Paper Review

Prophet Attention: Predicting Attention with Future Attention

Liu et al., NeurIPS, 2020

Myeongsup Kim

Integrated M.S./Ph.D. Student
Data Science & Business Analytics Lab.
School of Industrial Management Engineering
Korea University

Myeongsup_kim@korea.ac.kr

Introduction

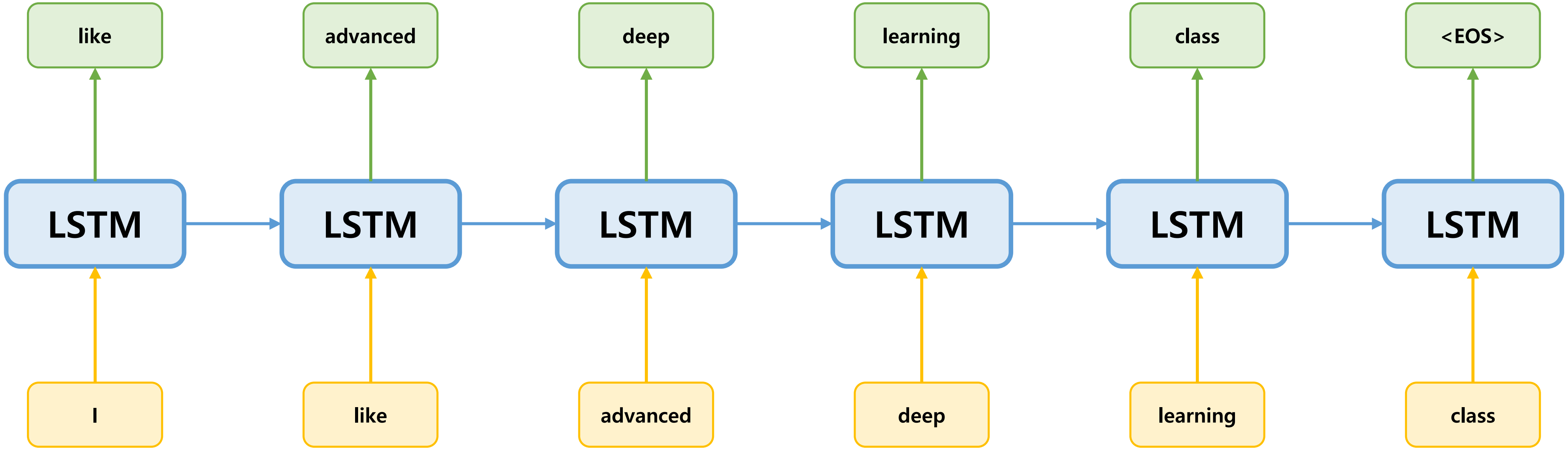
- Image Captioning
- Deviated Focus

<COCO 2015 Image Captioning Task>

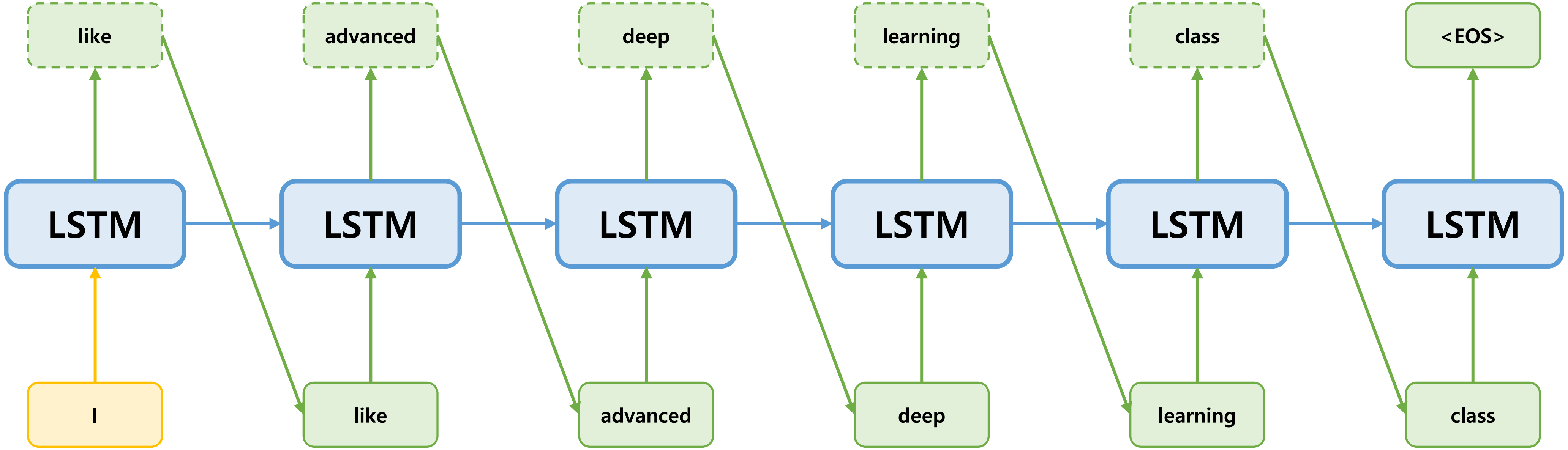
The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

