

# Learning a Deep Embedding Model for Zero-Shot Learning

CVPR, 2017

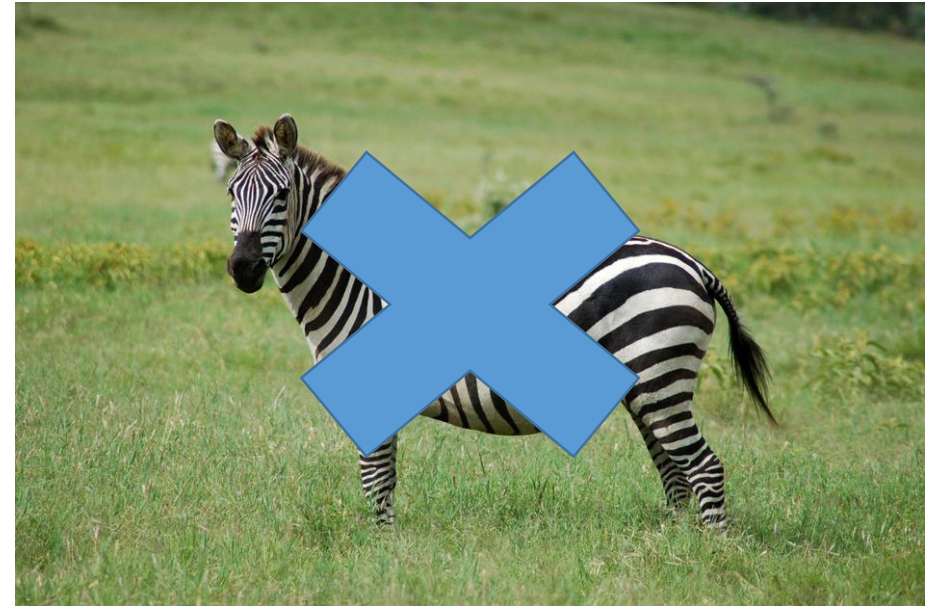
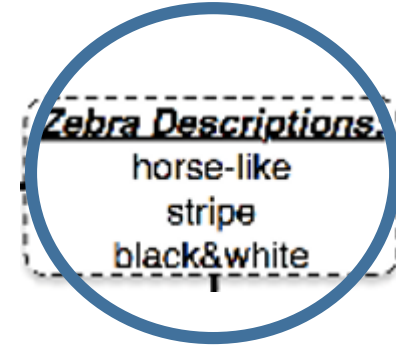
장두수

# Zero-Shot Learning?

Image는 본적 없음, 근데 뭘지는 들어봄

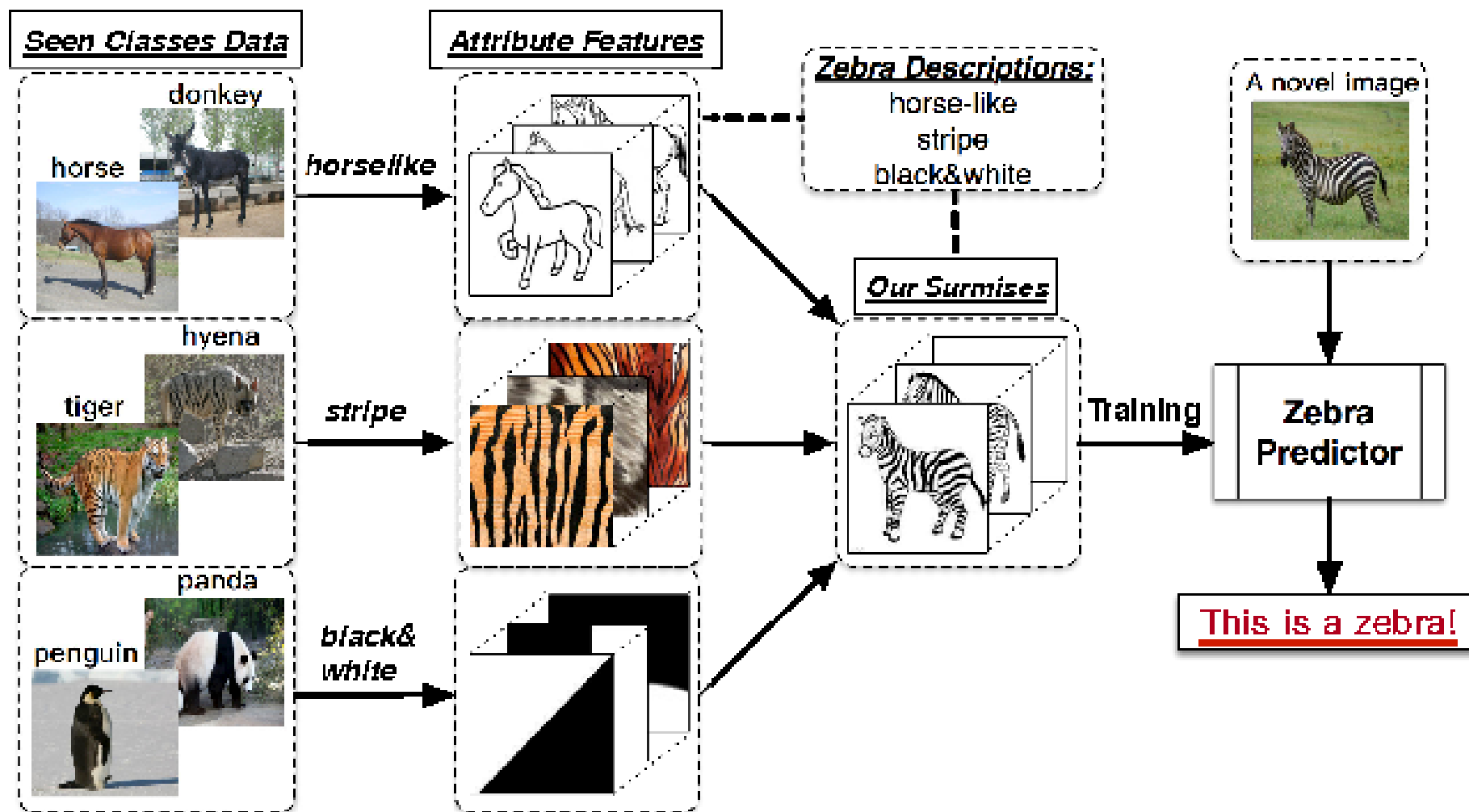


말(seen class)

