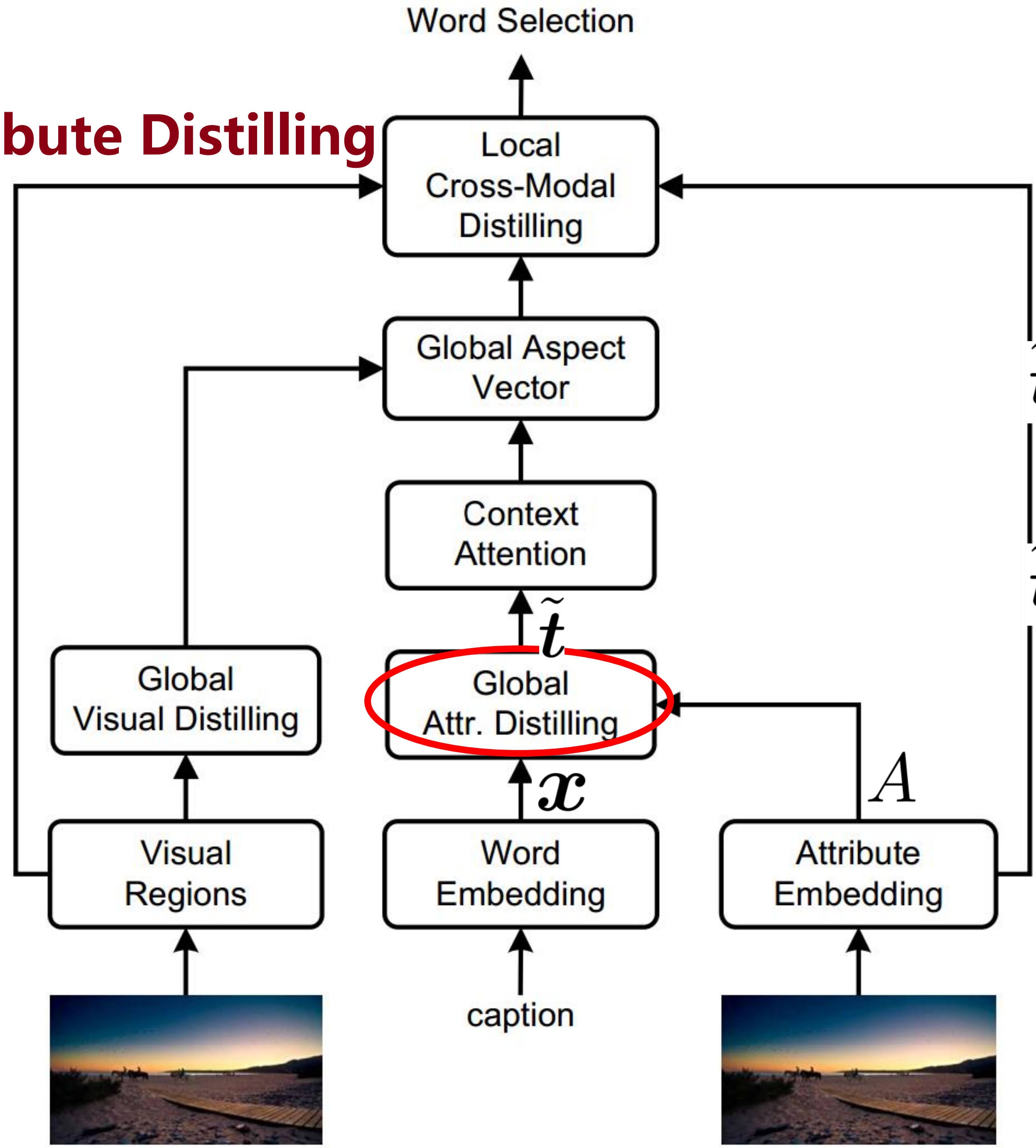


Global Attr. Distilling:

$$\tilde{t} = \mathcal{G}(\mathcal{H}_{\mathrm{ad}}(A, A, A), A)$$

Unlike image regions which are based on shapes or textures, simply combining the attributes may result in common collocations that do not actually appear in the image.