<Text Generation>



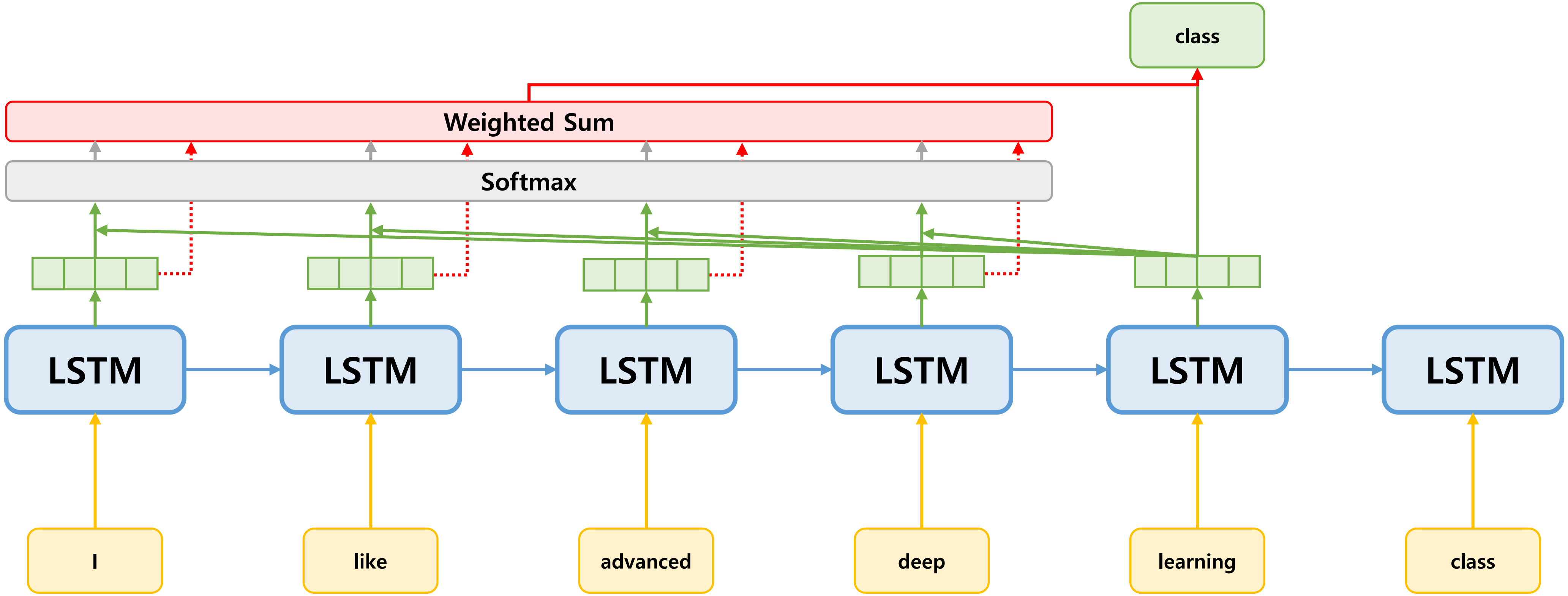
<Text Generation>



Introduction

-Image Captioning

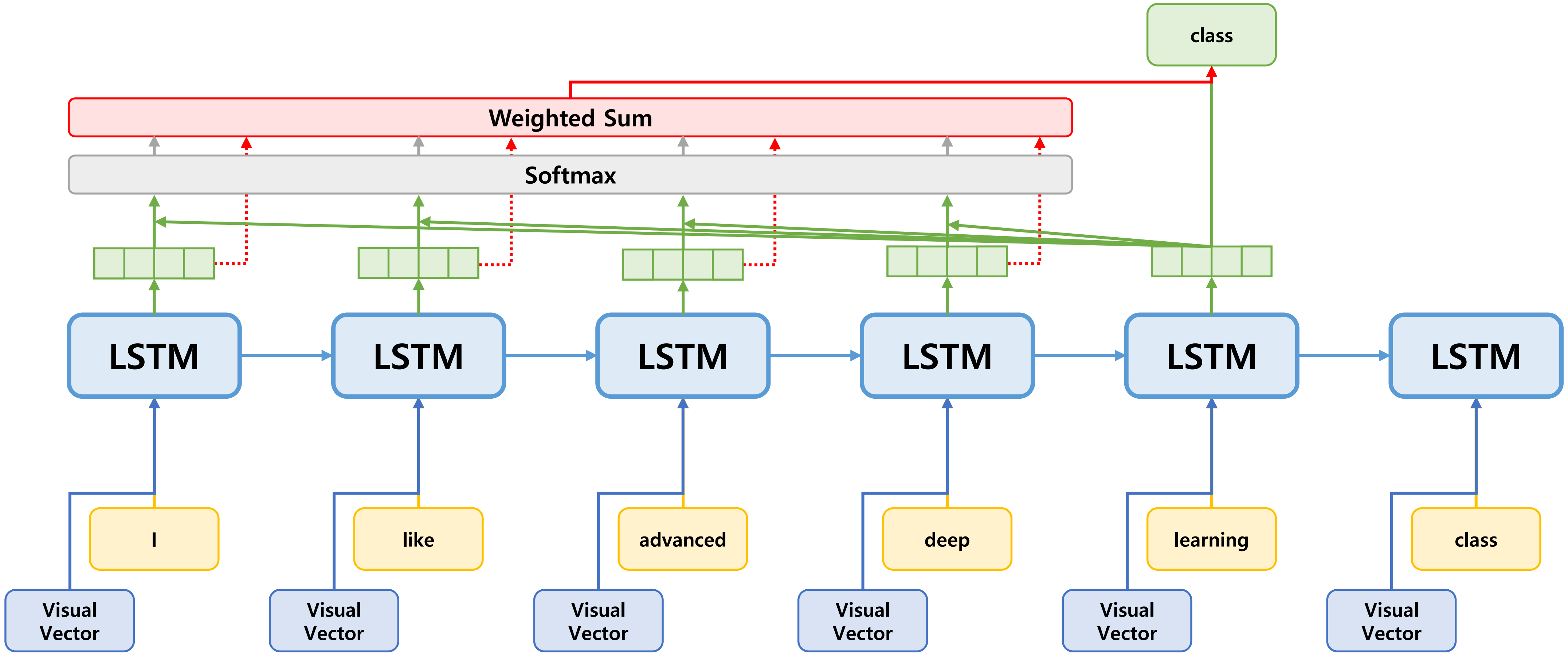
<Attention>



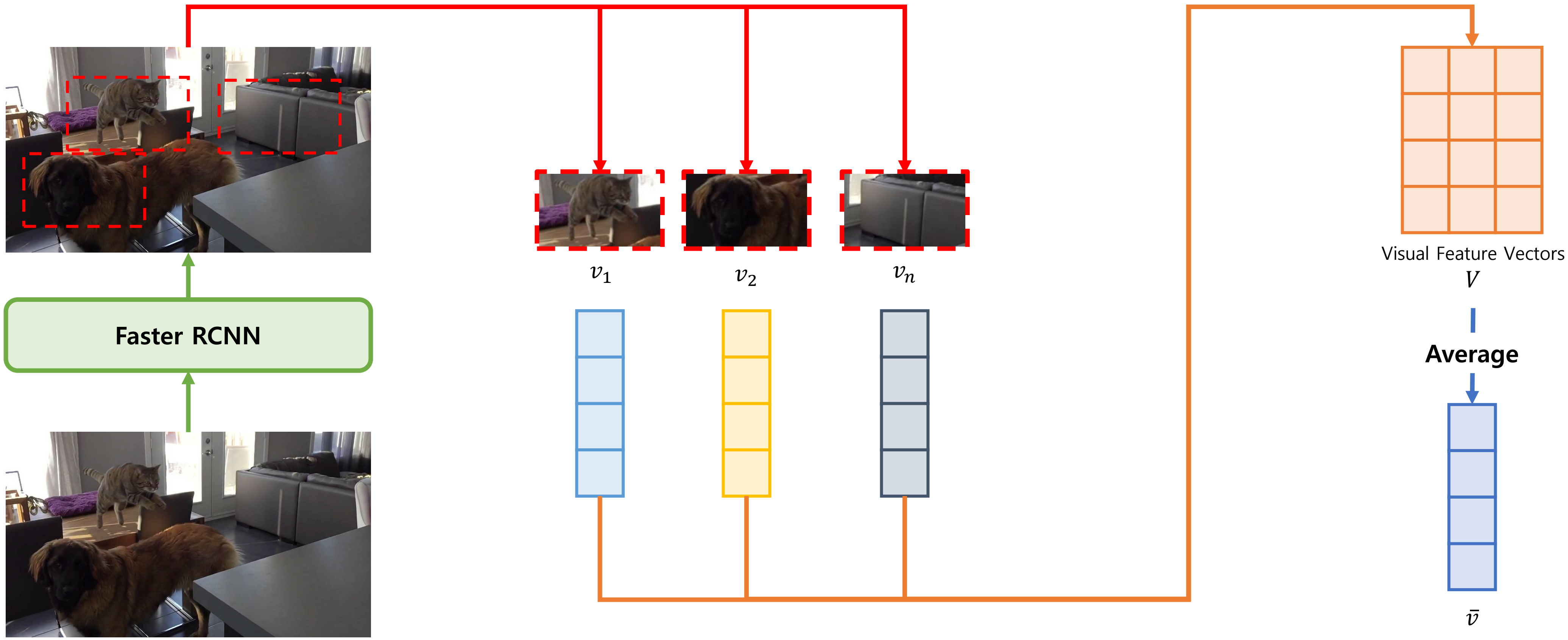
Introduction

-Image Captioning

<Visual Text Generation>



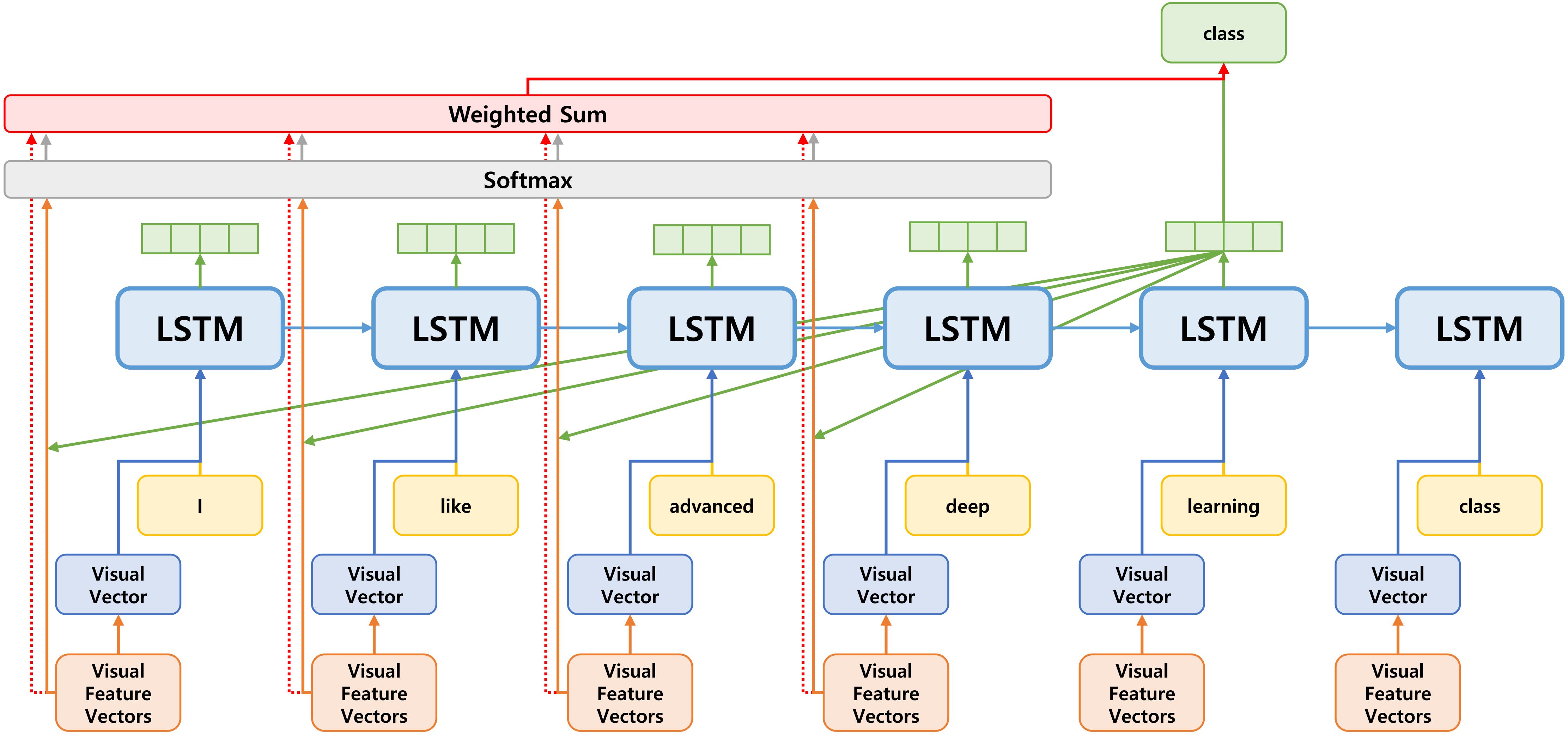
<Visual Encoder>



Introduction

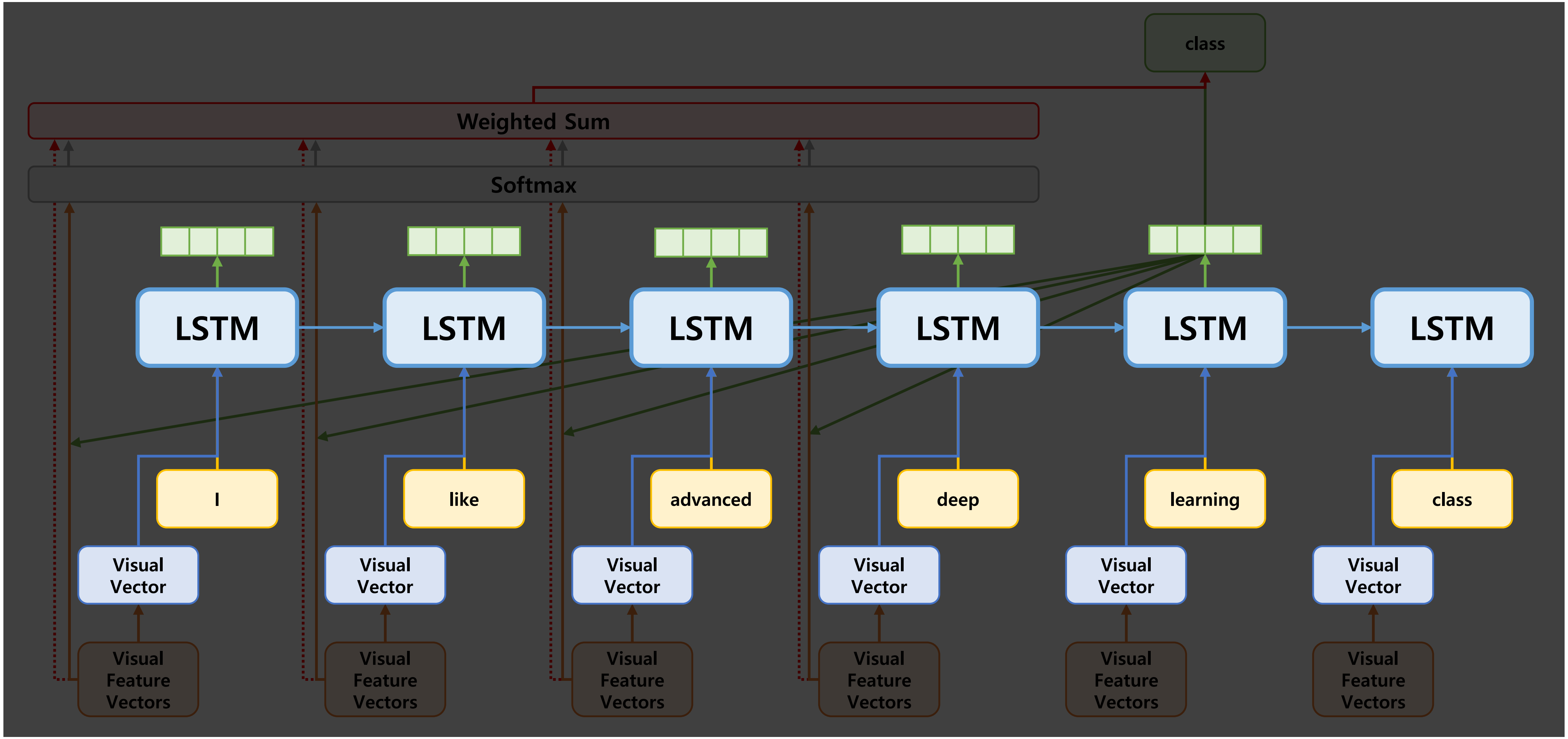
-Deviated Focus

<Visual Text Generation with Attention>



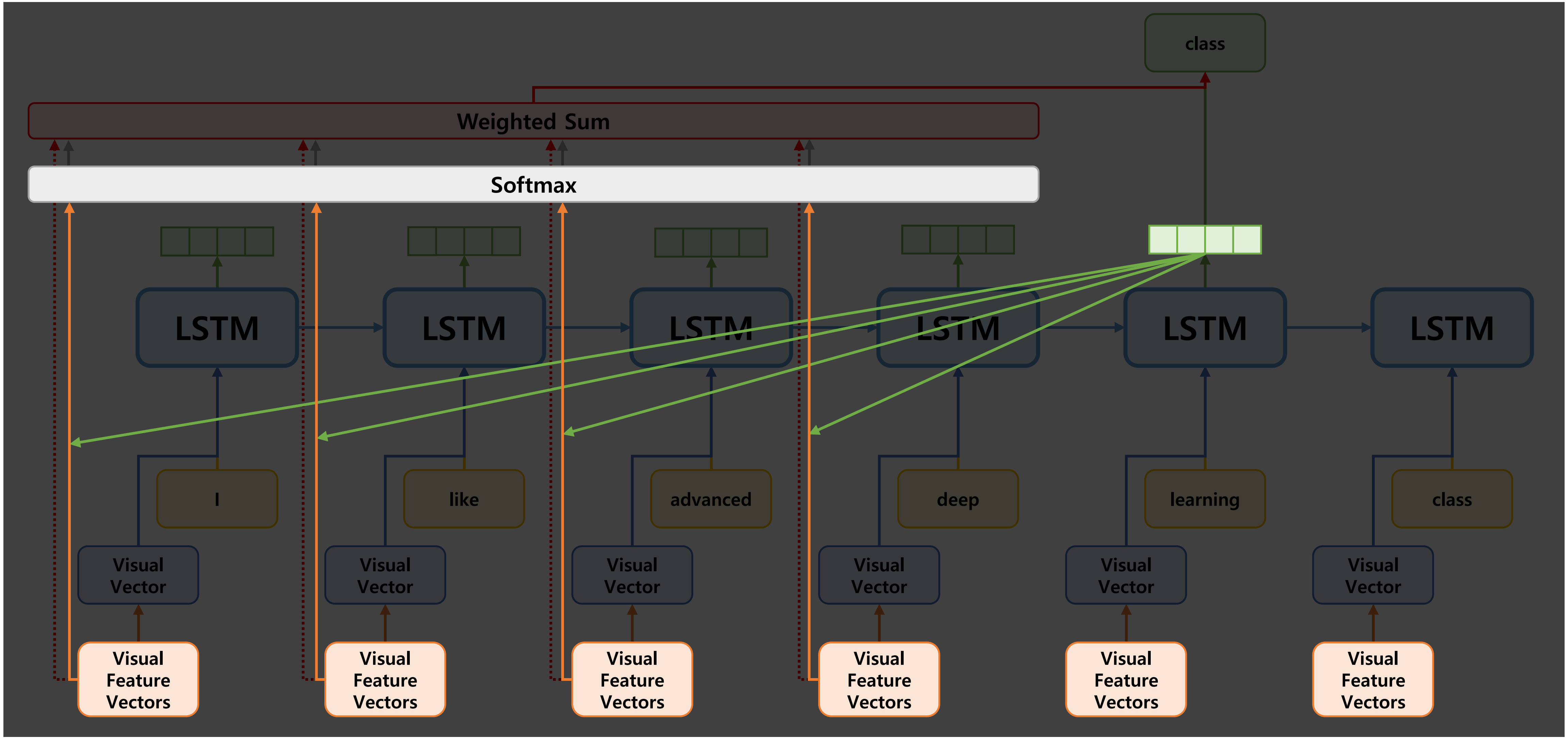
Introduction
-Deviated Focus

<Visual Text Generation with Attention>



Introduction
-Deviated Focus

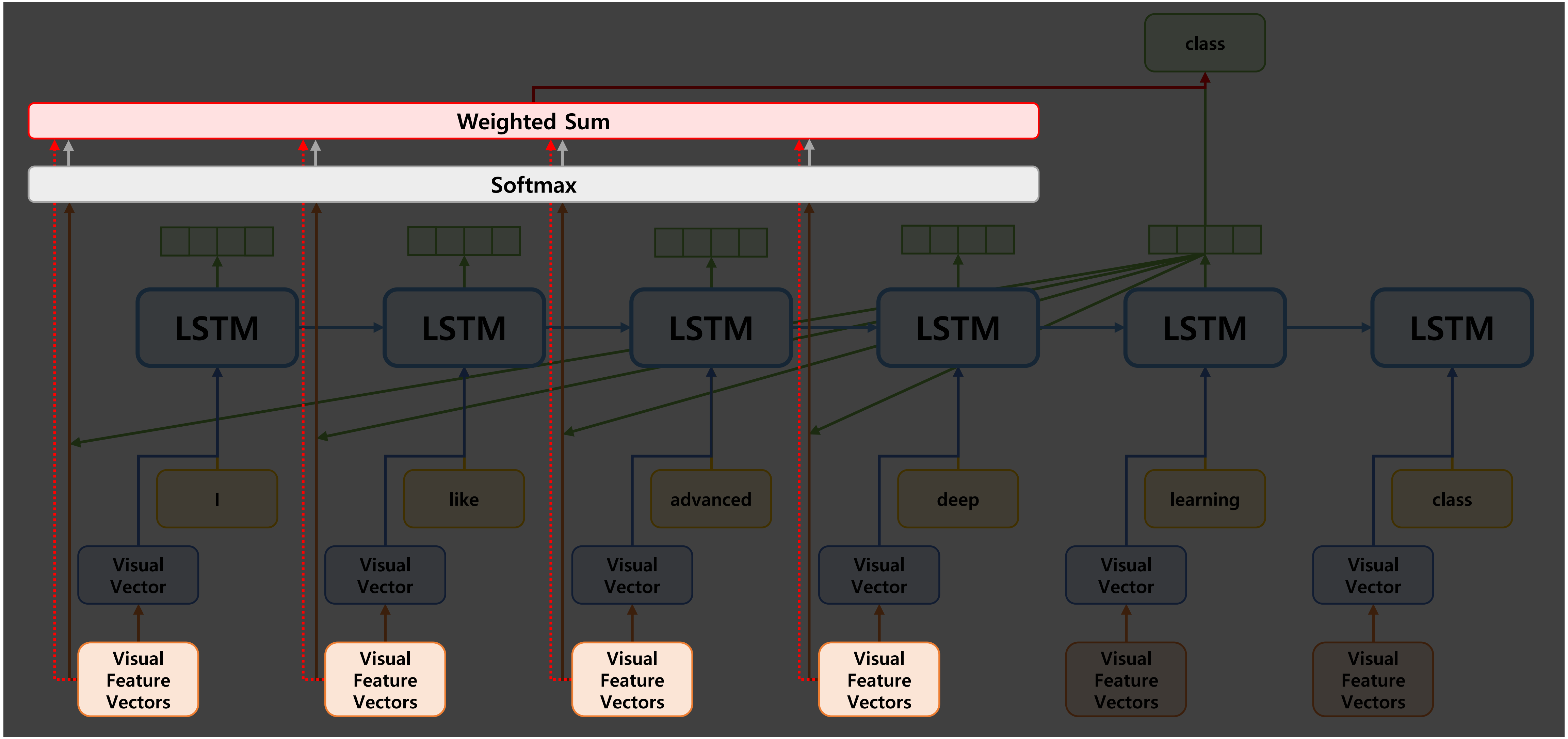
<Visual Text Generation with Attention>



Introduction

-Deviated Focus

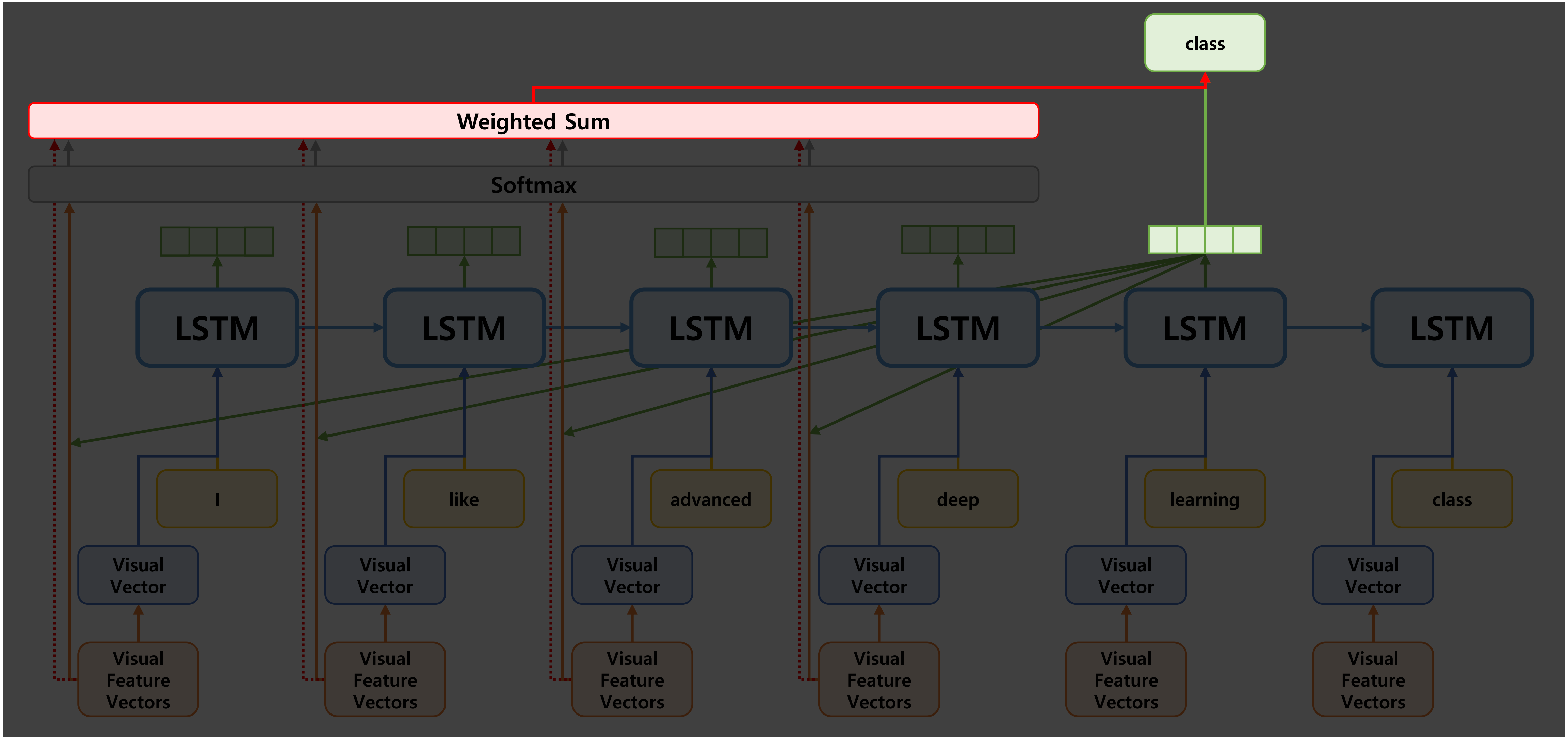
<Visual Text Generation with Attention>



Introduction

-Deviated Focus

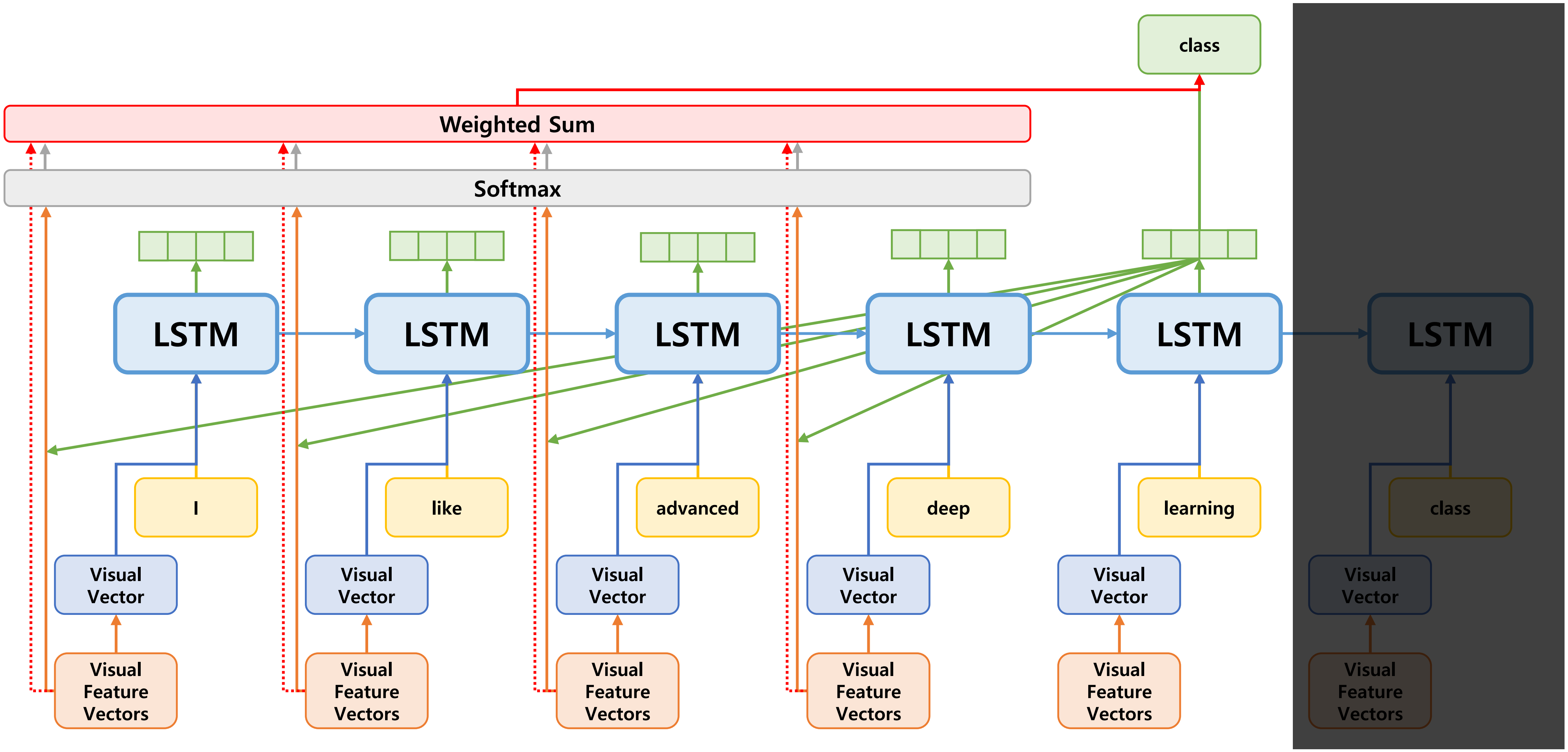
<Visual Text Generation with Attention>



Introduction

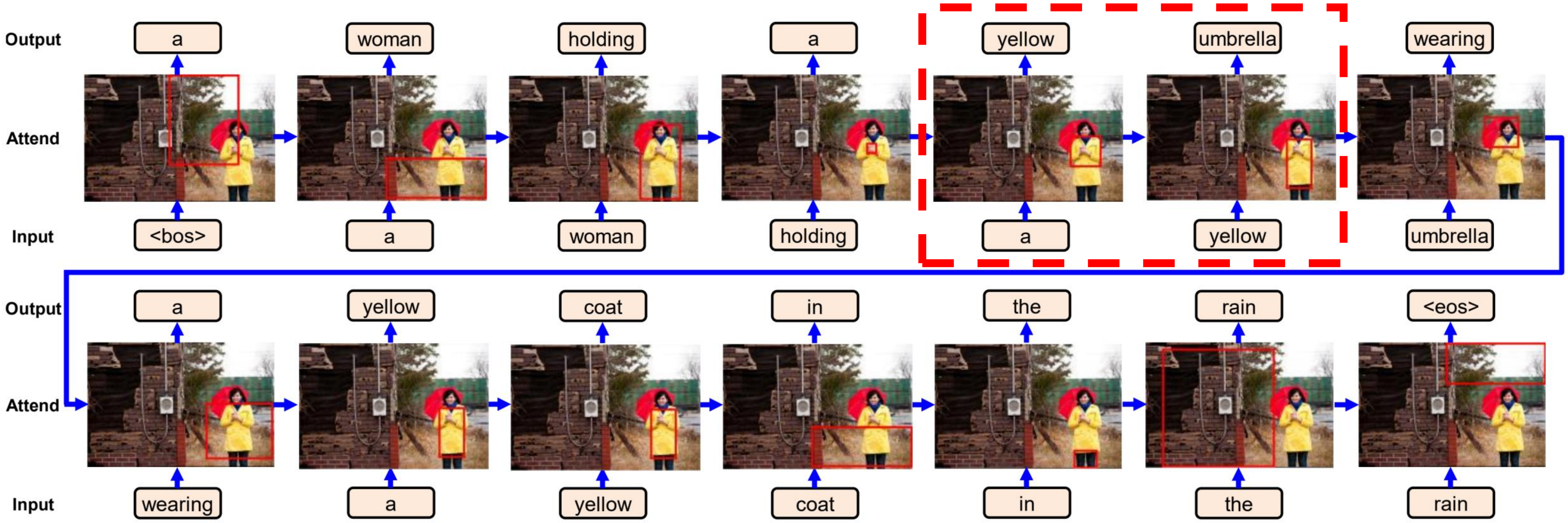