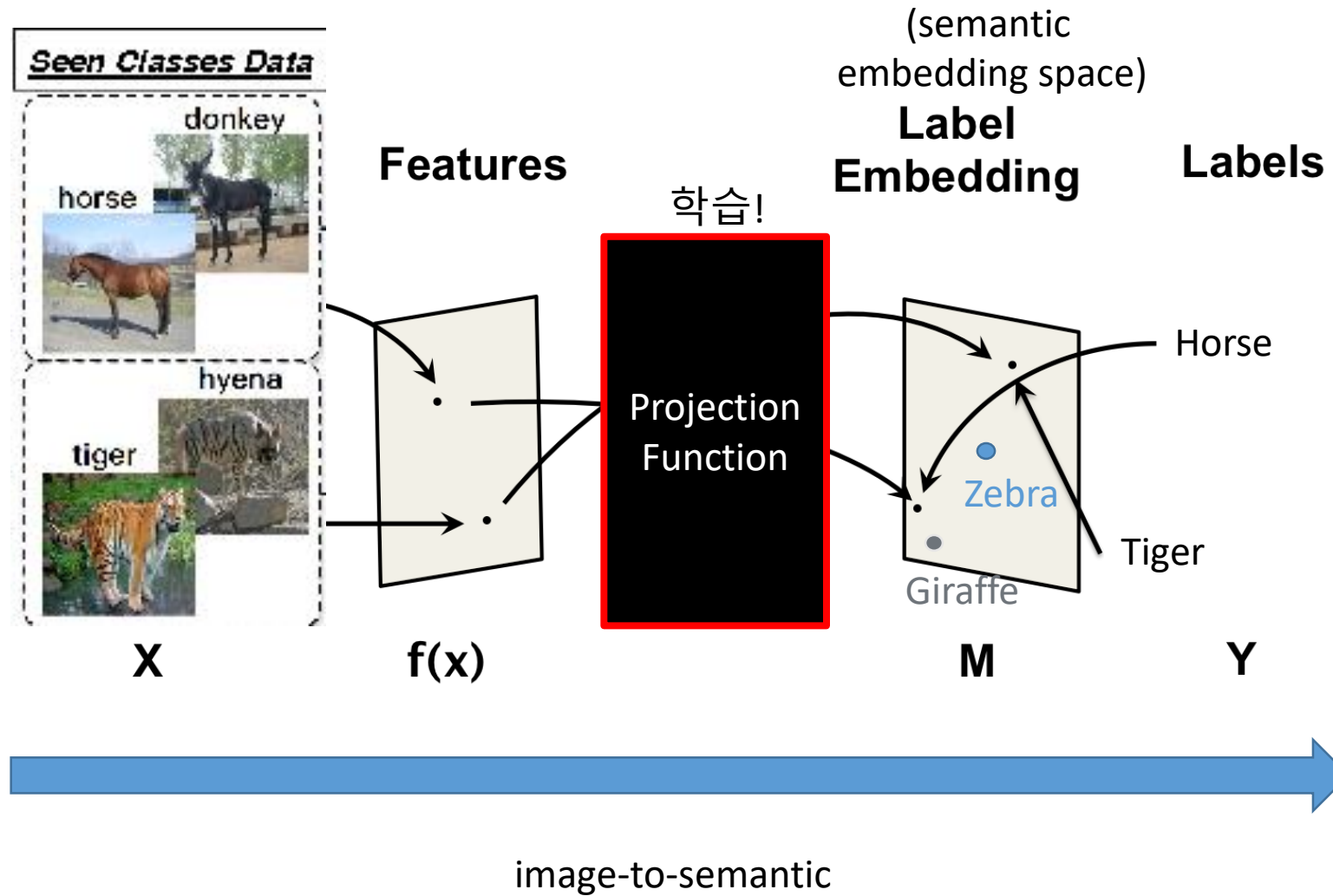


얼룩말(unseen class)

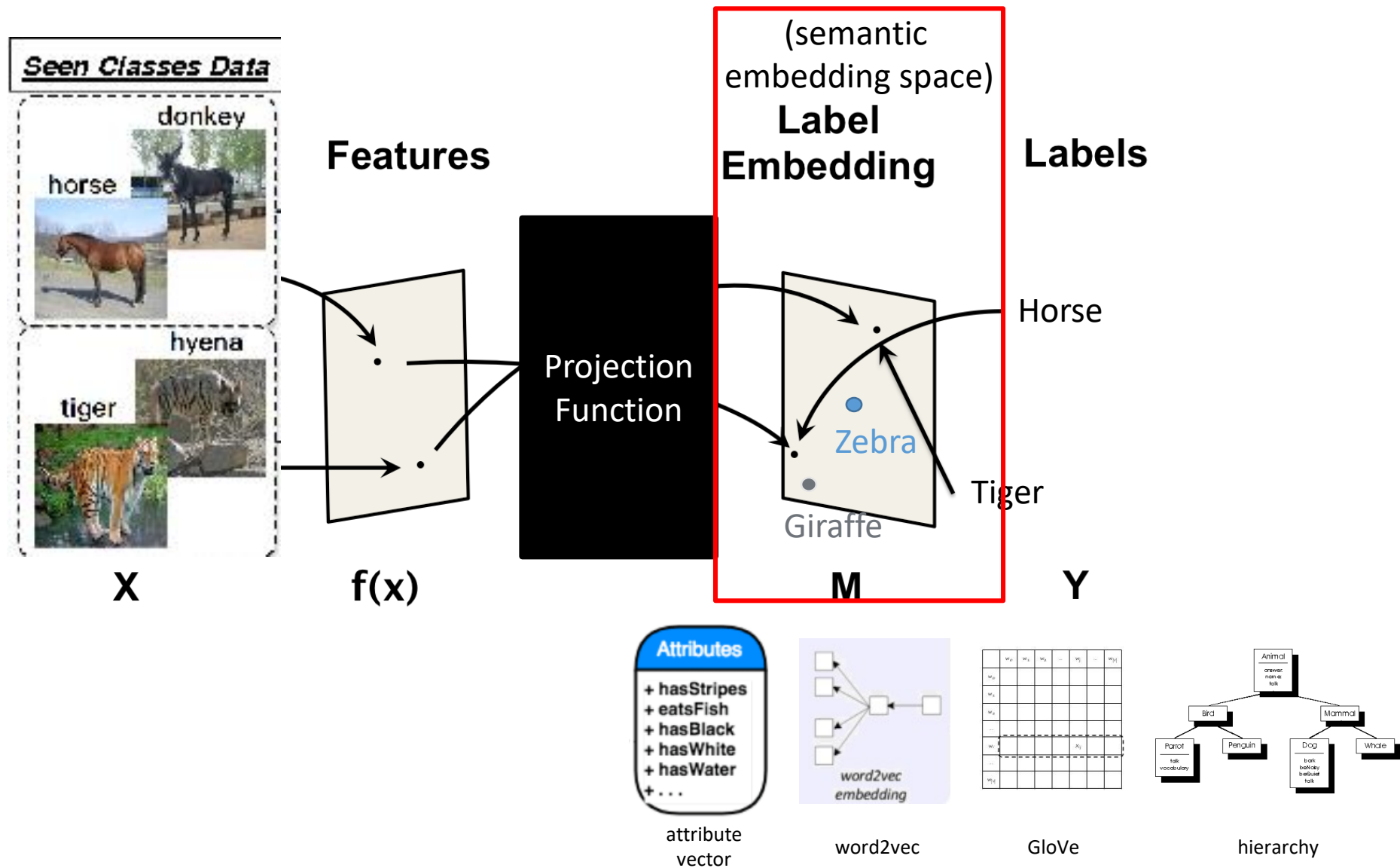
# Zero-Shot Learning?