GLIED: Global Attribute Distilling



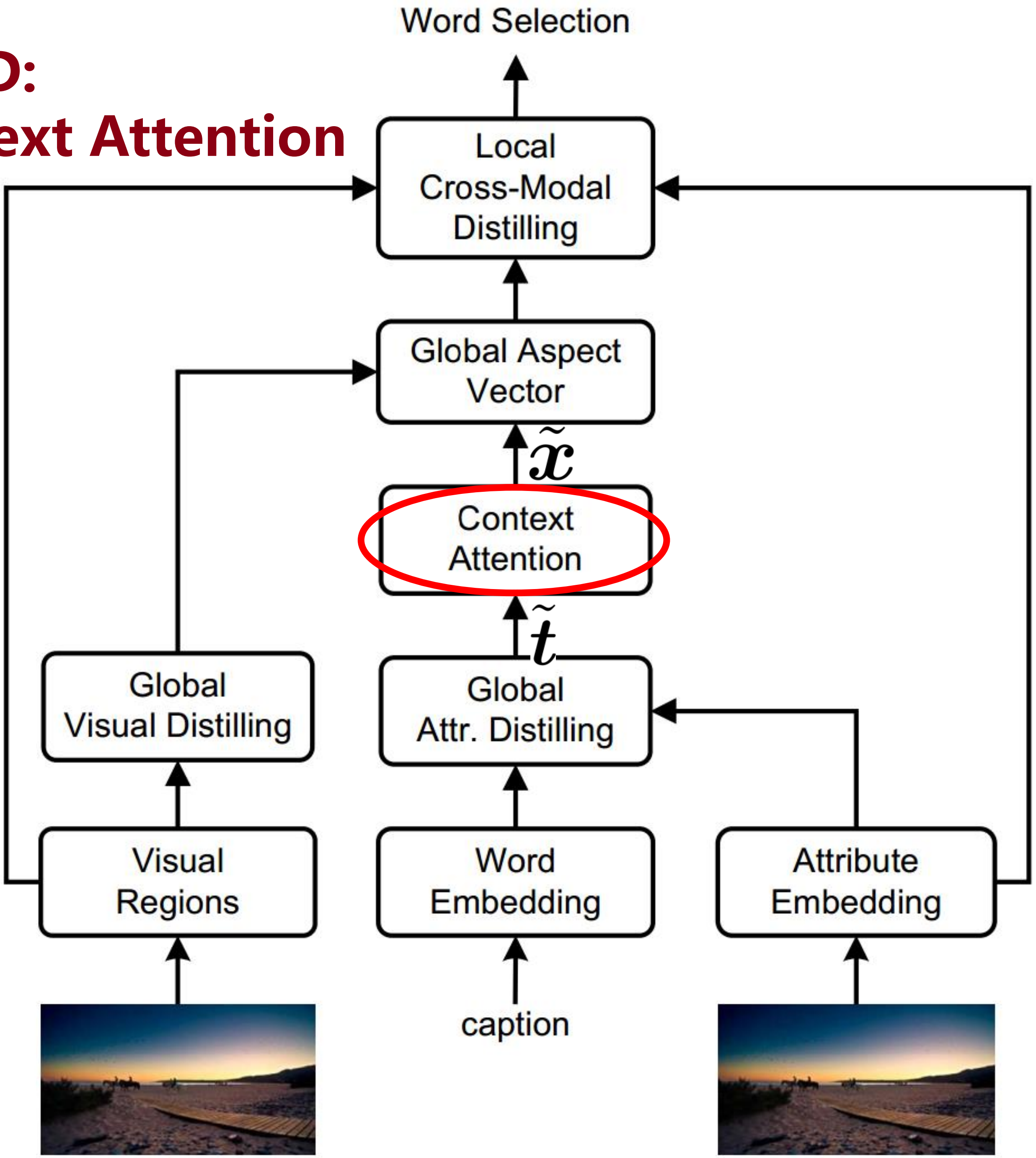
Global Attr. Distilling:

$$\tilde{t} = \mathcal{G}(\mathcal{H}_{\mathrm{ad}}(\cancel{A}, A, A), \cancel{A})$$

$$\tilde{t} = \mathcal{G}(\mathcal{H}_{\mathrm{ad}}(\underline{\boldsymbol{x}}, A, A), \underline{\boldsymbol{x}})$$

Using the input word as pivot to extract constrained collocations.