-Deviated Focus

<Visual Text Generation with Attention>



Introduction
-Deviated Focus

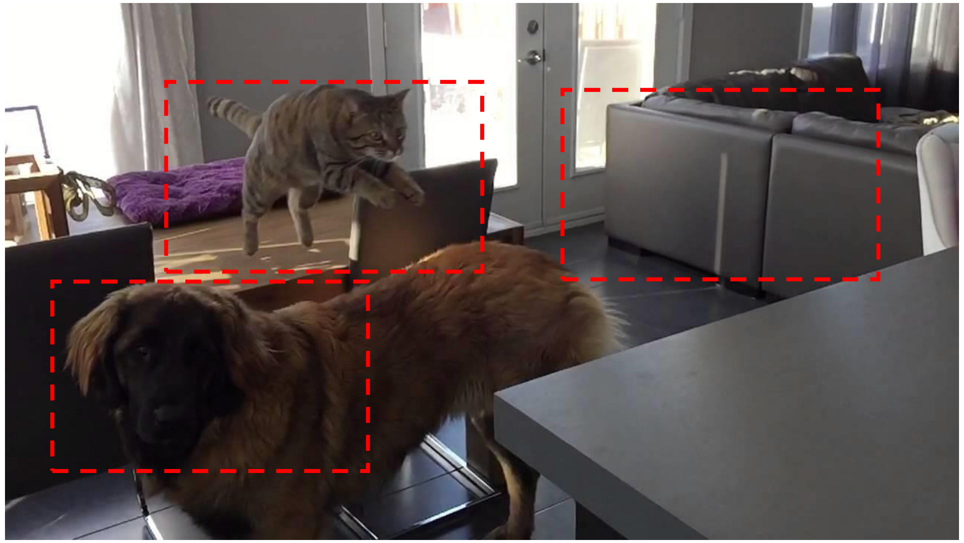
<Deviated Focus>



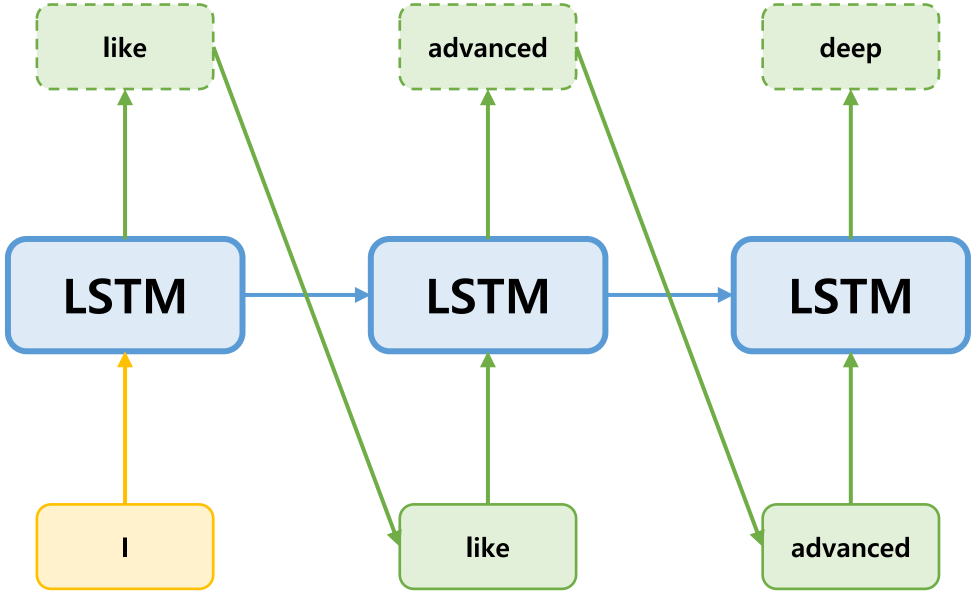
Introduction

-Deviated Focus

<Deviated Focus>



<Grounding>



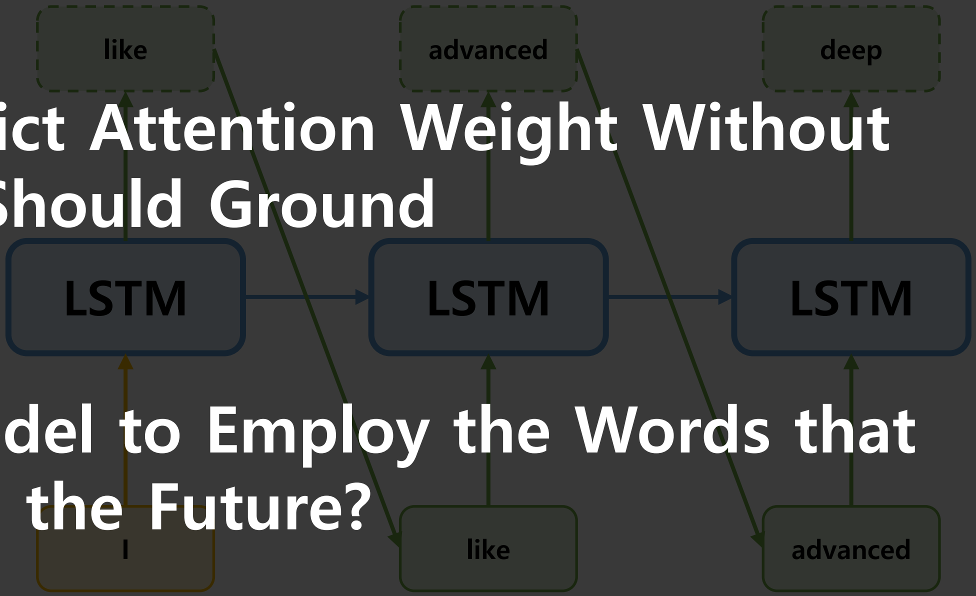
<Generation>

<Deviated Focus>

Current Attention Model Has to Predict Attention Weight Without Knowing the Word It Should Ground

How Can We Enable the Attention Model to Employ the Words that Will be Generated in the Future?

<Grounding>



<Generation>

Prophet Attention: Predicting Attention with Future Attention

Liu et al., NeurIPS, 2020

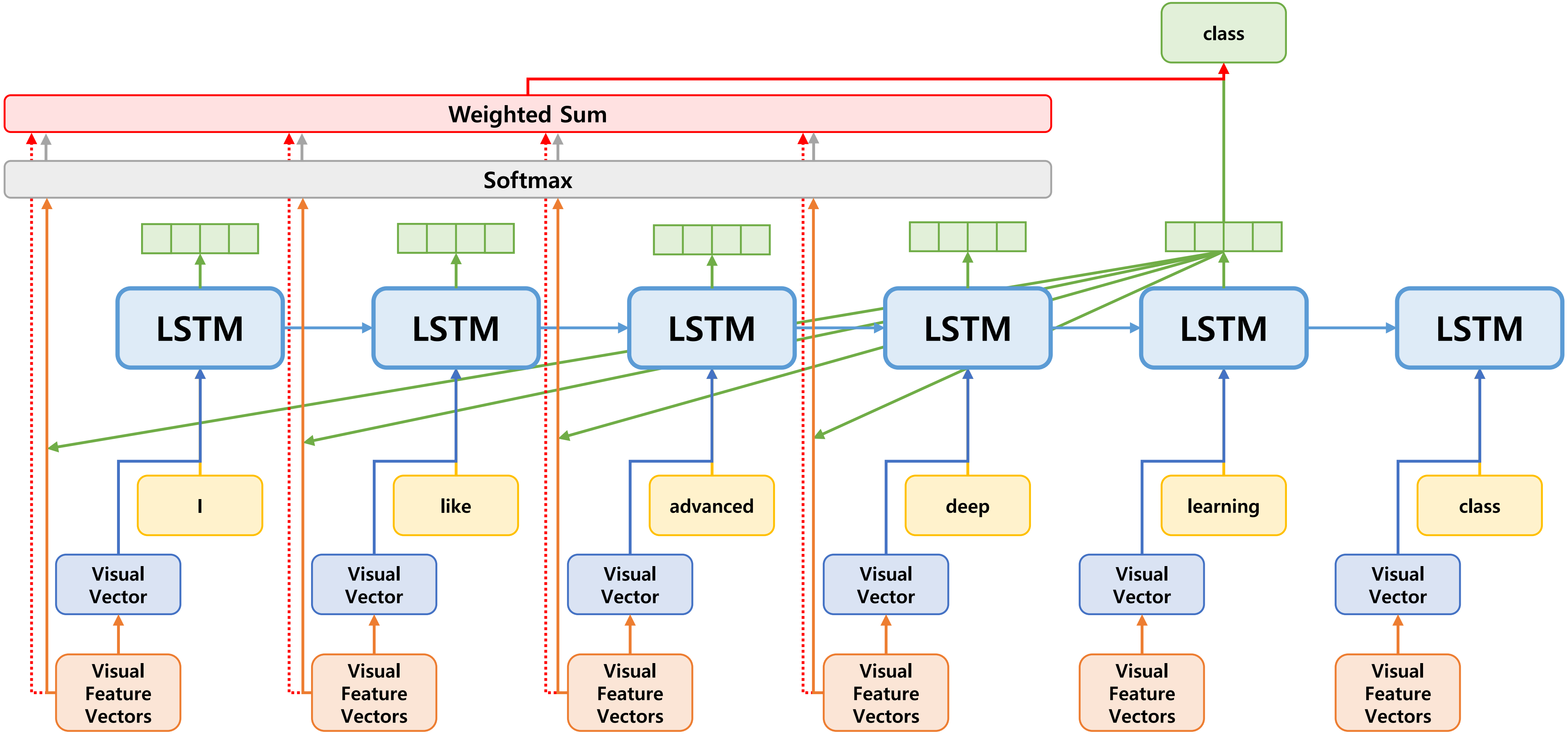
Prophet Attention: Predicting Attention with Future Attention

Liu et al., NeurIPS, 2020

Pre-requisites

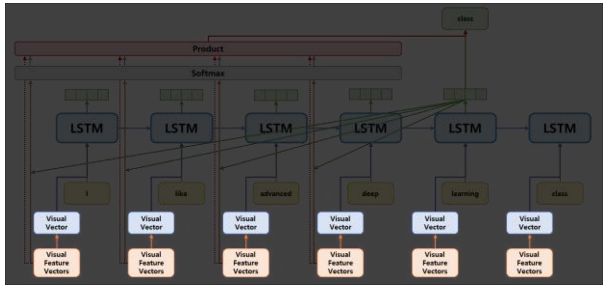
- Attention-Enhanced Encoder-Decoder Framework
- Visual Encoder
- Attention-Enhanced Caption Decoder