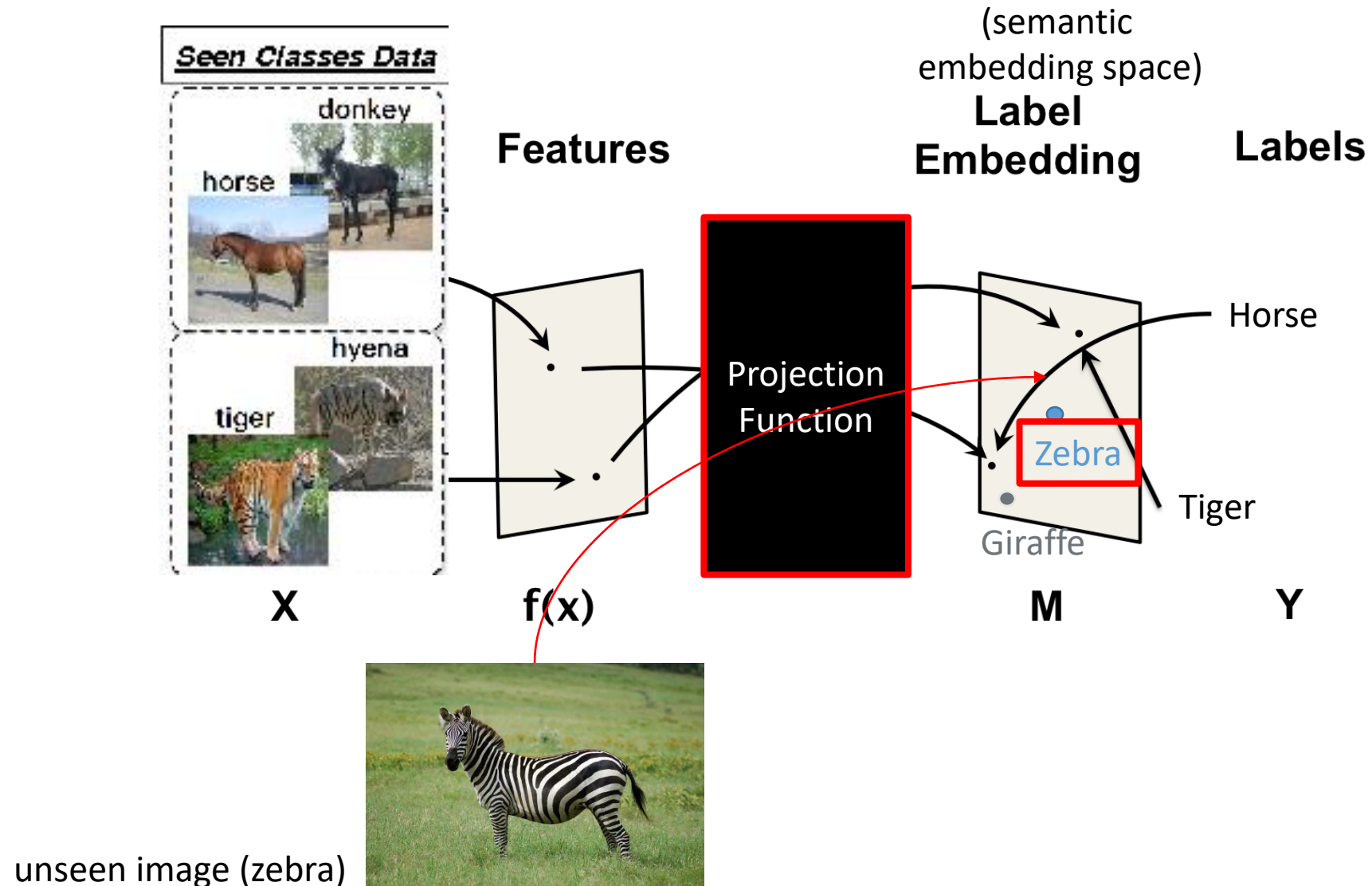
# Zero-Shot Learning?



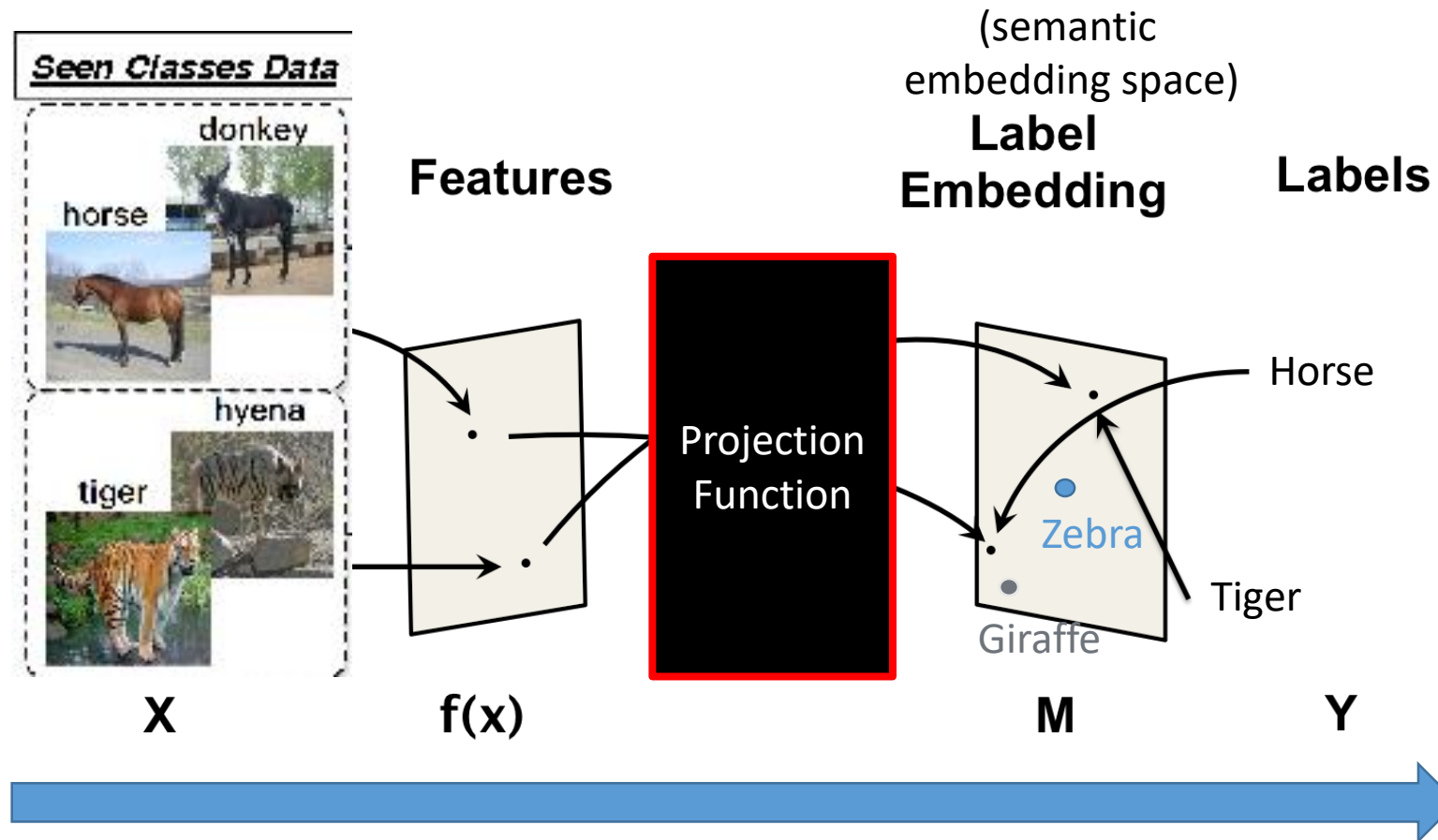
# Zero-Shot Learning?



