中国科学院信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS
中国科学院大学 University of Chinese Academy of Sciences
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY
THE UNIVERSITY OF ADELAIDE AUSTRALIA

WeChat Paper GitHub

Visual-Semantic Decomposition and Partial Alignment for Document-based Zero-Shot Learning

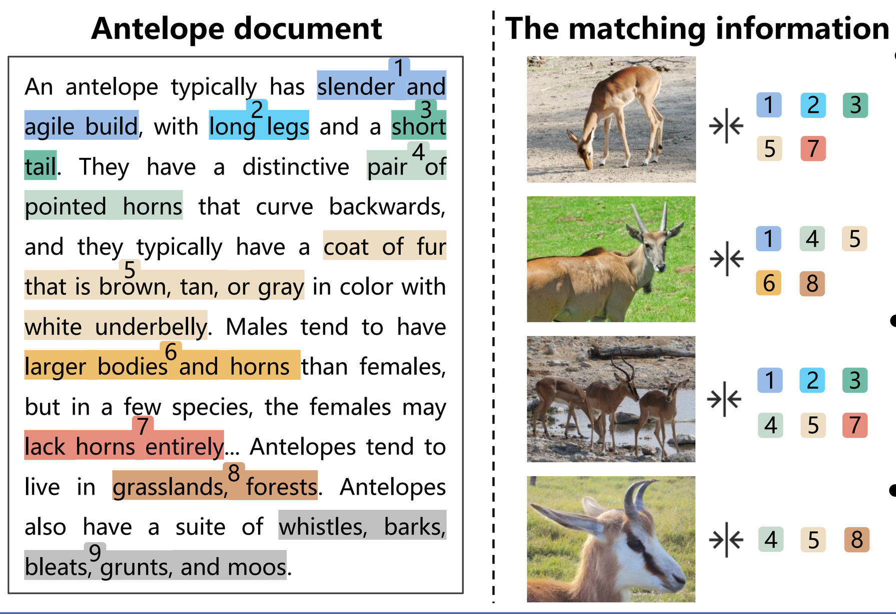
Xiangyan Qu, Jing Yu†, Keke Gai, Jiamin Zhuang, Yuanmin Tang, Gang Xiong, Gaopeng Gou, Qi Wu

Motivation

Document-based Zero-Shot Learning (ZSL)

- ZSL aims to identify unseen classes by training a set of seen classes.
- Document-based ZSL uses category-level text corpora from Wiki as auxiliary information, transferring knowledge by shared descriptions.

Partial Association between Images and Documents

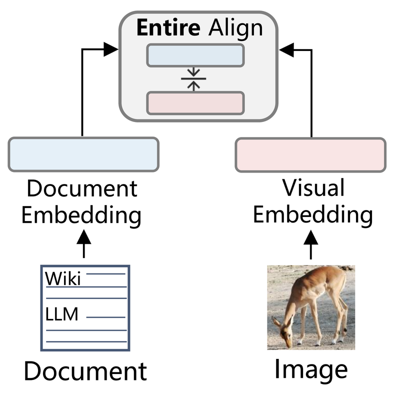


- diverse image content, noisy document, exhaustive description

result in

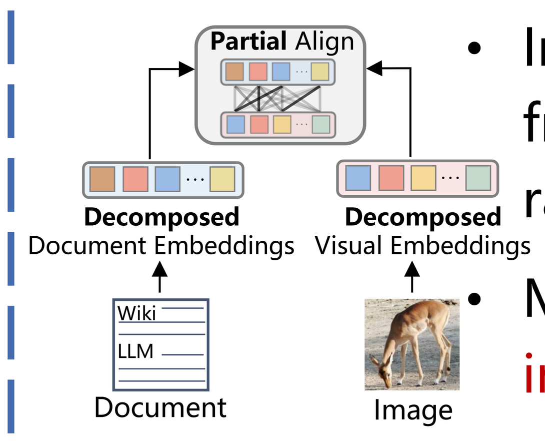
- Semantics in the document may partially be reflected in the image.
- Distinct images capture varying semantics within the document.