<Attention-Enhanced Encoder-Decoder Framework>



Pre-requisites

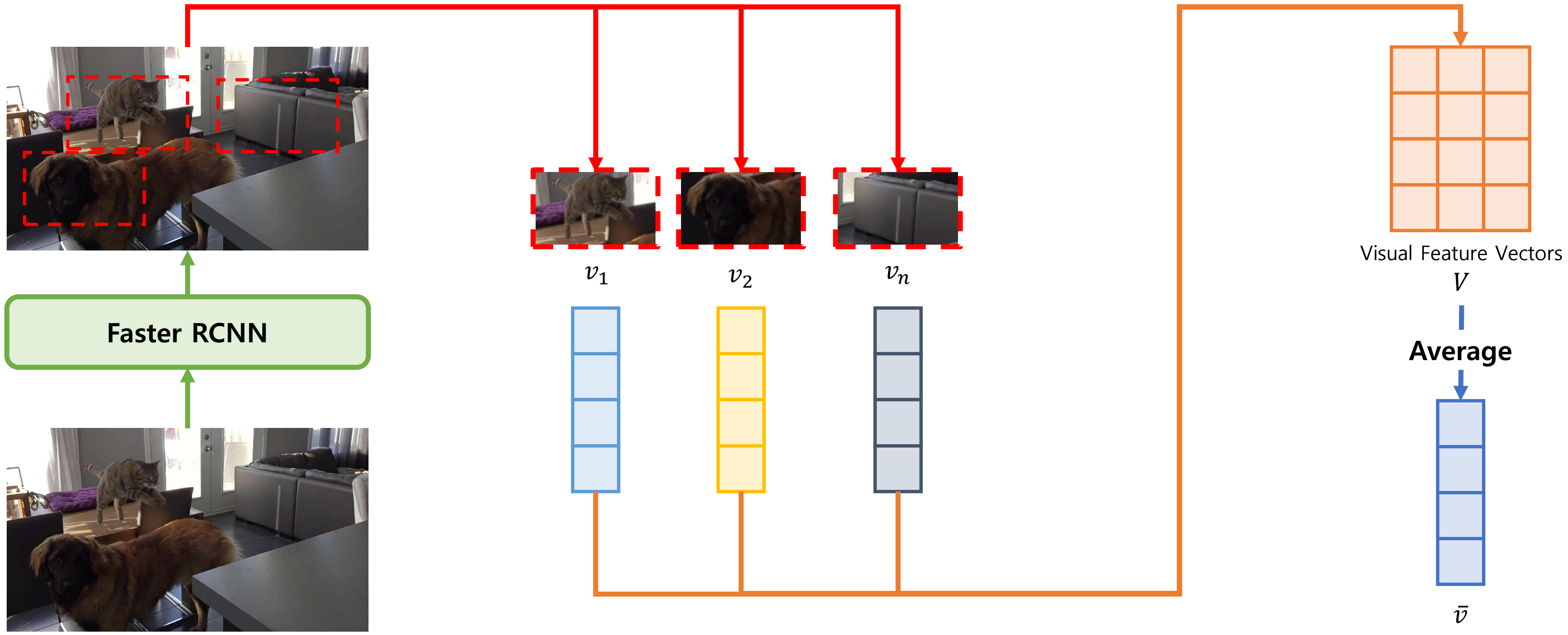
-Visual Encoder



<Visual Encoder>

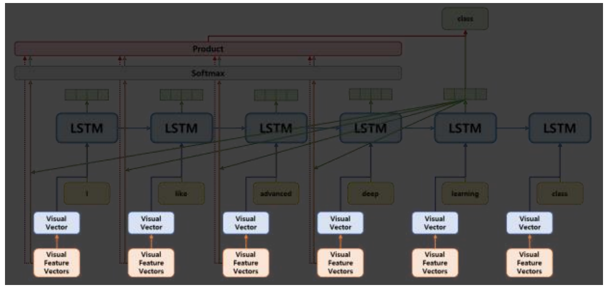
$$V = \{v_1, v_2, \cdots, v_N\} \in \mathbb{R}^{d \times N}$$

$$\bar{v} = \frac{1}{k}\sum_{i=1}^{k} v_i$$



Pre-requisites

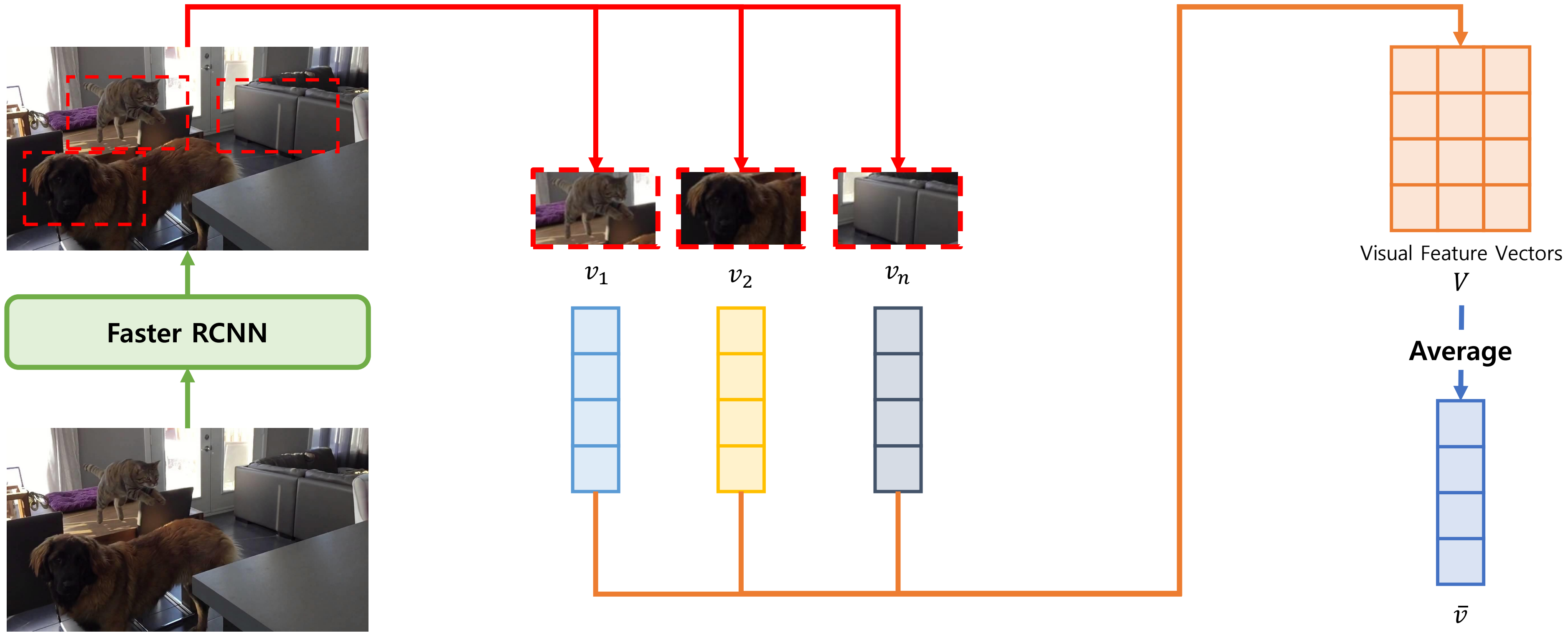
-Visual Encoder

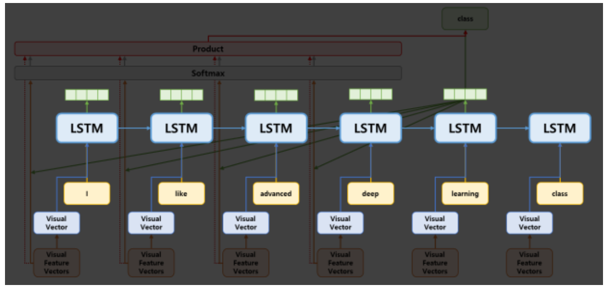


<Visual Encoder>

$$V = \{v_1, v_2, \cdots, v_N\} \in \mathbb{R}^{d \times N}$$

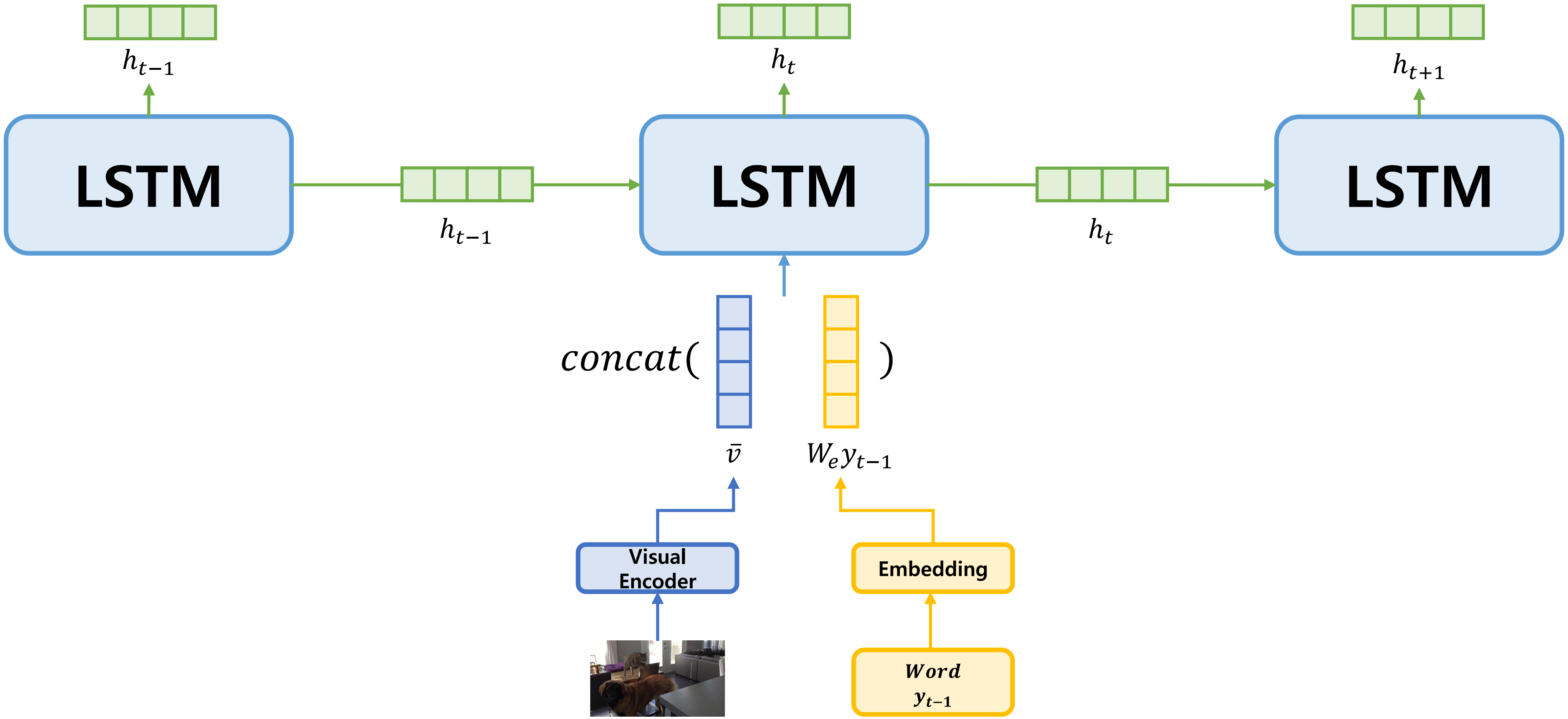
$$\bar{v} = \frac{1}{k}\sum_{i=1}^{k} v_i$$

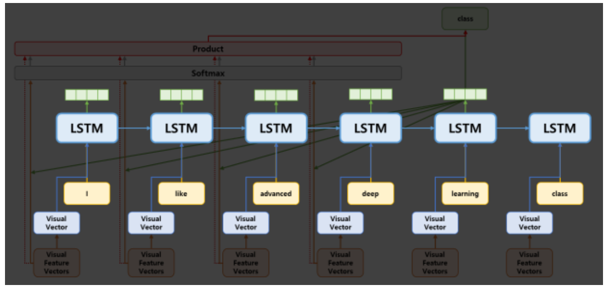




<Attention-Enhanced Caption Decoder>

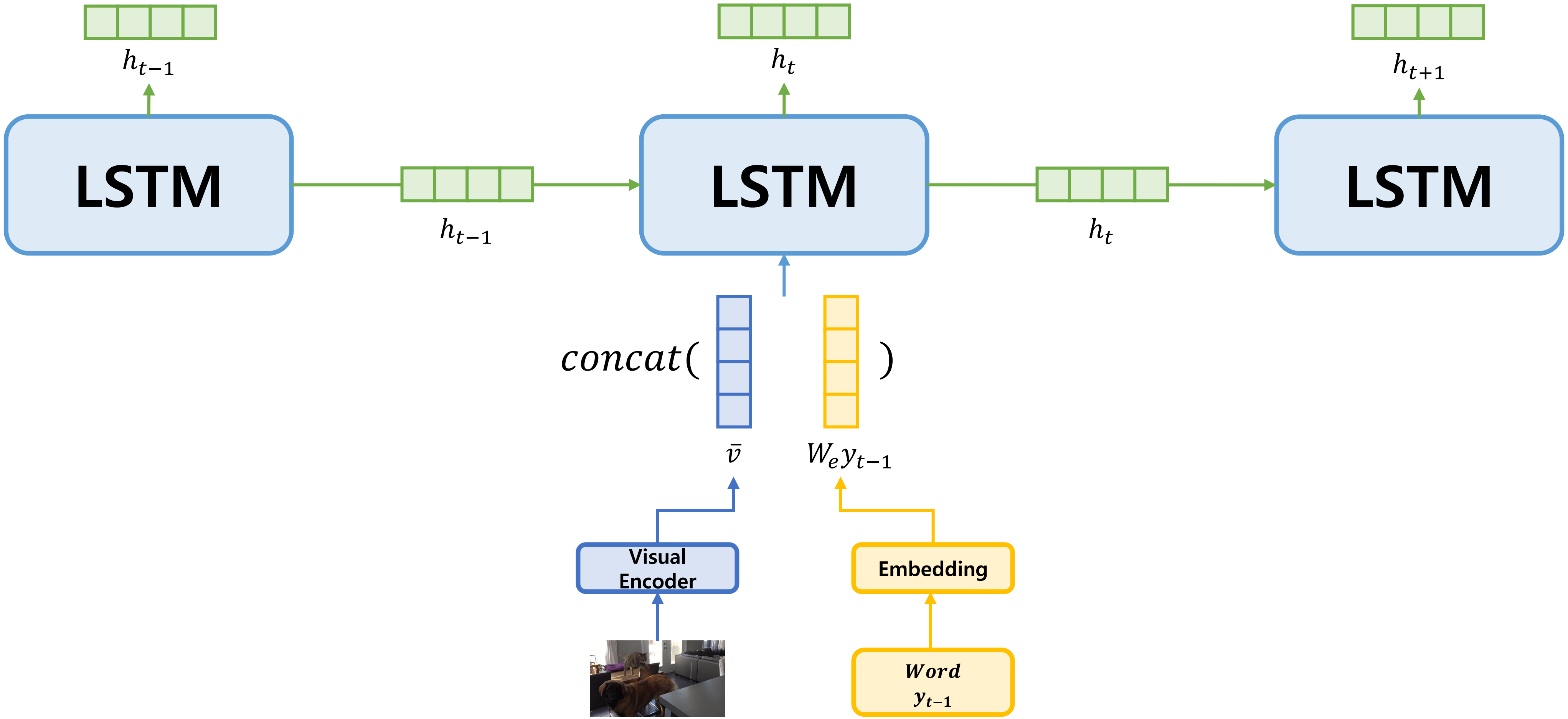
$$h_t = \mathrm{LSTM}(h_{t-1}, [\,W_e\, y_{t-1};\ \bar{v}])$$
$$W_e: Embedding\ Parameter$$



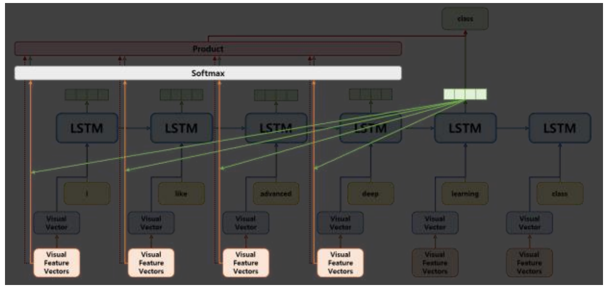


<Attention-Enhanced Caption Decoder>

$$h_t = \mathrm{LSTM}(h_{t-1}, [\, W_e\, y_{t-1};\ \bar{v}])$$
$$W_e : Embedding\ Parameter$$