# Zero-Shot Learning?



# Introduction

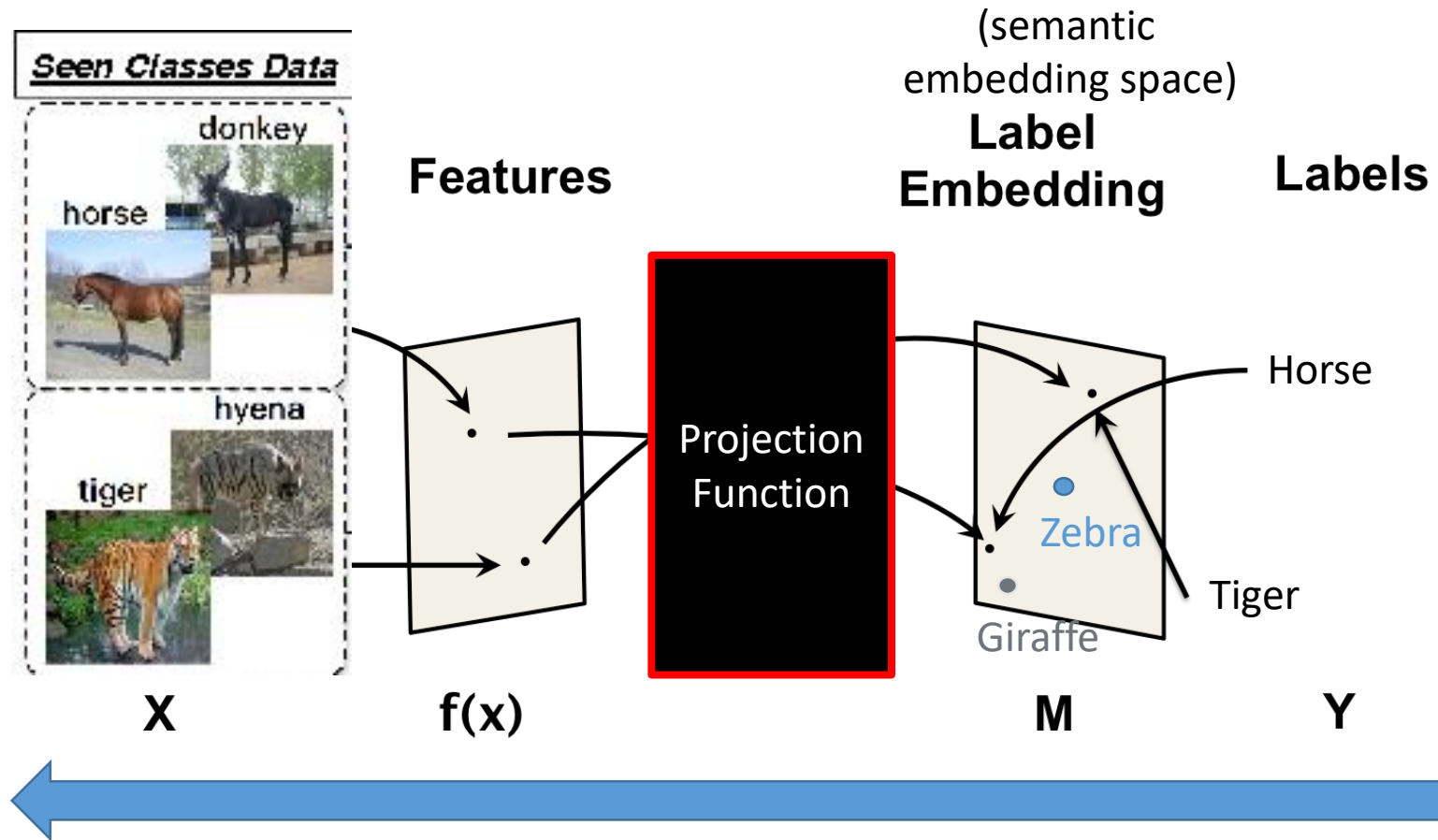


기존 visual-to-semantic embedding

- NN search
  - high dimensional embedding space
  - Less prototypes(labels)
- hubness problem! (특정 prototype에만 편중되어 projection 되는 현상)



# Introduction



1. semantic-to-image embedding



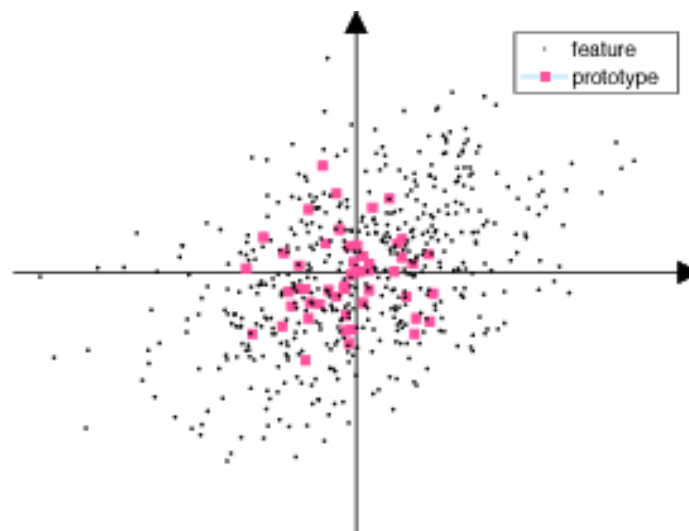
# Introduction

$$\mathcal{L}(\mathbf{W}) = \|\mathbf{B} - \mathbf{W}\mathbf{A}\|_F^2 + \lambda \|\mathbf{W}\|_F^2$$

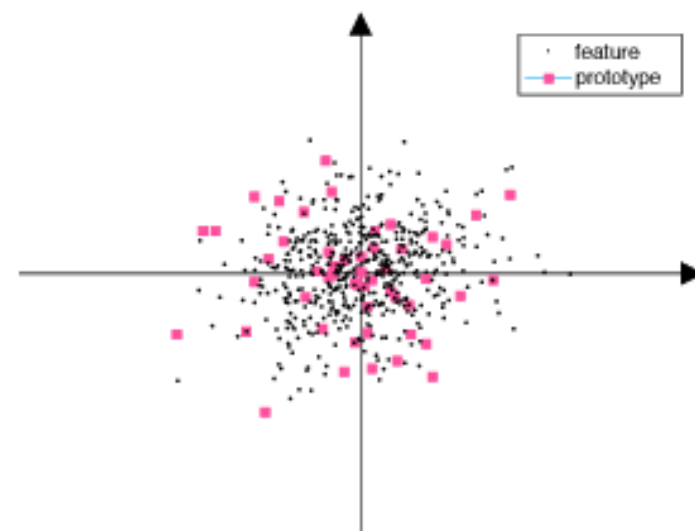
$$\begin{aligned} \|\mathbf{W}\mathbf{A}\|_2 &= \|\mathbf{B}\mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1} \mathbf{A}\|_2 \\ &\leq \|\mathbf{B}\|_2 \|\mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1} \mathbf{A}\|_2 \end{aligned}$$

$$\|\mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1} \mathbf{A}\|_2 = \frac{\sigma^2}{\sigma^2 + \lambda} \leq 1$$