Previous Work: Entire Align



- Existing methods align the entire semantics of a document with images to transfer knowledge.
- They disregard that semantics is not equivalent between them, resulting in a suboptimal alignment.

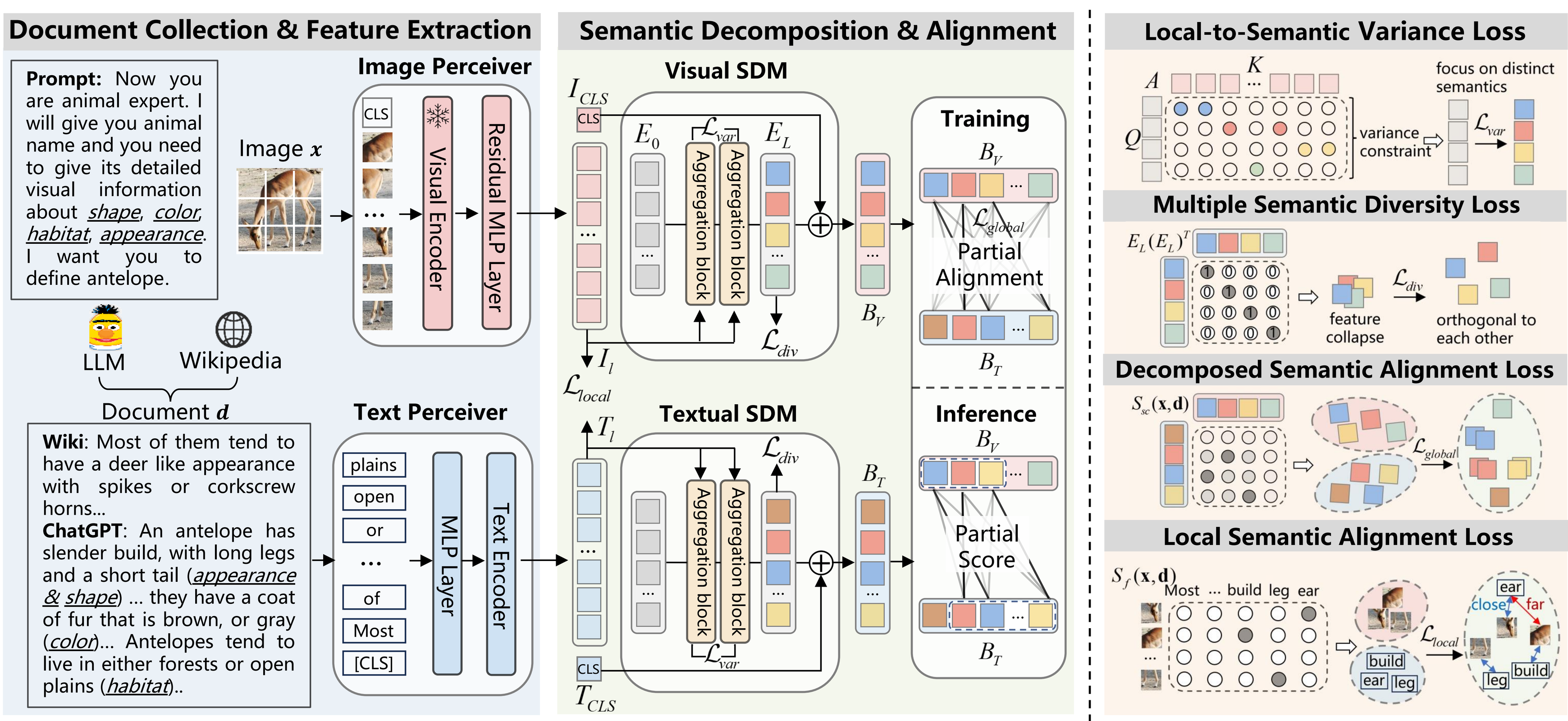
Our EmDepart: Partial Align



- In contrast, we extract multi-view semantic concepts from documents and images and align the matching rather than entire concepts.
- Moreover, two losses are proposed to solve issues of information redundancy caused by feature collapse.