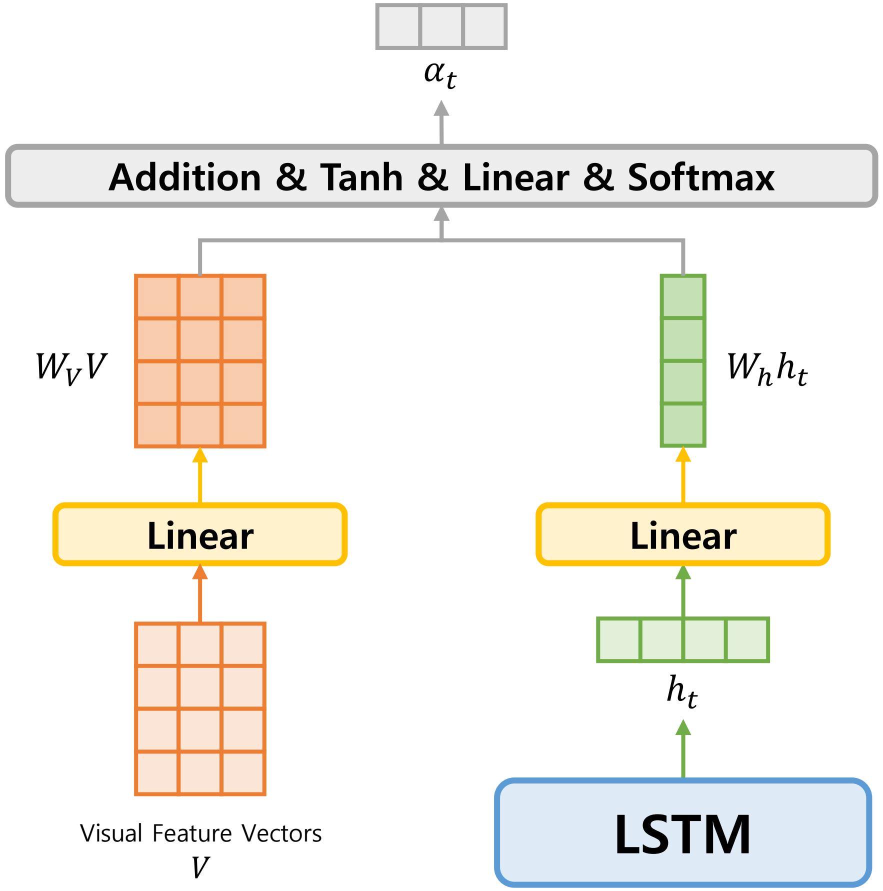


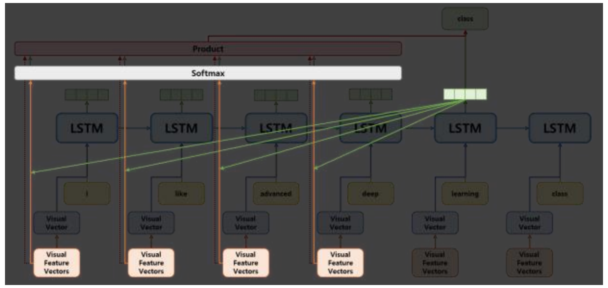
Pre-requisites
-Attention-Enhanced Caption Decoder



<Attention-Enhanced Caption Decoder>

$$\alpha_t = f_{Att}(h_t, V) = \mathrm{softmax}\,(w_\alpha \tanh(W_h h_t \oplus W_V V))$$
$$\oplus : Matrix\text{-}Vector\ Addition$$
$$W_h : Hidden\ State, W_V : Visual\ Feature, w_\alpha : Attention$$



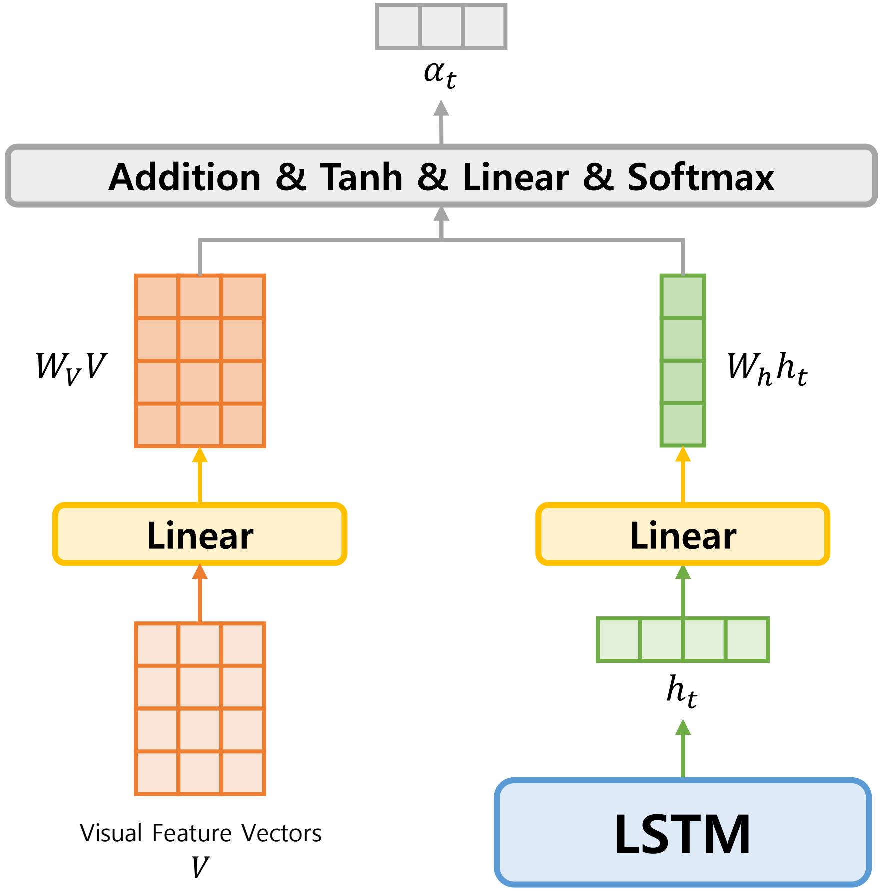


<Attention-Enhanced Caption Decoder>

$$\alpha_t = f_{Att}(h_t, V) = \mathrm{softmax}\,(w_\alpha \tanh(W_h h_t \oplus W_V V))$$

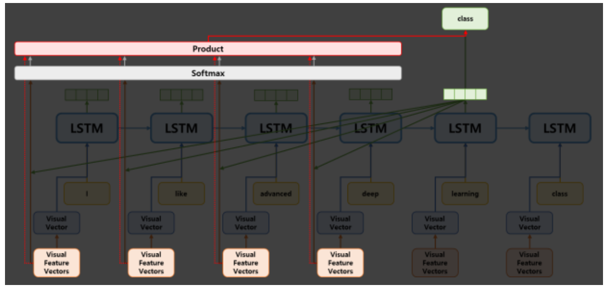
$$\oplus : Matrix\text{-}Vector\ Addition$$

$$W_h : Hidden\ State, W_V : Visual\ Feature, w_\alpha : Attention$$



Pre-requisites

-Attention-Enhanced Caption Decoder

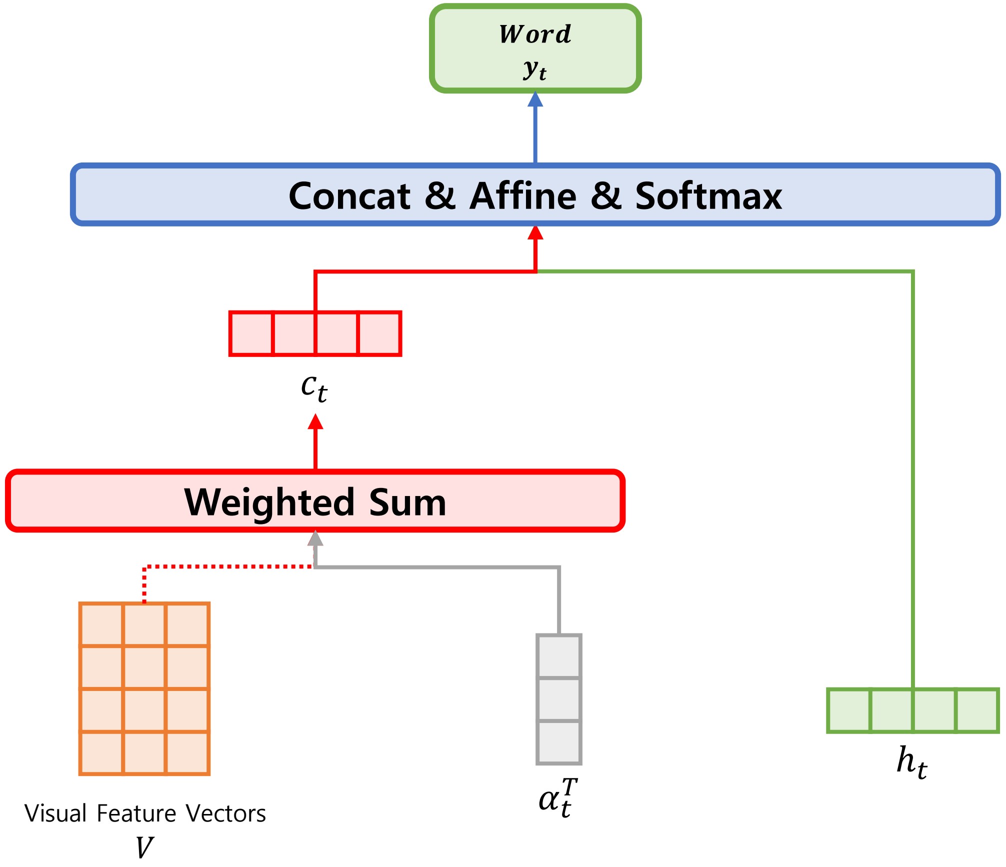


<Attention-Enhanced Caption Decoder>

$$c_t = V\alpha_t^T$$

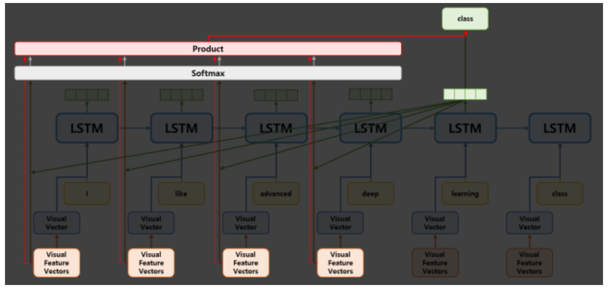
$$y_t \sim p_t = softmax\left(W_p[h_t; c_t] + b_p\right)$$

$$W_p, b_p : Prediction$$



Pre-requisites

-Attention-Enhanced Caption Decoder

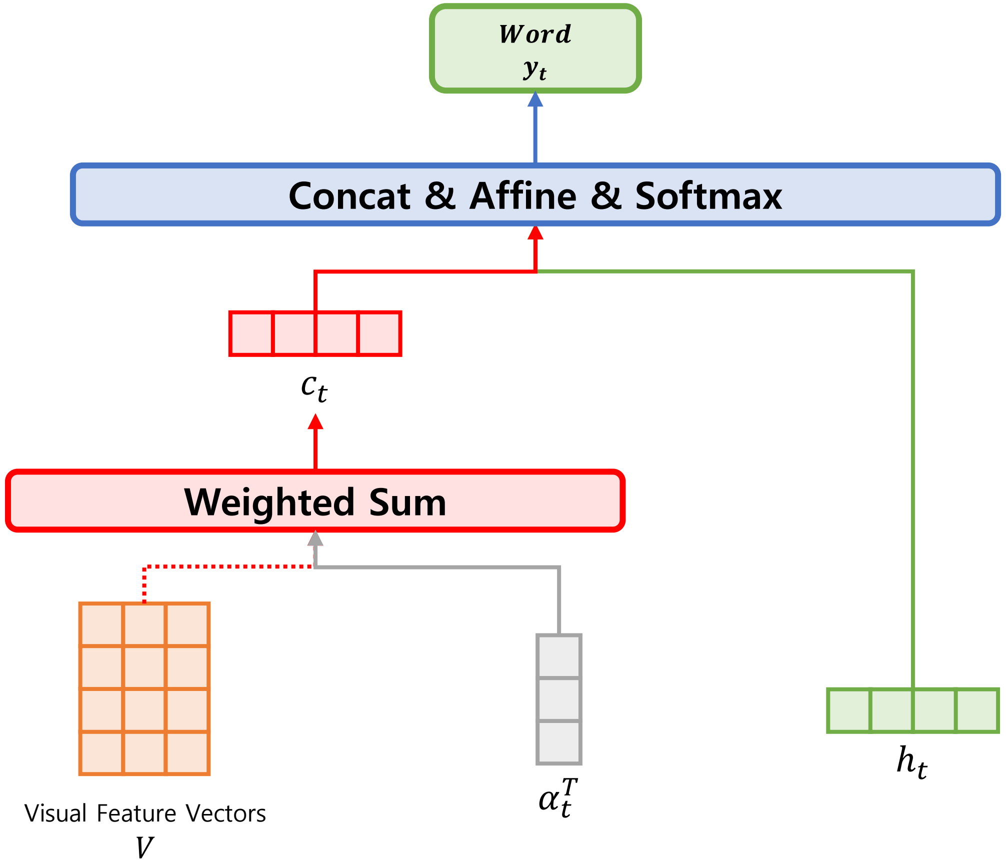


<Attention-Enhanced Caption Decoder>

$$c_t = V\alpha_t^T$$

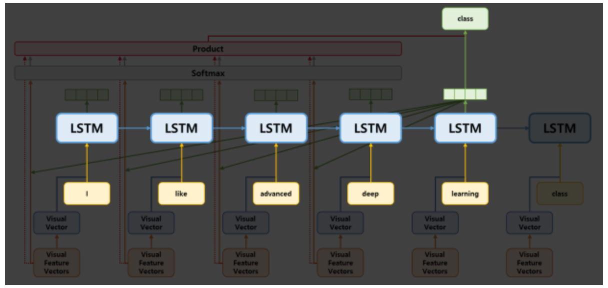
$$y_t \sim p_t = softmax\left(W_p[h_t; c_t] + b_p\right)$$

$$W_p, b_p : Prediction$$



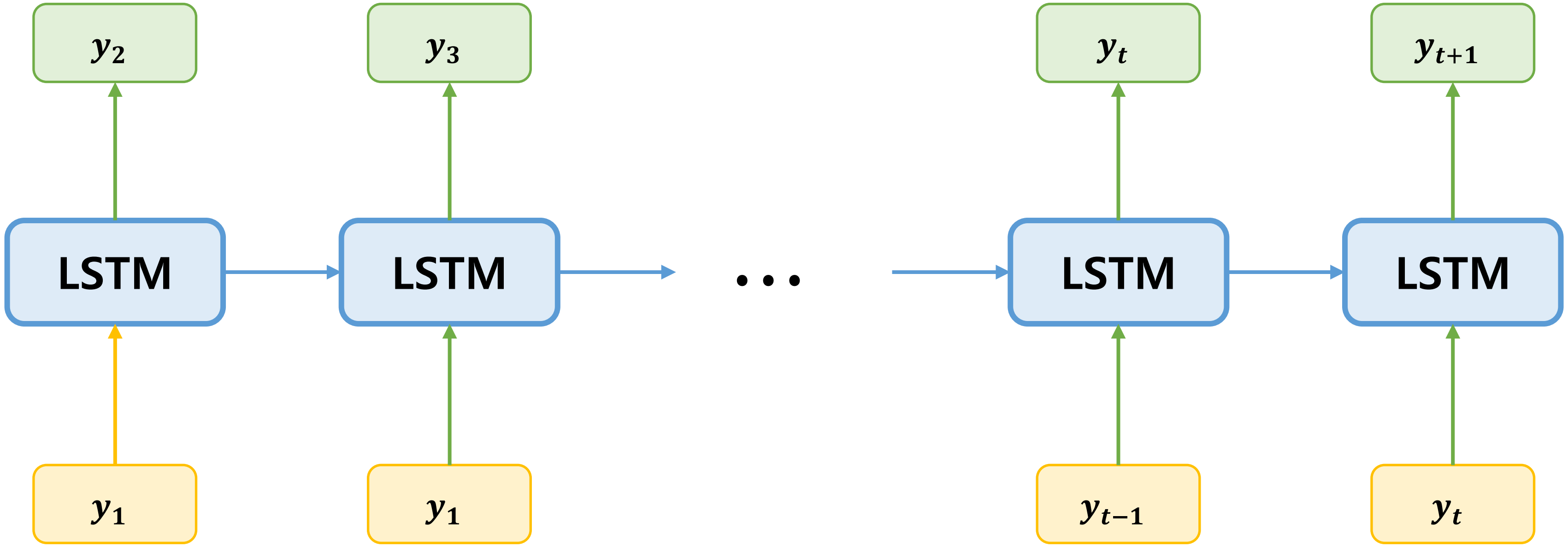
Pre-requisites

-Attention-Enhanced Caption Decoder



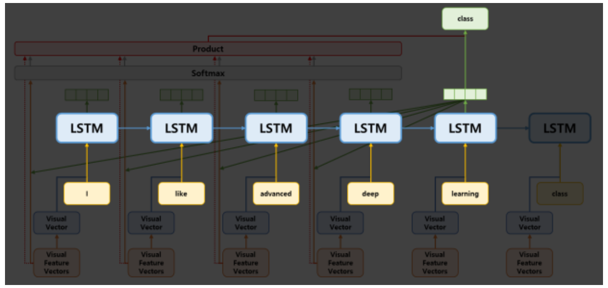
<Attention-Enhanced Caption Decoder>

$$\mathcal{L}_{CE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(\, y_t^* \mid y_{1:t-1}^* \,))$$