$$\boxed{\|\mathbf{W}\mathbf{A}\|_2 \leq \|\mathbf{B}\|_2}$$

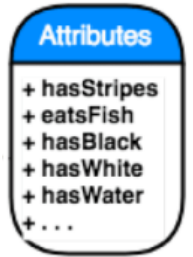
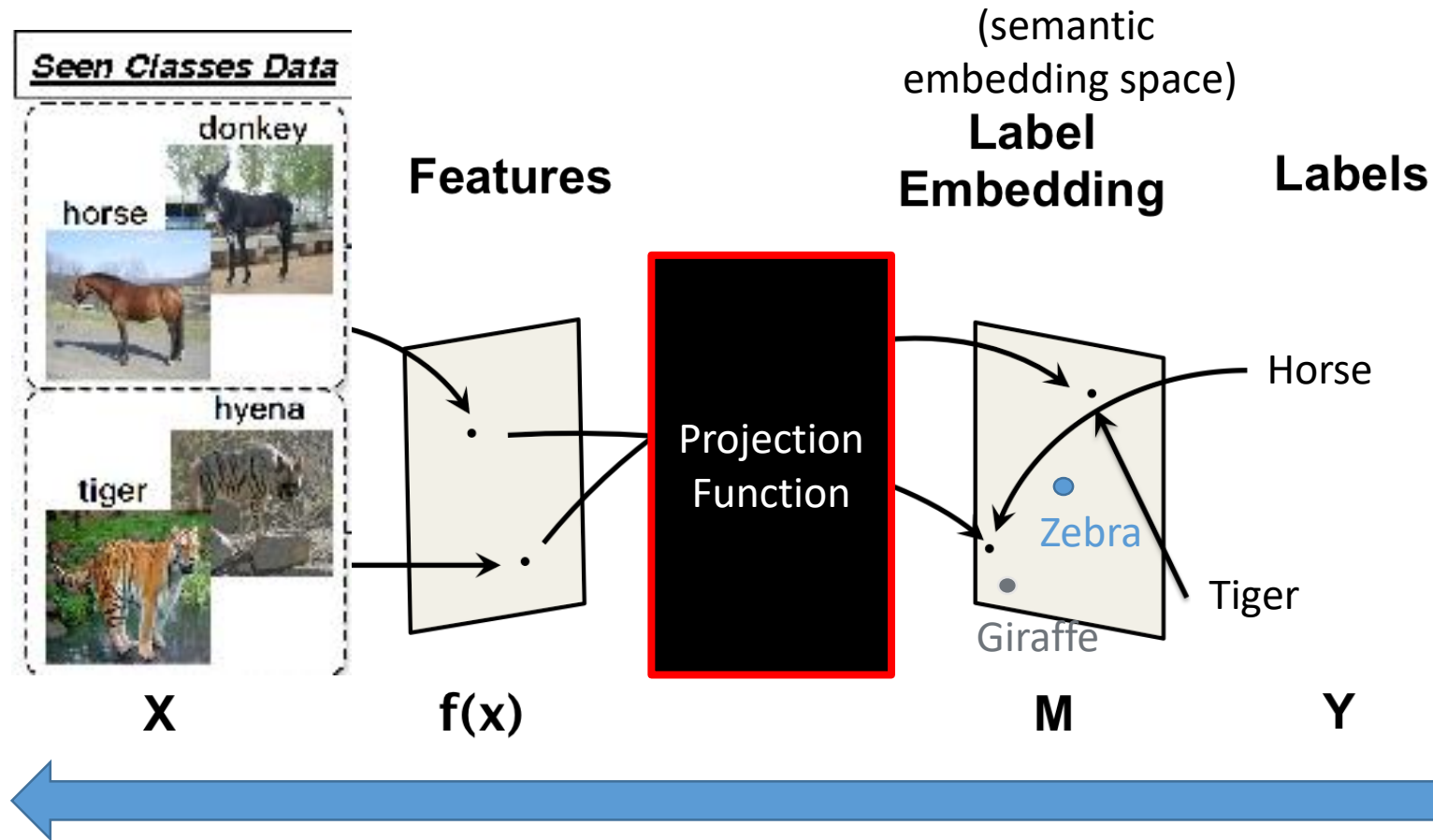


(a)  $S \rightarrow V$

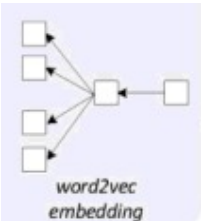


(b)  $V \rightarrow S$

# Introduction



attribute vector



word2vec

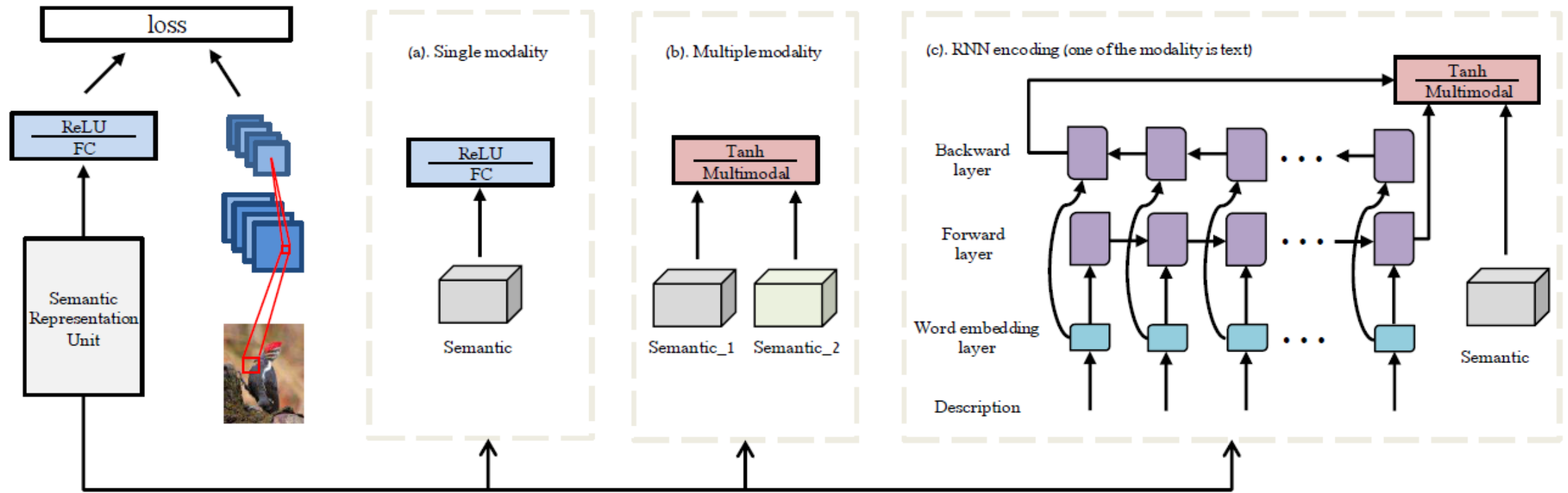
This bird is red in color with a short stubby brown beak and spotted eye ring.