Our Solutions

An overview of our EmDepart, which contains an image perceiver, a text perceiver, and visual and textual semantic decomposition modules.



(a) Overall architecture (b) Loss functions

Document Collection: Wiki + LLMs

- Collecting category documents from encyclopedia (*e.g.,* Wikipedia)
- Enriching less-described document by Large Language Models (LLMs)

Visual-Semantic Decomposition

Visual and textual SDM aggregate information and decompose them to generate multi-view semantic embeddings:

- This process introduces a set of learnable tokens and cross-attention mechanisms.

$$\mathbf{E}_t = \mathrm{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{r_h}}\right)\mathbf{VW}_o + \mathbf{E}^{t-1},$$

$$\mathbf{E}_t = \mathrm{MLP}(\mathbf{E}_t) + \mathbf{E}_t.$$

- In the last, we concatenate the output with global feature to maintain small set variance.

$$\mathbf{B}_V = \mathrm{LayerNorm}(\mathbf{E}_L + [\mathbf{I}_{CLS}]^{\times k}).$$

Distinct Semantic Information Learning

To solve the information redundancy caused by feature collapse (multiple embeddings with a slight variance), we introduce two losses:

- $\mathcal{L}_{var}$: encourage each view embedding to focus on unique local information.
- $\mathcal{L}_{div}$: penalize each view embedding orthogonal to others.

$$C(\mathbf{A}_V) = \sum_{t=1}^{l}\sum_{j=1}^{n}\max(0, \gamma - \sqrt{Var(\mathbf{a}_{tj}) + \epsilon}),$$

$$\mathcal{L}_{var} = \frac{1}{2}\left(C(\mathbf{A}_T) + C(\mathbf{A}_V)\right).$$

$$\mathcal{L}_{div} = \frac{1}{2k^2}(\| \mathbf{M}_T - \mathbb{I}\|_2 + \| \mathbf{M}_V - \mathbb{I}\|_2).$$

Partial Semantic Alignment

We assigns distinct weights to every document-image embedding pair based on similarity to model the partial association.

$$\mathrm{LSE}(\mathbf{b}_T, \mathbf{B}_V) = \log\Big(\sum_{\mathbf{b}_V \in \mathbf{B}_V} e^{\cos(\mathbf{b}_T, \mathbf{b}_V)}\Big), \qquad S_{sc}(\mathbf{x}, \mathbf{d}) = \frac{1}{2k}\Big(\sum_{\mathbf{b}_T \in \mathbf{B}_T}\mathrm{LSE}(\mathbf{b}_T, \mathbf{B}_V) + \sum_{\mathbf{b}_V \in \mathbf{B}_V}\mathrm{LSE}(\mathbf{b}_V, \mathbf{B}_T)\Big),$$

$$\mathcal{L}_{global} = -\log\frac{\exp(S_{sc}(\mathbf{x}, \mathbf{d})/\tau)}{\sum_{\mathbf{d}' \in \mathcal{D}^s}\exp(S_{sc}(\mathbf{x}, \mathbf{d}')/\tau)}, \qquad \mathcal{L}_{local} = -\log\frac{\exp(S_f(\mathbf{x}, \mathbf{d}))}{\sum_{\mathbf{d}' \in \mathcal{D}^s}\exp(S_f(\mathbf{x}, \mathbf{d}'))}.$$

- Final Loss: $\mathcal{L} = \mathcal{L}_{global} + \lambda_{local}\mathcal{L}_{local} + \lambda_{var}\mathcal{L}_{var} + \lambda_{div}\mathcal{L}_{div}$.