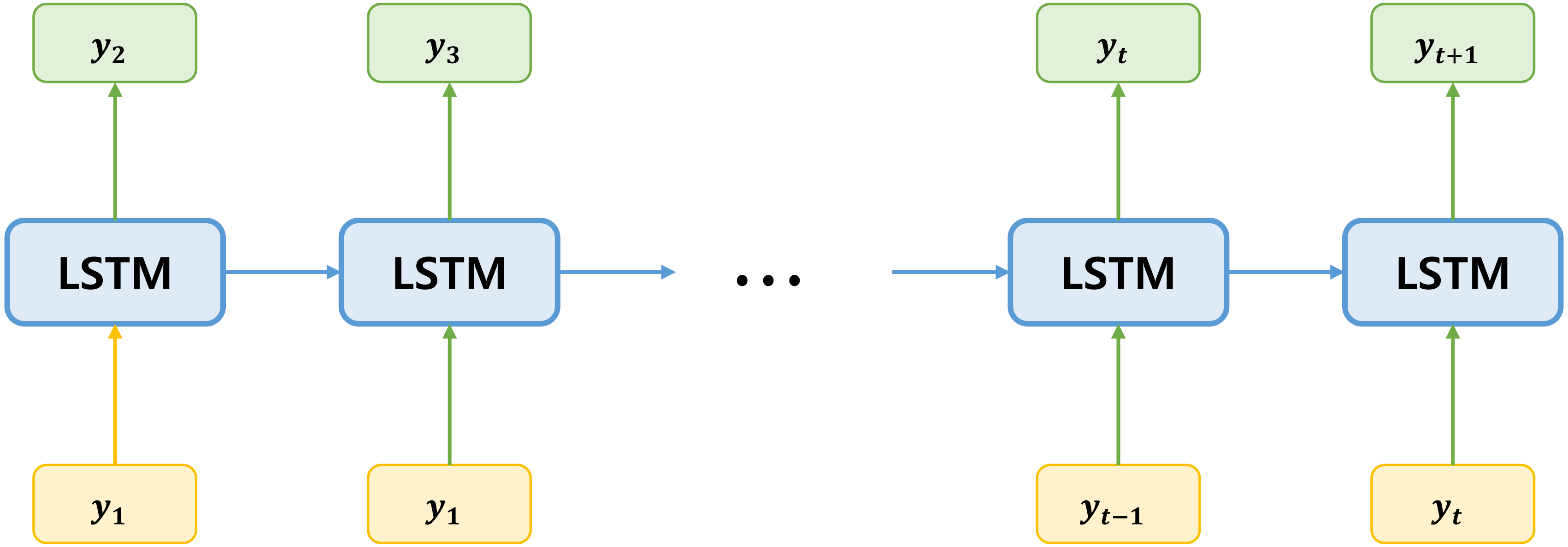
Pre-requisites

-Attention-Enhanced Caption Decoder



<Attention-Enhanced Caption Decoder>

$$\mathcal{L}_{CE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(\, y_t^* \mid y_{1:t-1}^* \,))$$



<Attention-Enhanced Encoder-Decoder Framework>

$$V = \{v_1, v_2, \cdots, v_N\} \in \mathbb{R}^{d \times N}$$

$$\bar{v} = \frac{1}{k}\sum_{i=1}^{k} v_i$$

$$h_t = \mathrm{LSTM}(h_{t-1}, [\, W_e\, y_{t-1};\ \bar{v}])$$

$$\alpha_t = f_{Att}(h_t, V) = \mathrm{softmax}\,(w_\alpha \tanh(W_h h_t \oplus W_V V))$$

$$c_t = V\alpha_t^T$$

$$y_t \sim p_t = softmax\left(W_p[h_t; c_t] + b_p\right)$$

$$\mathcal{L}_{CE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(\, y_t^* \mid y_{1:t-1}^* \,))$$

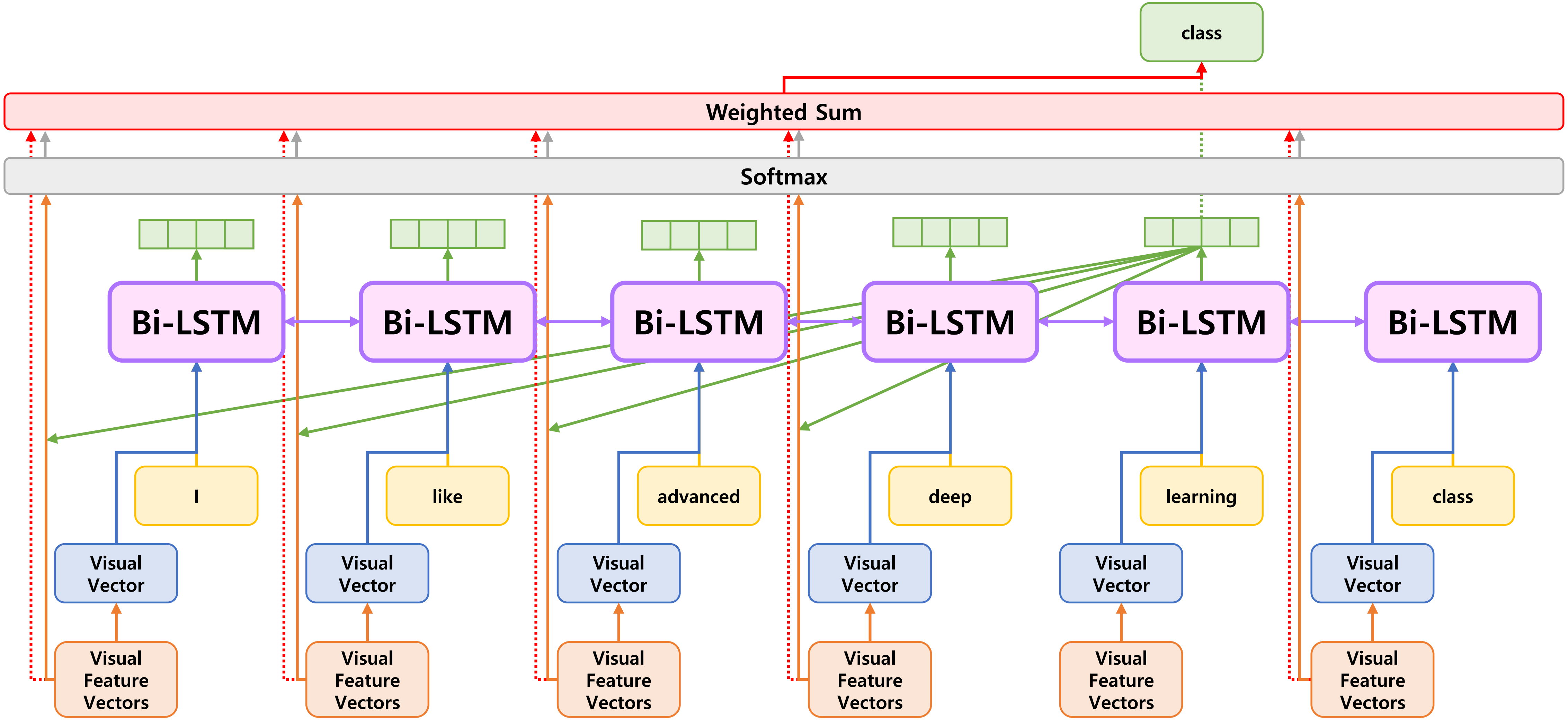
Model

- Prophet Attention
- Constant Prophet Attention
- Dynamic Prophet Attention

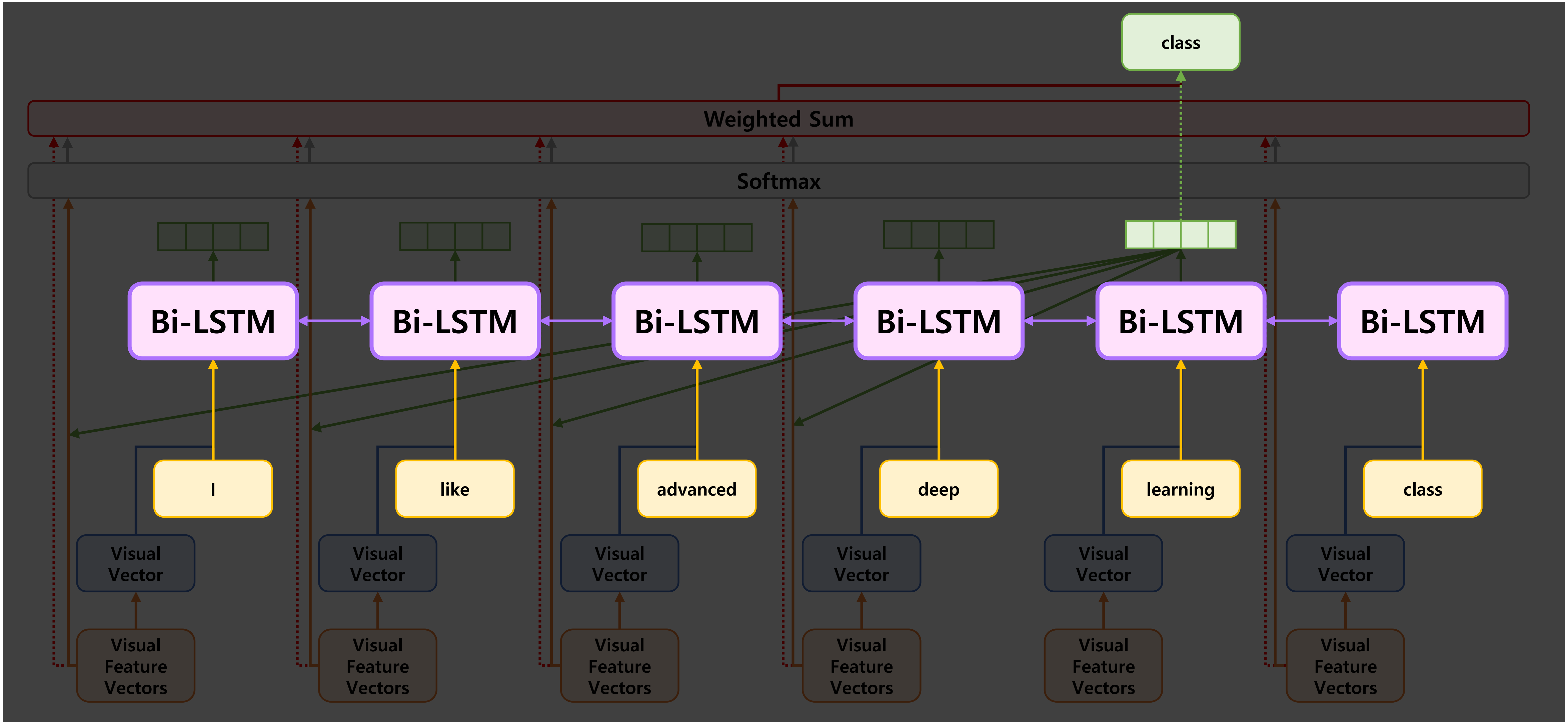
Pre-requisites

-Attention-Enhanced Encoder-Decoder Framework

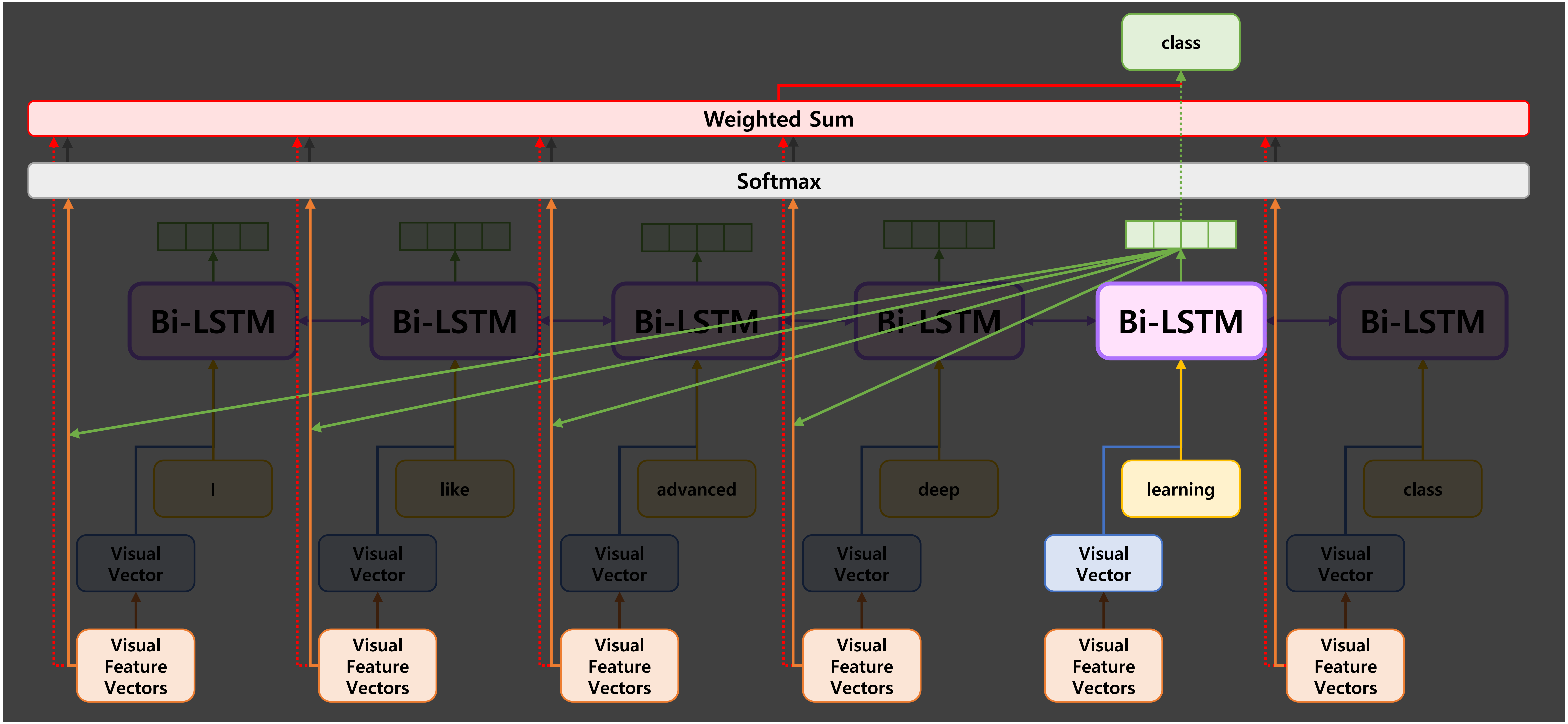
<Attention-Enhanced Encoder-Decoder Framework>