GLIED: Context Attention



(Same as the base model)

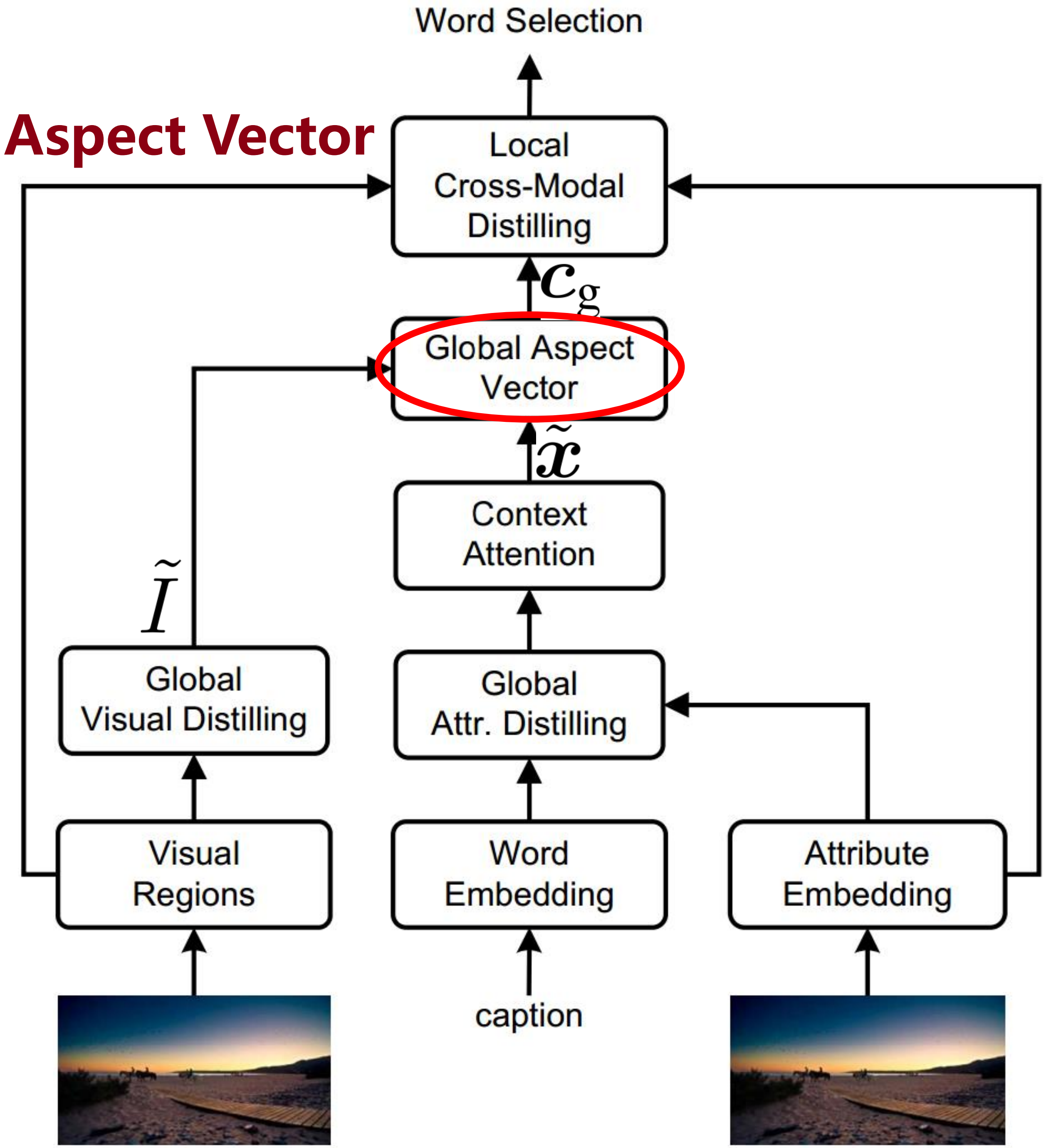
Context Attention:

$$\tilde{x} = \mathcal{G}(\mathcal{H}_{\mathrm{X}}(\tilde{t}, \tilde{T}, \tilde{T}), \tilde{t})$$

$\tilde{t}$ is the current input word enriched by attribute collocations.