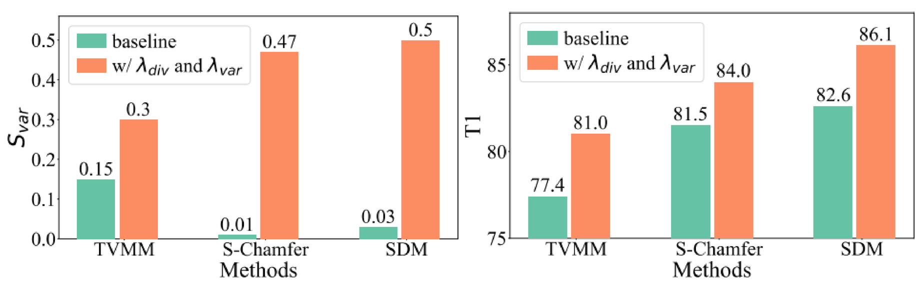
Experiments

Achieved SOTA in document-based ZSL

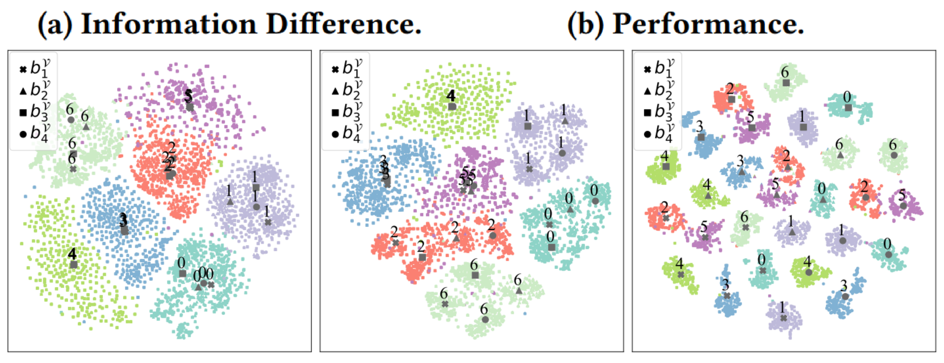
Our EmDepart improves performance by 6.0% and 5.8% on average across all metrics under Wiki and Wiki+LLM documents.

Model	Auxiliary Information	Zero-Shot Learning			Generalized Zero-Shot Learning								
		AWA2	CUB	FLO	AWA2			CUB			FLO		
		T1	T1	T1	U	S	H	U	S	H	U	S	H
GloVe [42]	CLSN	52.1	20.4	21.6	42.1	75.3	54.0	16.2	43.6	23.6	14.4	88.3	24.8
GloVe [42]	Wiki	61.6	29.0	25.8	49.5	78.1	60.6	23.8	62.6	34.5	14.7	91.0	25.3
LongFormer [6]	Wiki	44.2	22.6	8.8	41.6	81.8	55.2	19.9	41.0	26.8	8.8	89.8	16.0
MPNet [48]	Wiki	61.8	25.8	26.3	58.0	76.4	66.0	20.6	44.3	28.2	22.2	96.7	36.1
TF-IDF [45]	Wiki	46.4	39.9	34.0	29.6	87.6	44.2	29.0	52.1	37.3	28.9	94.8	44.3
VGSE [63]	CLSN+IMG	69.6	37.1	-	56.9	82.8	67.4	27.6	**70.6**	39.7	-	-	-
I2DFormer [38]	Wiki	76.4	45.4	40.0	66.8	76.8	71.5	35.3	57.6	43.8	35.8	91.9	51.5
I2MVFormer [37]	Wiki	73.6	42.1	41.3	66.6	82.9	73.8	32.4	63.1	42.8	34.9	96.1	51.2
EmDepart (Ours)	Wiki	81.4 $^{+5.0}$	50.2 $^{+4.8}$	47.2 $^{+5.9}$	76.0	87.8	81.5 $^{+7.7}$	42.6	56.3	48.5 $^{+4.7}$	42.7	**97.6**	59.5 $^{+8.0}$
I2DFormer [38]	Wiki+LLM	77.3	47.0	43.0	68.6	77.4	72.7	38.5	59.3	46.7	40.4	80.1	53.8
I2MVFormer [37]	Wiki+LLM	79.6	51.1	46.2	75.7	79.6	77.6	42.5	59.9	49.7	41.6	91.0	57.1
EmDepart (Ours)	Wiki+LLM	**86.1** $^{+6.5}$	**52.8** $^{+1.7}$	**53.3** $^{+7.1}$	**81.4**	**88.5**	**84.8** $^{+7.2}$	**45.0**	61.4	**51.9** $^{+2.2}$	**52.3**	94.4	**67.3** $^{+10.2}$

Analysis of Feature Collapse



(a) Information Difference. (b) Performance.