<Attention-Enhanced Encoder-Decoder Framework>



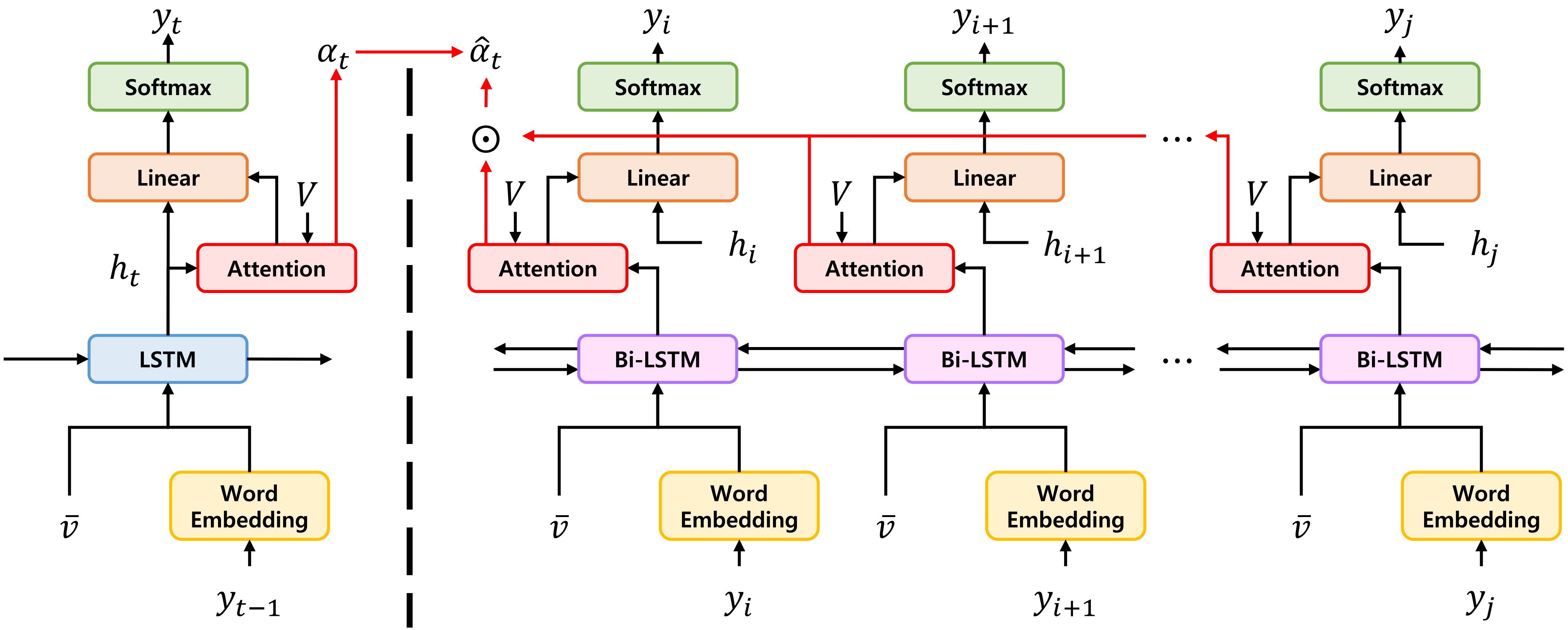
<Attention-Enhanced Encoder-Decoder Framework>



Model

- Prophet Attention

<Overall Architecture>



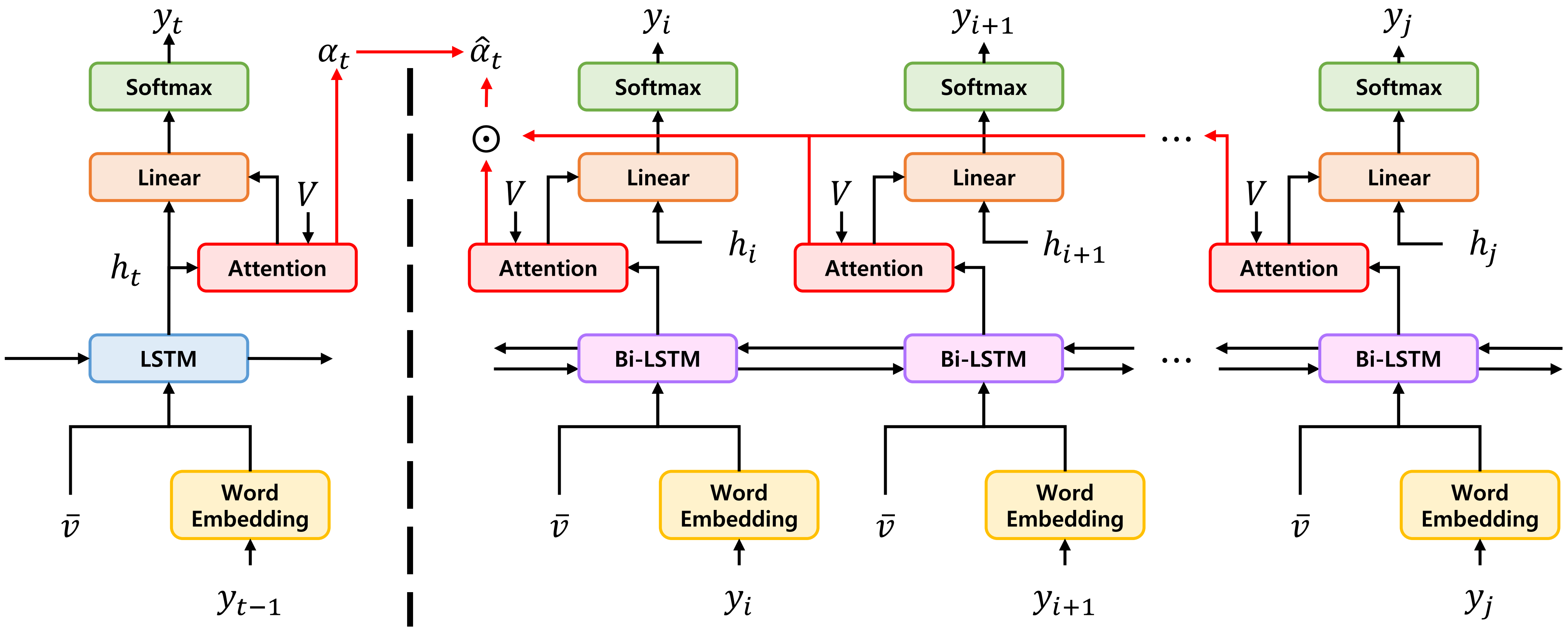
<Visual Attention> <Prophet Attention>

Model

- Prophet Attention

<Prophet Attention>

$$\hat{a}_t = f_{Prophet}\left(h'_{i:j}, V\right) = \frac{1}{j-i+1}\sum_{k=i}^{j} f_{Att}(h'_k, V)\,, where\ j \geq t$$



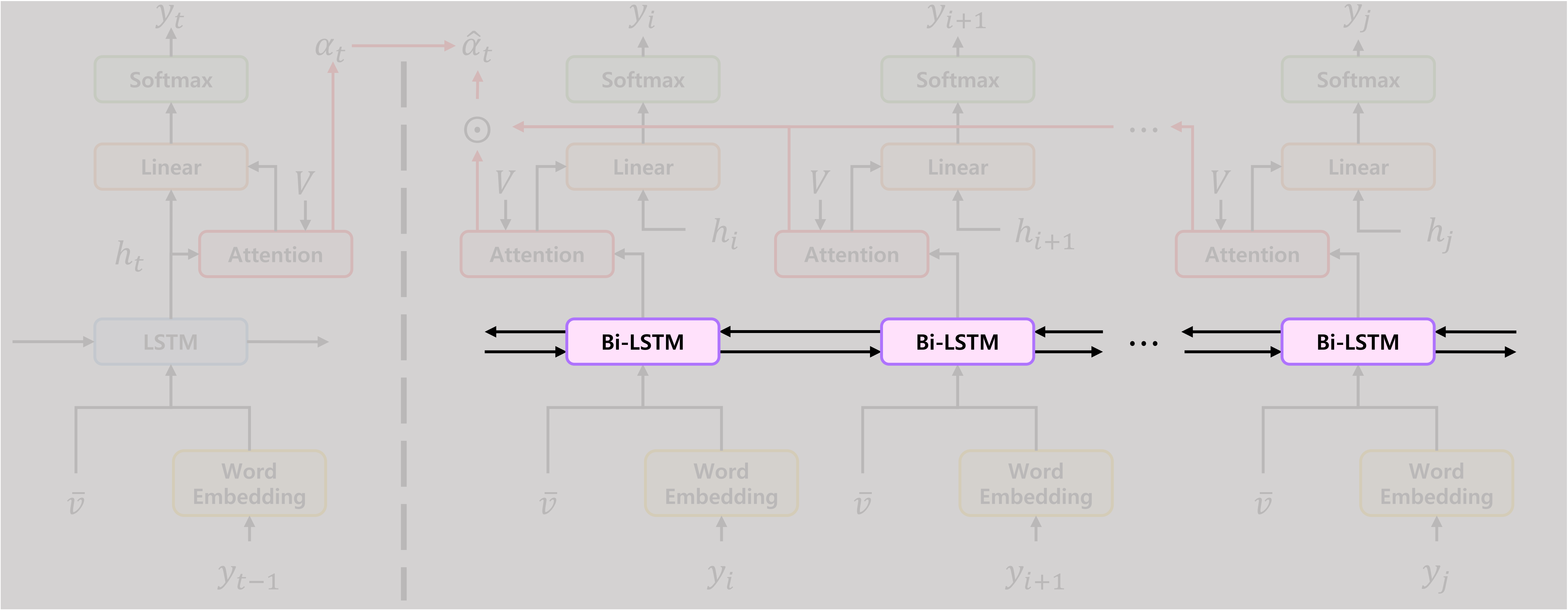
<Visual Attention> <Prophet Attention>

Model

- Prophet Attention

<Prophet Attention>

$$\hat{a}_t = f_{Prophet}\left(\boldsymbol{h}'_{i:j}, V\right) = \frac{1}{j-i+1}\sum_{k=i}^{j} f_{Att}(\boldsymbol{h}'_k, V), where\ j \geq t$$

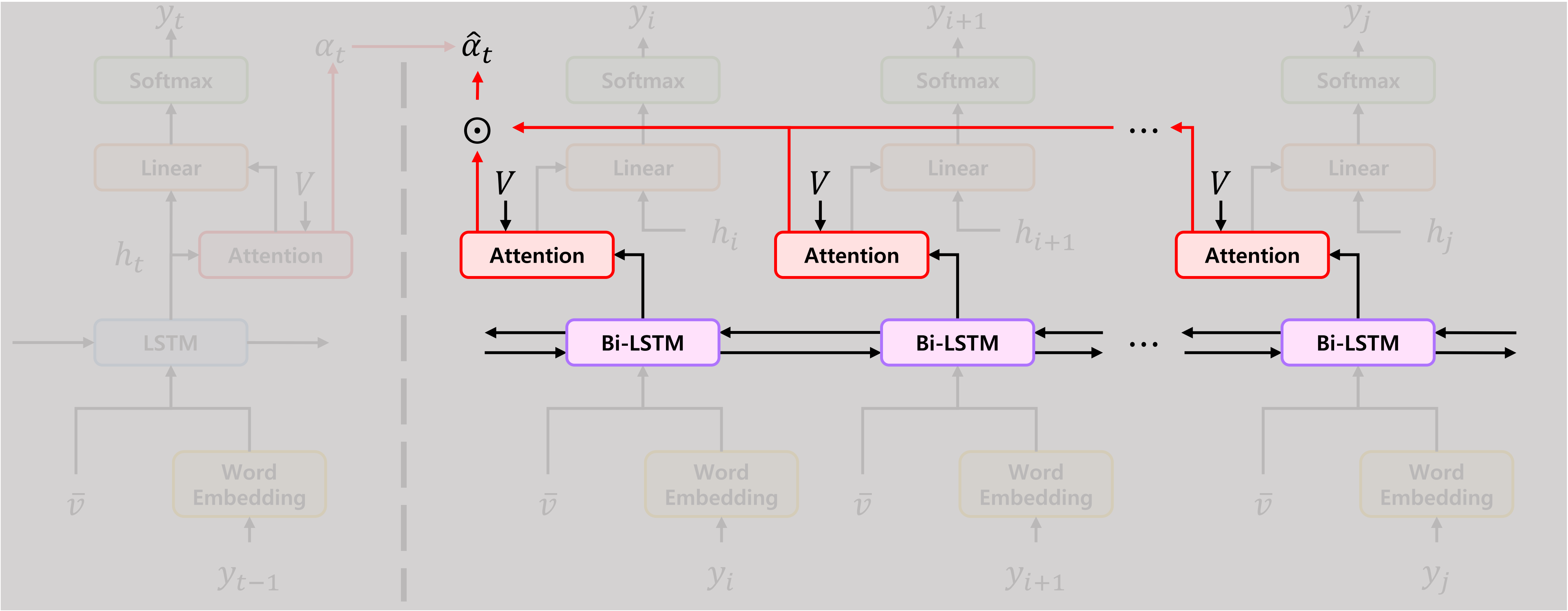


<Visual Attention> <Prophet Attention>

Model
- Prophet Attention

<Prophet Attention>

$$\hat{\alpha}_t = f_{Prophet}(h'_{i:j}, V) = \frac{1}{j-i+1}\sum_{k=i}^{j} f_{Att}(h'_k, V), where\ j \geq t$$