$\tilde{T}$ is the pack of $\tilde{t}$.

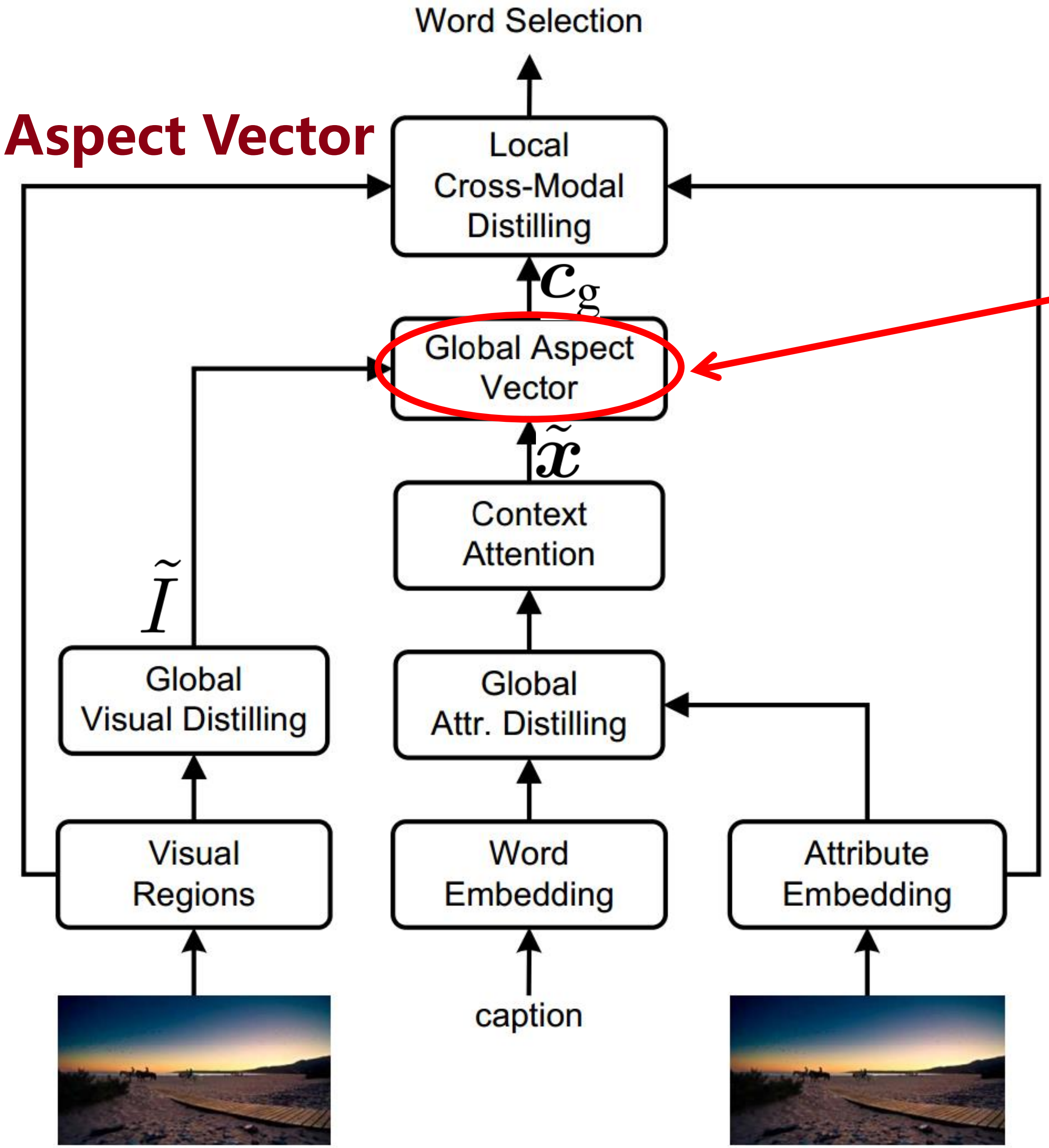
GLIED:
Global Aspect Vector



Further incorporate the visual region groups:

$$c_g = \mathcal{G}(\mathcal{H}_v(\tilde{x}, \tilde{I}, \tilde{I}), \tilde{x})$$

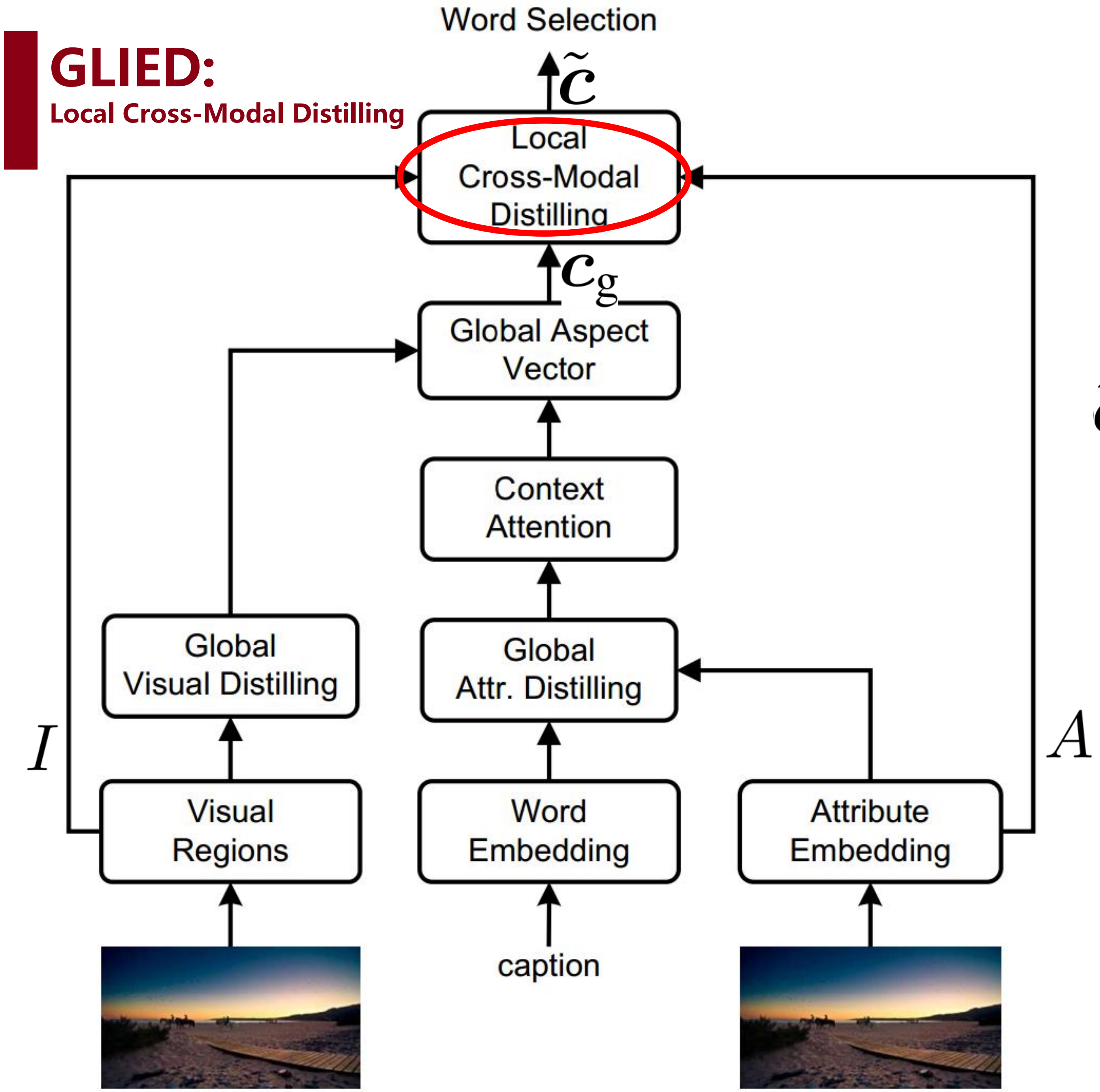
GLIED: Global Aspect Vector



The global aspect vector is a powerful basis for description, but it could be too general for word selection