Text description

## 2. Multi-modality fusion method

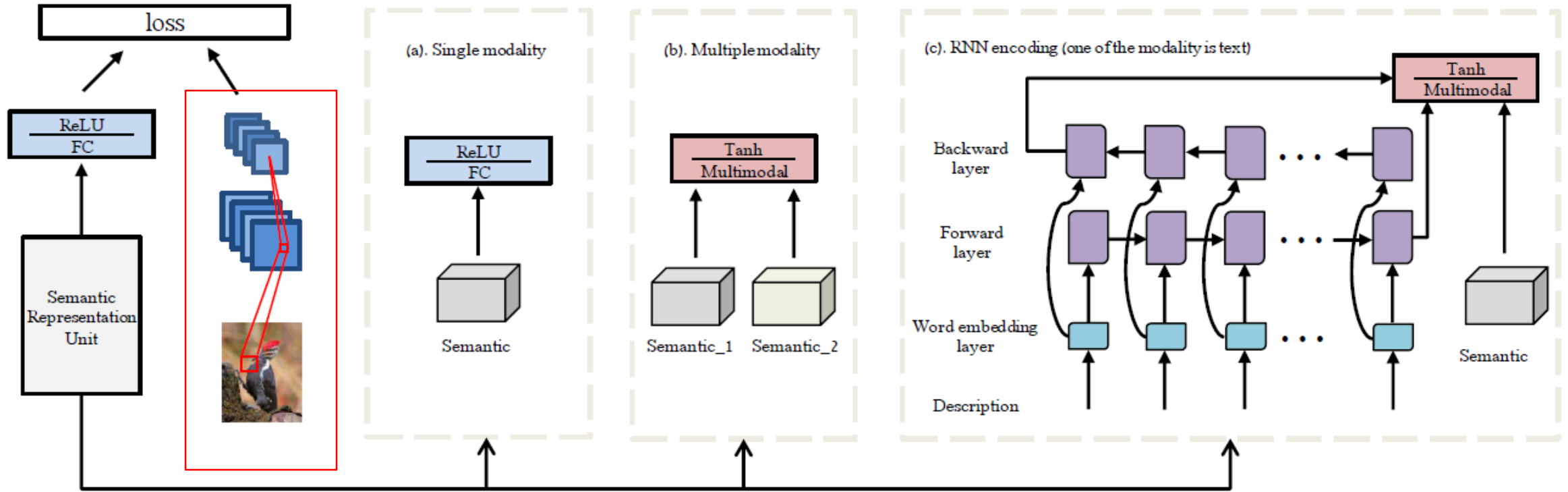
- Combine multiple semantic representation
- Enables end-to-end learning of the semantic space representation

# Methodology



Two main branches

# Methodology

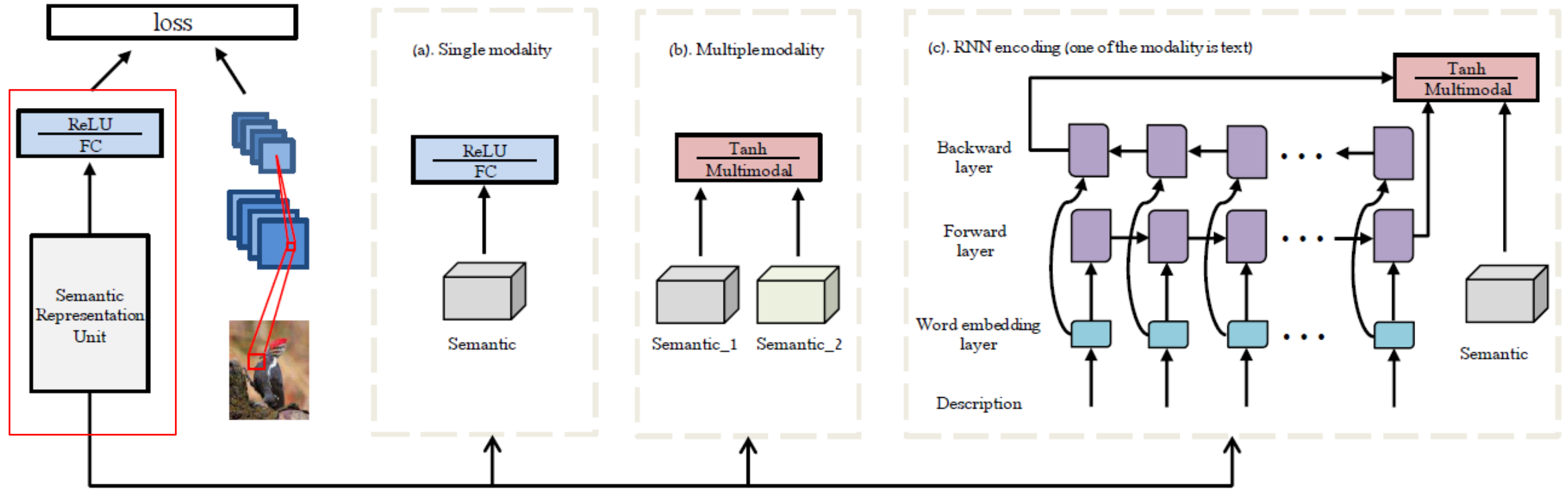


Two main branches

1. Visual encoding

- CNN subnet
- Input : image
- Output :  $D$ -dim feature vector

# Methodology

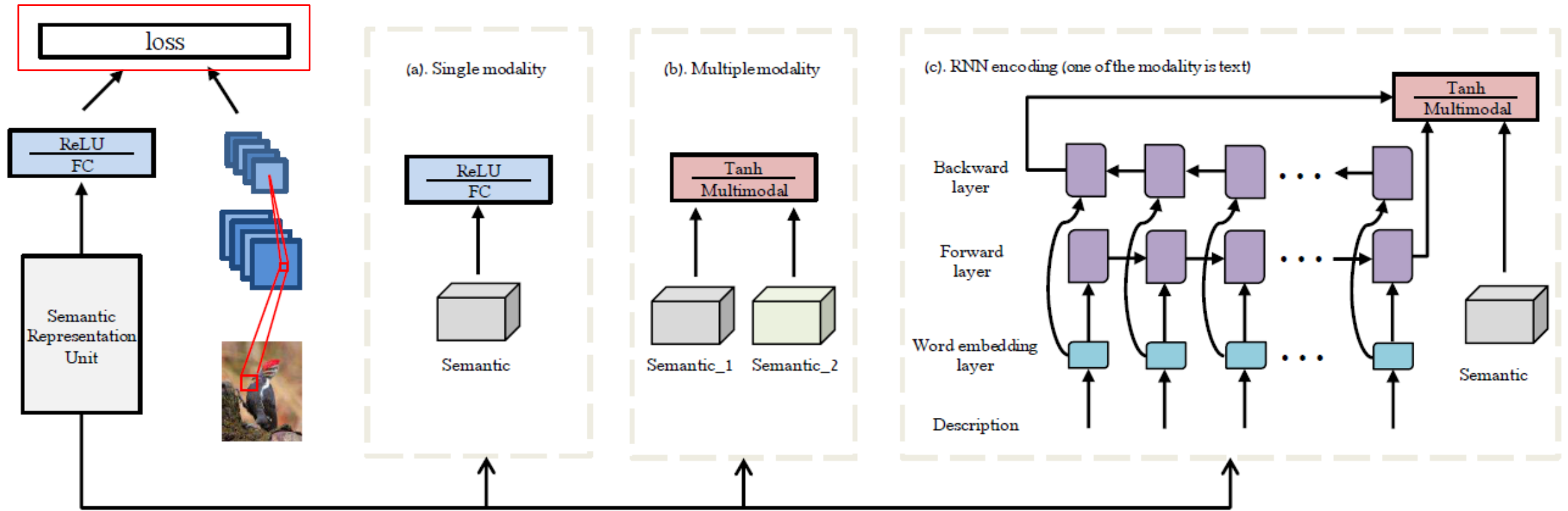


Two main branches

2. Semantic encoding

- Two fully connected layers (ReLU)
- Input :  $L$ -dim semantic representation vector
- Output :  $D$ -dim semantic embedding vector

# Methodology



$$\mathcal{L}(\mathbf{W}_1, \mathbf{W}_2) = \frac{1}{N} \sum_{i=1}^N \|\phi(\mathbf{I}_i) - f_1(\mathbf{W}_2 f_1(\mathbf{W}_1 \mathbf{y}_i^u))\|^2$$