(a) w/o $\mathcal{L}_{div}$ and $\mathcal{L}_{var}$. (b) w/o $\mathcal{L}_{div}$. (c) EmDepart.

- We improve previous methods performance and increase the information difference among view embeddings.

Model	AWA2		CUB		FLO	
	T1	H	T1	H	T1	H
TVMM [33]	77.4	74.4	41.6	43.1	42.3	54.2
$+\mathcal{L}_{var} + \mathcal{L}_{div}$	81.0	77.5	45.6	47.4	46.8	59.5
Gain	+3.6	+3.1	+4.0	+4.3	+4.5	+5.3
S-Chamfer [29]	81.5	80.6	45.6	45.2	43.5	57.3
$+\mathcal{L}_{var} + \mathcal{L}_{div}$	84.0	82.9	49.1	49.9	48.9	63.6
Gain	+2.5	+2.3	+3.5	+4.7	+5.4	+6.3
EmDepart(Ours)	**86.1**	**84.8**	**52.8**	**51.9**	**53.3**	**67.3**

Ablation Studies: Model and Document

Ablation on Modules

Model	AWA2	CUB	FLO
	T1	T1	T1
a) full model	**86.1**	**52.8**	**53.3**
Ablation on Loss Function			
b) w/o $\mathcal{L}_{local}$	85.8	45.9	41.7
c) w/o $\mathcal{L}_{div}$	83.5	47.7	41.5
d) w/o $\mathcal{L}_{var}$	85.5	50.1	49.9
e) w/o $\mathcal{L}_{div} + \mathcal{L}_{var}$	82.6	47.5	39.3
f) w/o $\mathcal{L}_{local} + \mathcal{L}_{div} + \mathcal{L}_{var}$	80.1	45.4	37.2
Ablation on Score Function			
g) w/o Partial Score in Eq.12	85.7	52.6	53.0
h) w/ average distance in Eq.7	80.0	39.4	45.7
i) w/ maximum distance in Eq.7	82.2	45.4	44.8
Ablation on Module			
j) w/o global feature in Eq.3	71.6	37.7	39.6
k) w/o SDM	79.7	46.0	45.1
l) w/o residual connection	81.4	49.7	48.3

Ablation on Different Documents

Auxiliary Information	AWA2		FLO	
	T1	H	T1	H
Wiki	81.4	81.5	47.2	59.5
Wiki+GPT3 [7]	$82.3_{\pm 0.45}$	$82.2_{\pm 0.61}$	$53.2_{\pm 0.78}$	$65.5_{\pm 0.87}$
Wiki+LLaMa2 [51]	$82.1_{\pm 0.37}$	$82.8_{\pm 0.27}$	$49.5_{\pm 0.65}$	$62.8_{\pm 0.61}$
Wiki+ChatGPT [20]	$\mathbf{86.1}_{\pm 0.16}$	$\mathbf{84.8}_{\pm 0.29}$	$\mathbf{53.3}_{\pm 0.41}$	$\mathbf{67.3}_{\pm 0.82}$

Computation Cost Analysis

Model	Params ($\times 10^6$)	Train (min)	Inference (ms)	FLO (H)
I2DFormer [38]	2.18	0.72	4.7	53.8
I2MVFormer [37]	**3.86**	0.80	**5.3**	57.1
EmDepart w/o SDM	1.52	0.67	4.6	57.9
EmDepart	3.10	**0.98**	5.2	**67.3**

Partial Association Visualization

It contains the visual-semantic decomposition to offer basic concepts and partial semantic alignment according to the matching information.