that is precise and detailed, since the basic unit of its sources is the learned groupings of regions and attributes.

北京大學 PEKING UNIVERSITY

GLIED:
Local Cross-Modal Distilling



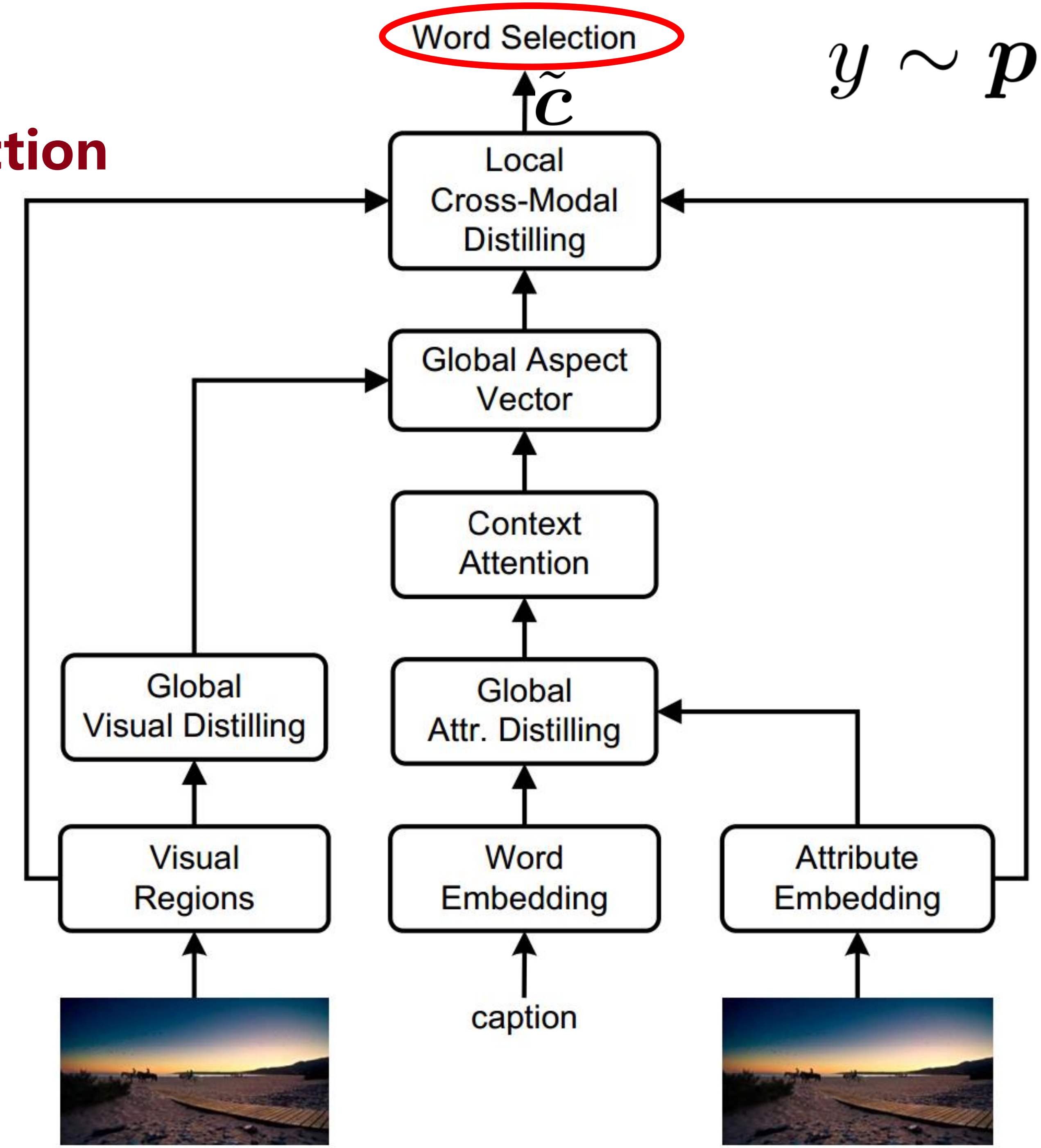
Local Cross-Modal Distilling:

$$\tilde{c} = \mathcal{G}(\mathcal{H}_{\mathrm{vl}}(c_{\mathrm{g}}, I, I) + \mathcal{H}_{\mathrm{al}}(c_{\mathrm{g}}, A, A), c_{\mathrm{g}})$$

The local cross-modal distilling method to make the decoding revisit the fine-grained source information so that the exact aspect could be retrieved.

GLIED:
Word Selection

$$y \sim \boldsymbol{p} = \mathrm{softmax}(W^{\mathrm{C}}\tilde{\boldsymbol{c}})$$



3 Experiments

Experiments

Dataset

Microsoft COCO(MSCOCO)

- ✓ Sparrow bird on branch, with beak inspecting leaves on branch.
- ✓ A bird sitting on the branch of a tree near leaves.
- ✓ A bird that is sitting in a tree.
- ✓ A bird sitting on a branch of a tree.
- ✓ A bird that is on a small branch of a tree.

Evaluation Metrics

- ✓ CIDEr
- ✓ SPICE
- ✓ BLEU
- ✓ METEOR
- ✓ ROUGE

Experiments: Quantitative Comparisons

Strong baseline

Suggesting the cross-modal point of view helps to generate coherent captions.

Cross-Entropy	B-1	B-4	M	R	C	S
SCST$^{\Sigma}$	-	32.8	26.7	55.1	106.5	-
Up-Down	77.2	36.2	27.0	56.4	113.5	20.3
RFNet$^{\Sigma}$	77.4	37.0	27.9	57.3	116.3	20.8
GCN-LSTM	77.4	37.1	28.1	57.2	117.1	21.1
Base	77.0	36.3	27.6	56.6	113.5	20.6
GLIED	**77.8**	**37.9**	**28.3**	**57.6**	**118.2**	**21.2**

Table 1: Comparisons with the the existing models on the COCO Karpathy test split. The symbol Σ denotes model ensemble.

Experiments: Quantitative Comparisons

Showing that the intrinsic associations of source information provides a solid basis for describing images.

Cross-Entropy	B-1	B-4	M	R	C	S
SCST$^{\Sigma}$	-	32.8	26.7	55.1	106.5	-
Up-Down	77.2	36.2	27.0	56.4	113.5	20.3
RFNet$^{\Sigma}$	77.4	37.0	27.9	57.3	116.3	20.8
GCN-LSTM	77.4	37.1	28.1	57.2	117.1	21.1
Base	77.0	36.3	27.6	56.6	113.5	20.6
GLIED	**77.8**	**37.9**	**28.3**	**57.6**	**118.2**	**21.2**

Table 1: Comparisons with the the existing models on the COCO Karpathy test split. The symbol Σ denotes model ensemble.

Experiments: Quantitative Comparisons

Fine tuning with Reinforcement Learning

RL on CIDEr	B-1	B-4	M	R	C	S
SCST$^{\Sigma}$	-	35.4	27.1	56.6	117.5	-
Up-Down	79.8	36.3	27.7	56.9	120.1	21.4
RFNet$^{\Sigma}$	80.4	37.9	28.3	58.3	125.7	21.7
GCN-LSTM	**80.9**	38.3	28.6	58.5	128.7	22.1
GLIED	80.4	**39.6**	**28.9**	**58.8**	**129.3**	**22.6**

Table 1: Comparisons with the the existing models on the COCO Karpathy test split. The symbol Σ denotes model ensemble.