*Embedding loss*

$$+ \lambda (\|\mathbf{W}_1\|^2 + \|\mathbf{W}_2\|^2)$$

*parameter regularization loss*

$\mathbf{W}_1$  : 1<sup>st</sup> FC layer weights ( $L \times M$ )

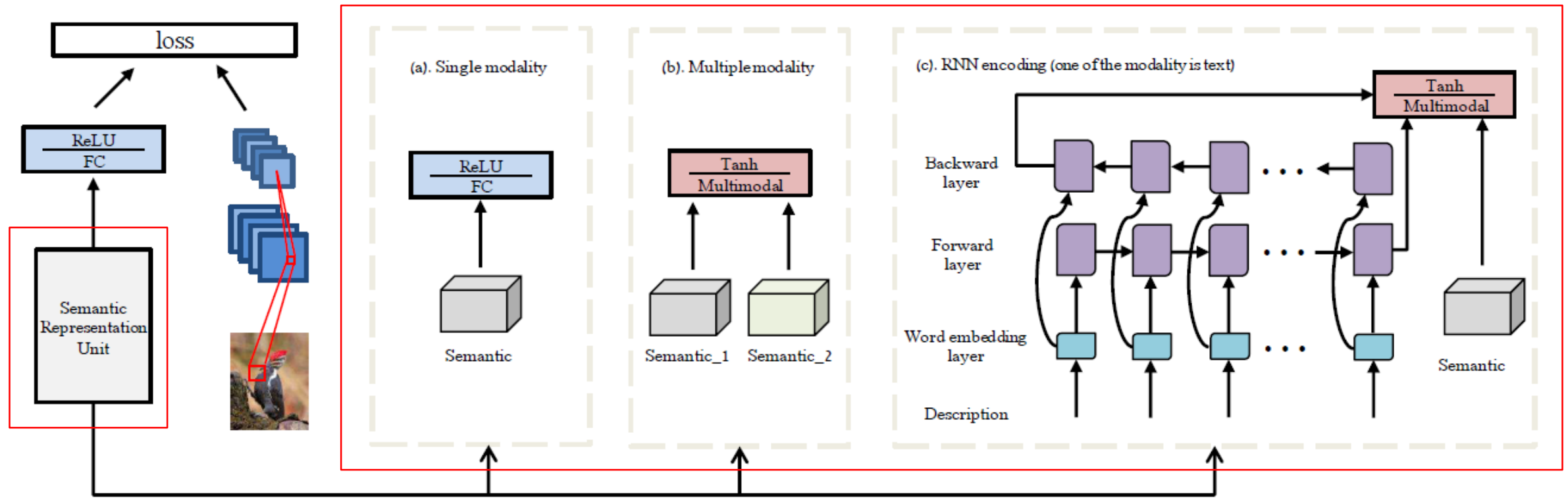
$\mathbf{W}_2$  : 2<sup>nd</sup> FC layer weights ( $M \times D$ )

$\phi(\mathbf{I}_i)$  : image feature vector

$f_1$  : ReLU

$\mathbf{y}$  : semantic representation vector

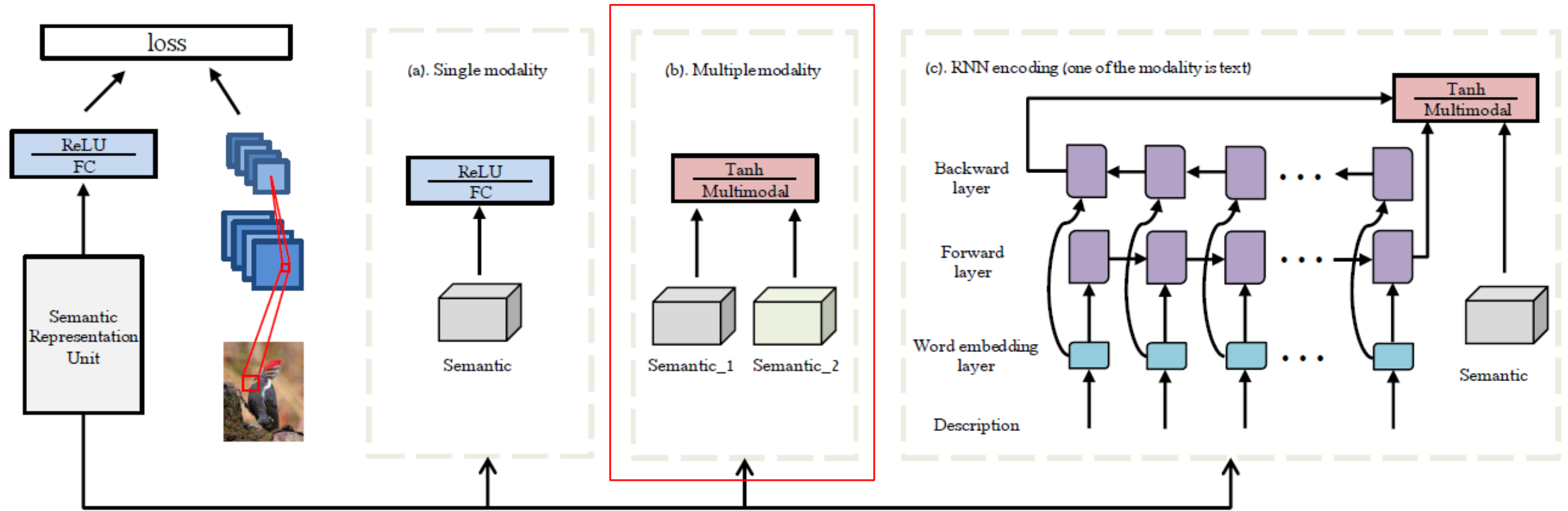
# Methodology



Semantic representation unit  
= semantic representation + 1<sup>st</sup> FC and ReLU