Model Complexity and Computation Speed

Similar number of parameters.

Our base model exhibits strongest performance, and thanks to the help of multi-head attention, Base model is time-efficient in both the training and inference stages.

Methods	#Parameters	Train Time (h)	Inference Speed (ips)	CIDEr
LSTM	**11.5M**	16.8	28.6	105.7
SoftAtt	12.1M	20.4	23.3	111.5
Up-Down	50.1M	24.9	14.8	113.2
CT†	27.5M	22.7	12.9	115.1
Base	12.3M	13.2	**37.9**	113.5
Ours	18.3M	**11.9**	34.5	**118.2**

Table 2: Comparisons of model complexity and speed. #Parameters are estimated. Time and Speed is measured on a single NVIDIA GeForce GTX 1080 Ti. ips stands for images per second. The symbol † denotes the result reported from original papers.

Model Complexity and Computation Speed

Our cross-modal base model is comparable with Up-Down in accuracy, yet 4x smaller and 2x faster.

Methods	#Parameters	Train Time (h)	Inference Speed (ips)	CIDEr
LSTM	**11.5M**	16.8	28.6	105.7
SoftAtt	12.1M	20.4	23.3	111.5
Up-Down	50.1M	24.9	14.8	113.2
CT†	27.5M	22.7	12.9	115.1
Base	12.3M	13.2	**37.9**	113.5
Ours	18.3M	**11.9**	34.5	**118.2**

Table 2: Comparisons of model complexity and speed. #Parameters are estimated. Time and Speed is measured on a single NVIDIA GeForce GTX 1080 Ti. ips stands for images per second. The symbol † denotes the result reported from original papers.

Model Complexity and Computation Speed

GLIDE achieves faster training speed (faster convergence) compared to the Base model, at the cost of only moderate increase in parameters and slight inference speed regression.

It attests to the effectiveness of the cross-modal point of view.

Methods	#Parameters	Train Time (h)	Inference Speed (ips)	CIDEr
LSTM	**11.5M**	16.8	28.6	105.7
SoftAtt	12.1M	20.4	23.3	111.5
Up-Down	50.1M	24.9	14.8	113.2
CT†	27.5M	22.7	12.9	115.1
Base	12.3M	13.2	**37.9**	113.5
Ours	18.3M	**11.9**	34.5	**118.2**

Table 2: Comparisons of model complexity and speed. #Parameters are estimated. Time and Speed is measured on a single NVIDIA GeForce GTX 1080 Ti. ips stands for images per second. The symbol † denotes the result reported from original papers.

Model Complexity and Computation Speed

The basic Transformer model CT and our model based on the same multi-head attention structure.

Ours > CT

Methods	#Parameters	Train Time (h)	Inference Speed (ips)	CIDEr
LSTM	**11.5M**	16.8	28.6	105.7
SoftAtt	12.1M	20.4	23.3	111.5
Up-Down	50.1M	24.9	14.8	113.2
CT†	27.5M	22.7	12.9	115.1
Base	12.3M	13.2	**37.9**	113.5
Ours	18.3M	**11.9**	34.5	**118.2**

Table 2: Comparisons of model complexity and speed. #Parameters are estimated. Time and Speed is measured on a single NVIDIA GeForce GTX 1080 Ti. ips stands for images per second. The symbol † denotes the result reported from original papers.

4 Conclusion

Conclusion

- We present a simple yet effective approach exploring and distilling the cross-modal source information.
- The global distilling methods learn to capture salient region groupings and attribute collocations and explore a spatial and relational coarse-grained representation of the image.
- The local distilling method in contrast makes the decoder revisit the fine-grained source representation so that related and specific details can be retrieved.
- Our approach outperforms previous works with fewer parameters and faster computation.

Thank you!

If you have any questions about our paper, you can send an email to fenglinliu98@pku.edu.cn