# Methodology



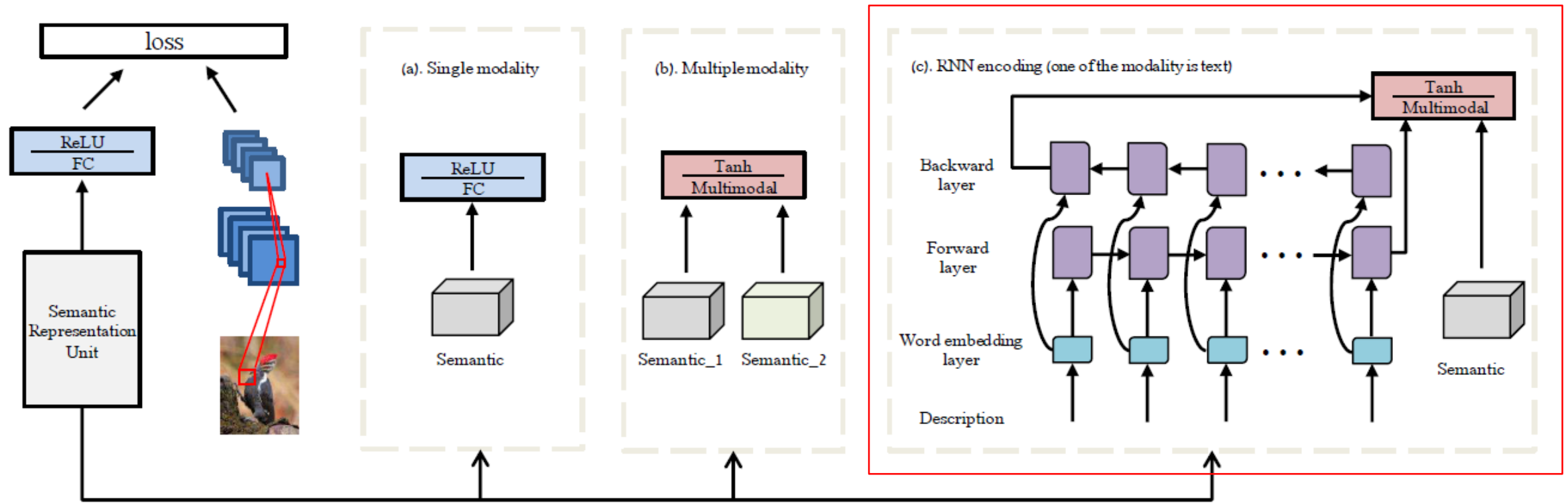
(b) multi-modality

- e.g. an attribute vector and a word vector

$$f_2(\mathbf{W}_1^{(1)} \cdot \mathbf{y}_i^{u_1} + \mathbf{W}_1^{(2)} \cdot \mathbf{y}_i^{u_2})$$

- f2 : element-wise scaled tanh

# Methodology



(c) Bi-LSTM encoder for text description

$$f(\mathbf{W}_{\vec{h}} \cdot \vec{h} + \mathbf{W}_{\overleftarrow{h}} \cdot \overleftarrow{h})$$

- f : f1, ReLU (single) , f2, tanh (multiple)

# Experiment

## 1. AWA and CUB

- AWA : 40 training classes, 10 test classes
  - 1000 dim word vec
  - 85 dim attribute vec
- CUB : 150 training classes, 50 test classes
  - 312 dim attribute vec
  - 10 descriptions per image
- CNN subnet : Inception-V2, 1024 dim

Model	F	SS	AwA	CUB
AMP [14]	$F_O$	A+W	66.0	-
SJE [2]	$F_G$	A	66.7	50.1
SJE [2]	$F_G$	A+W	73.9	51.7
ESZSL [37]	$F_G$	A	76.3	47.2
SSE-ReLU [47]	$F_V$	A	76.3	30.4
JLSE [48]	$F_V$	A	80.5	42.1
SS-Voc [13]	$F_O$	A/W	78.3/68.9	-
SynC-struct [4]	$F_G$	A	72.9	54.5
SEC-ML [3]	$F_V$	A	77.3	43.3
DeViSE [10]	$N_G$	A/W	56.7/50.4	33.5
Socher <i>et al.</i> [43]	$N_G$	A/W	60.8/50.3	39.6
MTMDL [46]	$N_G$	A/W	63.7/55.3	32.3
Ba <i>et al.</i> [24]	$N_G$	A/W	69.3/58.7	34.0
DS-SJE [34]	$N_G$	A/D	-	50.4/56.8
Ours	$N_G$	A/W(D)	86.7/78.8	58.3/53.5
Ours	$N_G$	A+W(D)	88.1	59.0

# Experiment

## 1. ImageNet

- ILSVRC 2010 1K : 800 train, 200 test
- ILSVRC 2012/2010 : 1000 train(2012)  
360 test (2010, disjoint)
- Train word vectors on 4.6M Wikipedia corpus
- Alexnet

Model	hit@5
ConSE [31]	28.5
DeViSE [10]	31.8
Mensink <i>et al.</i> [27]	35.7
Rohrbach [36]	34.8
PST [35]	34.0
AMP [14]	41.0
Ours	<b>46.7</b>
Gaussian Embedding [30]	45.7
PDDM [18]	48.2
Ours	<b>60.7</b>

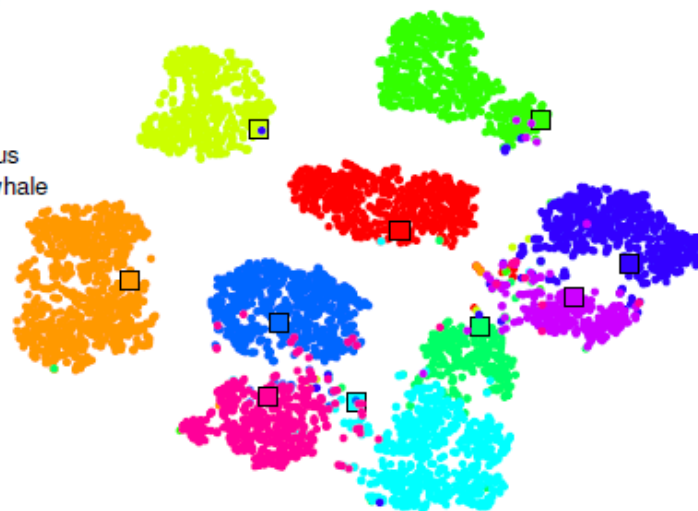
ILSVRC 2010

Model	hit@1	hit@5
ConSE [31]	7.8	15.5
DeViSE [10]	5.2	12.8
AMP [14]	6.1	13.1
SS-Voc [13]	9.5	16.8
Ours	<b>11.0</b>	<b>25.7</b>

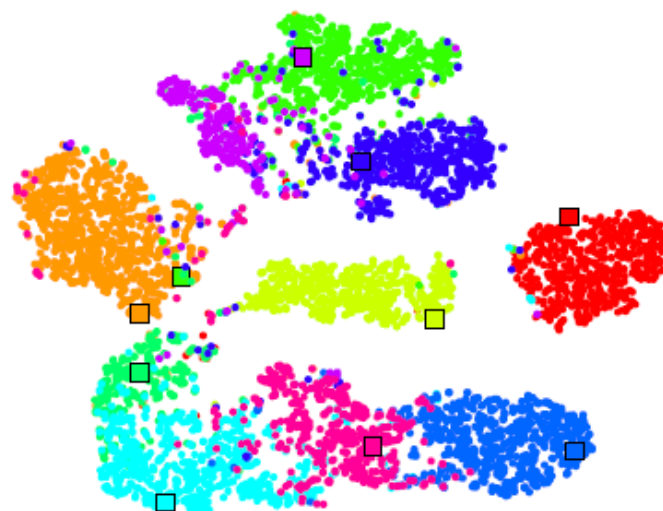
ILSVRC 2012/2010

# Experiment

- chimpanzee
- giant panda
- leopard
- persian cat
- pig
- hippopotamus
- humpback whale
- raccoon
- rat
- seal



(a)  $S \rightarrow V$



(b)  $V \rightarrow S$

$N_1$ skewness	AwA	CUB
Visual $\rightarrow$ Semantic	0.4162	8.2697
Semantic $\rightarrow$ Visual	<b>-0.4834</b>	<b>2.2594</b>

Model	AwA	CUB
Linear regression ( $V \rightarrow S$ )	54.0	40.7
Linear regression ( $S \rightarrow V$ )	74.8	45.7
Ours	<b>86.7</b>	<b>58.3</b>

$$(N_k \text{ skewness}) = \frac{\sum_{i=1}^l (N_k(i) - E[N_k])^3 / l}{Var[N_k]^{\frac{3}{2}}}$$

Loss	Visual $\rightarrow$ Semantic	Semantic $\rightarrow$ Visual
Least square loss	60.6	<b>86.7</b>
Hinge loss	57.7	72.8