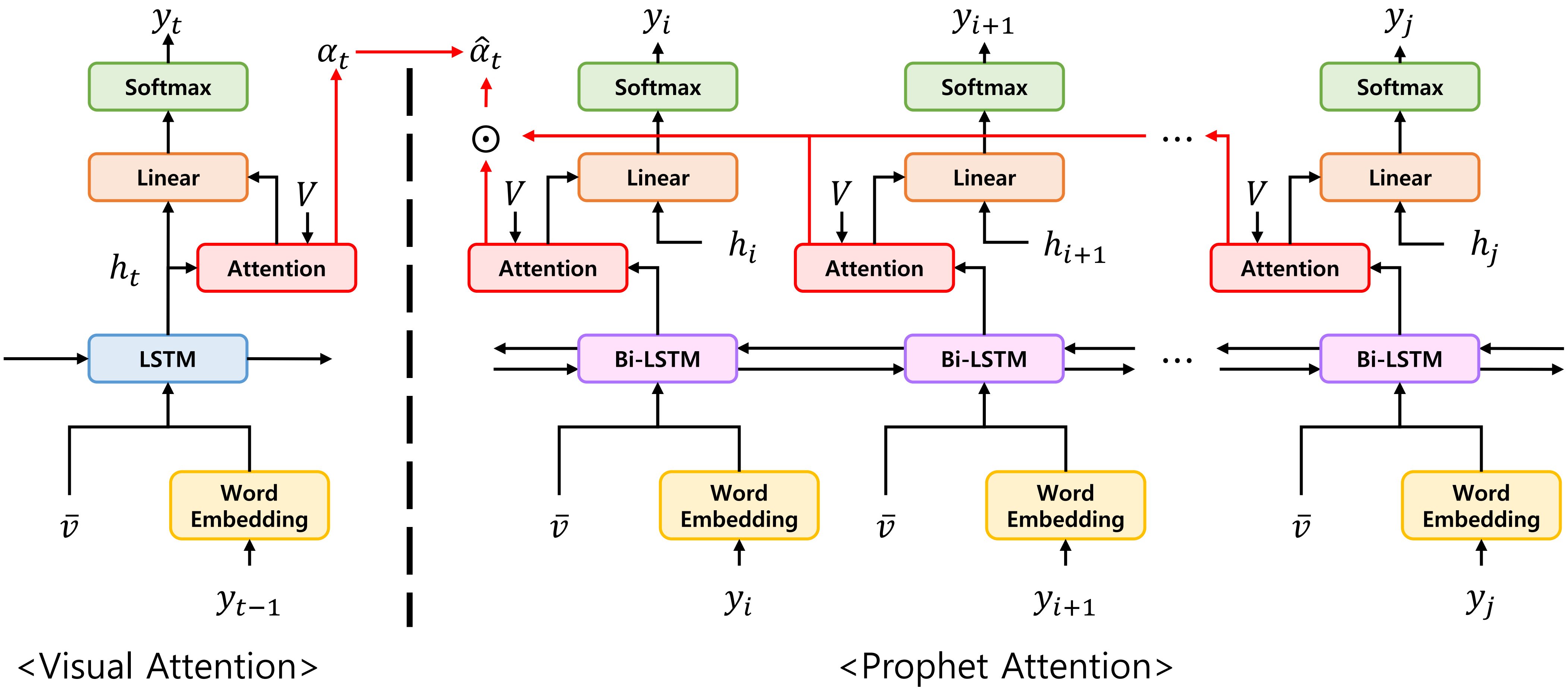
<Visual Attention> <Prophet Attention>

Model

- Prophet Attention

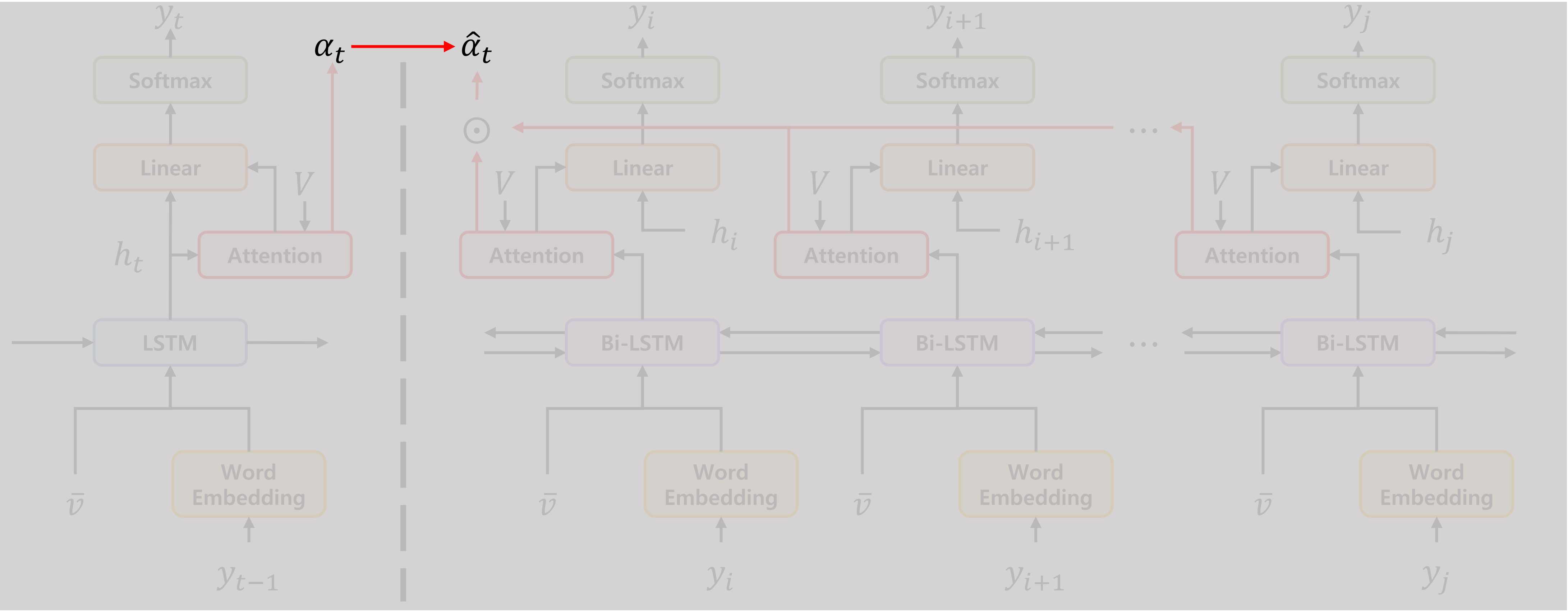
<Prophet Attention>

$$\mathcal{L}_{Att}(\theta) = \sum_{t=1}^{T} \|\alpha_t - \hat{\alpha}_t\|_1$$



<Prophet Attention>

$$\mathcal{L}_{Att}(\theta) = \sum_{t=1}^{T} \|\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t\|_1$$

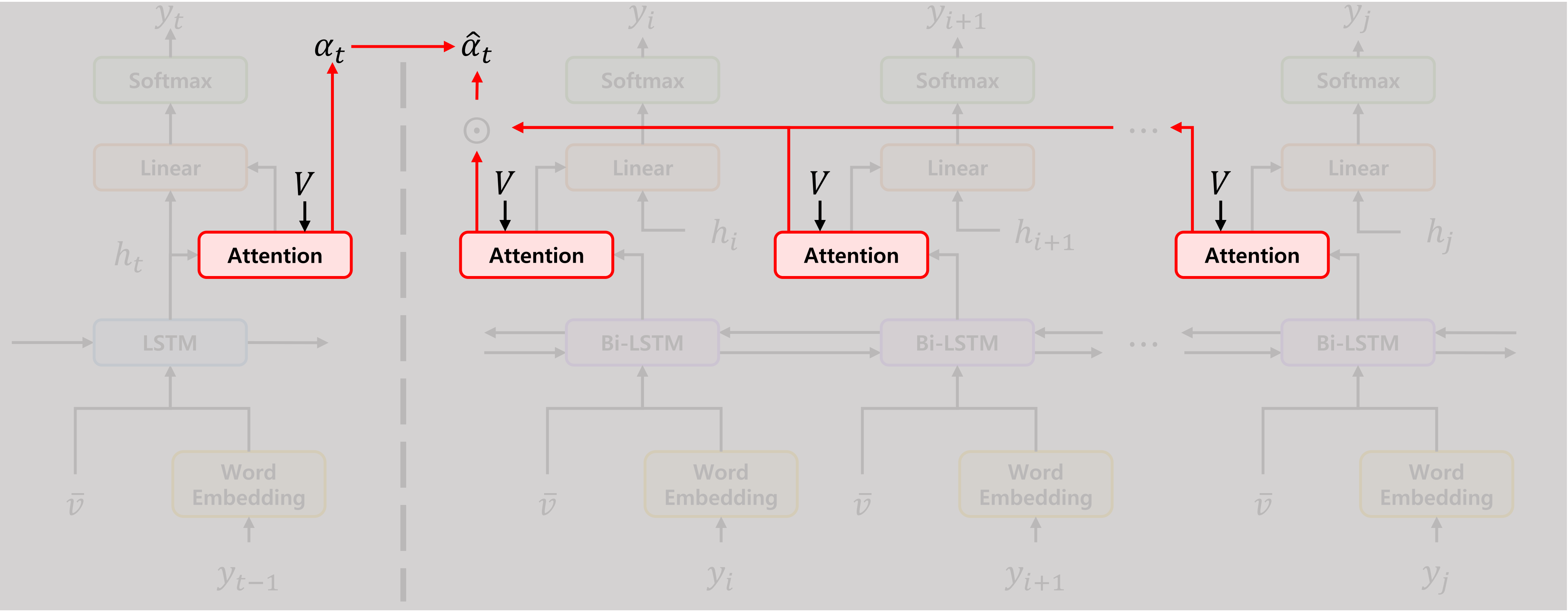


<Visual Attention> <Prophet Attention>

Model
- Prophet Attention

<Prophet Attention>

$$\mathcal{L}_{Att}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \|\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t\|_1$$



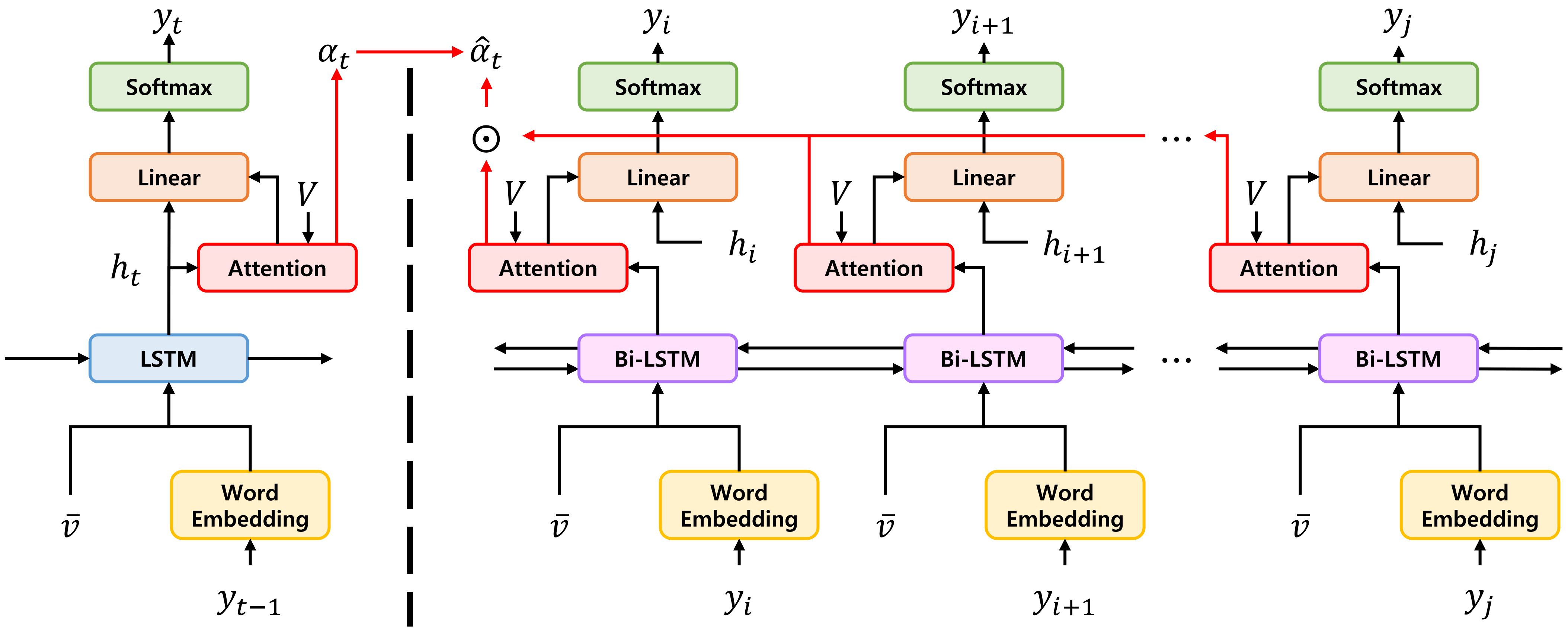
<Visual Attention> <Prophet Attention>

Model

- Prophet Attention

<Prophet Attention>

$$\hat{c} = V\hat{\alpha}_t^T, \qquad y_t \sim p_t = \mathrm{softmax}\left(W_p[h_t; \hat{c}_t] + b_p\right), \qquad \hat{\mathcal{L}}_{CE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(\, y_t^* \mid y_{1:t-1}^* \,))$$

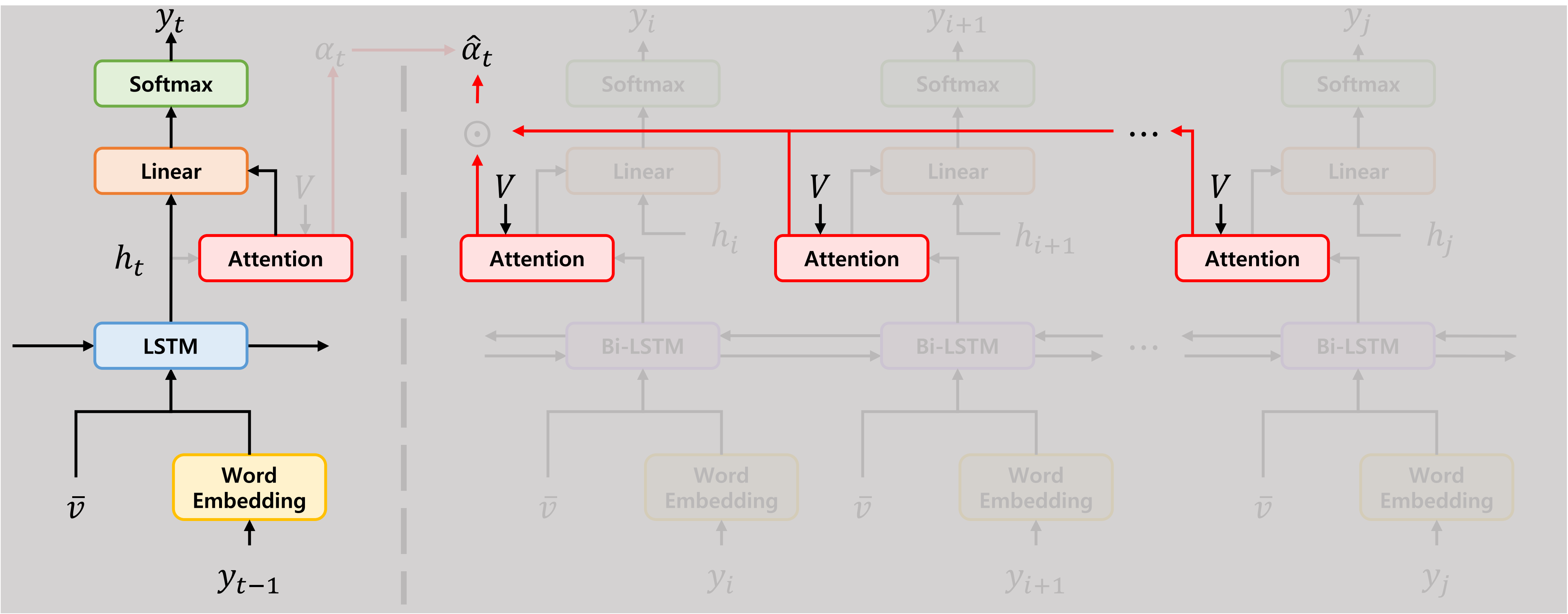


<Visual Attention>

<Prophet Attention>

<Prophet Attention>

$$\hat{c} = V\hat{\alpha}_t^T, \qquad y_t \sim p_t = \mathrm{softmax}\big(W_p[h_t; \hat{c}_t] + b_p\big), \qquad \hat{\mathcal{L}}_{CE}(\theta) = -\sum_{t=1}^{T} \log\big(p_\theta(\, y_t^* \mid y_{1:t-1}^* \,)\big)$$



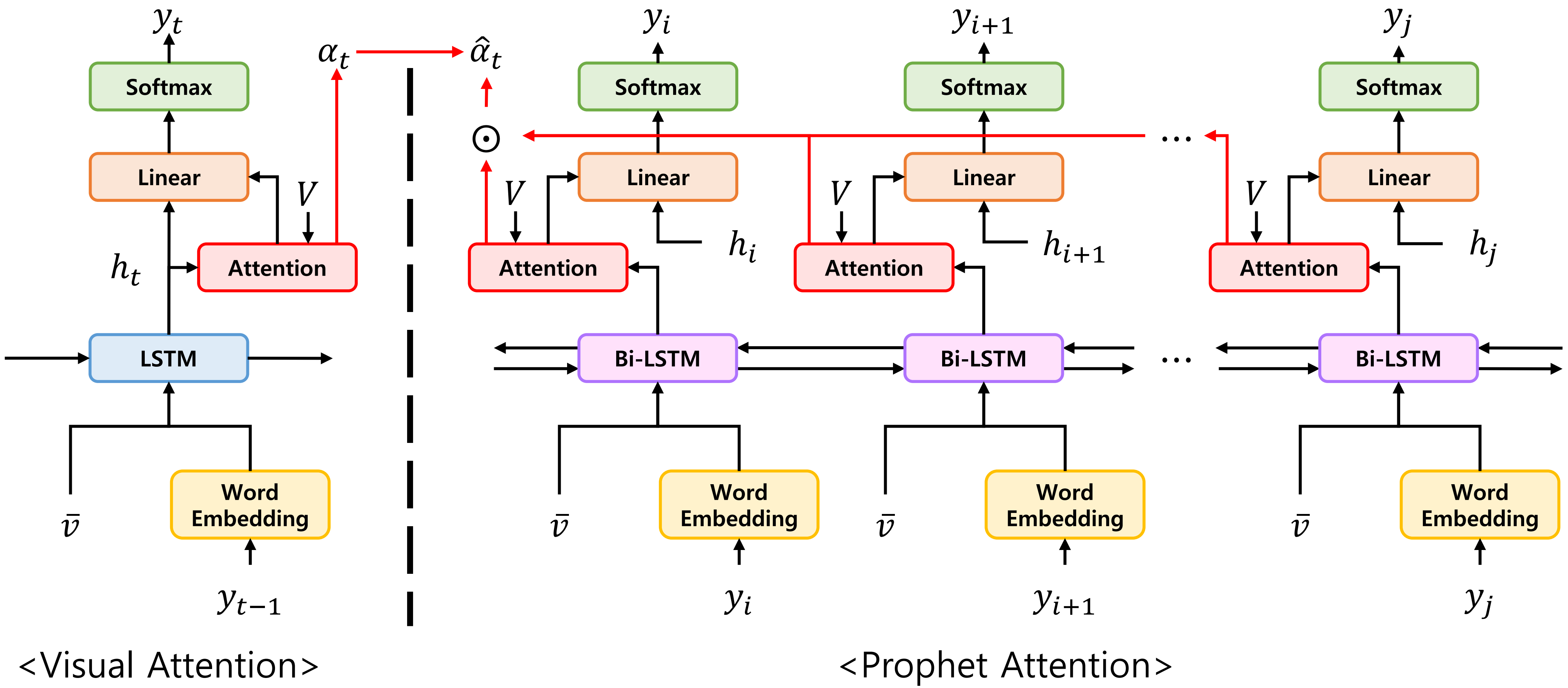
<Visual Attention> <Prophet Attention>

Model

- Prophet Attention

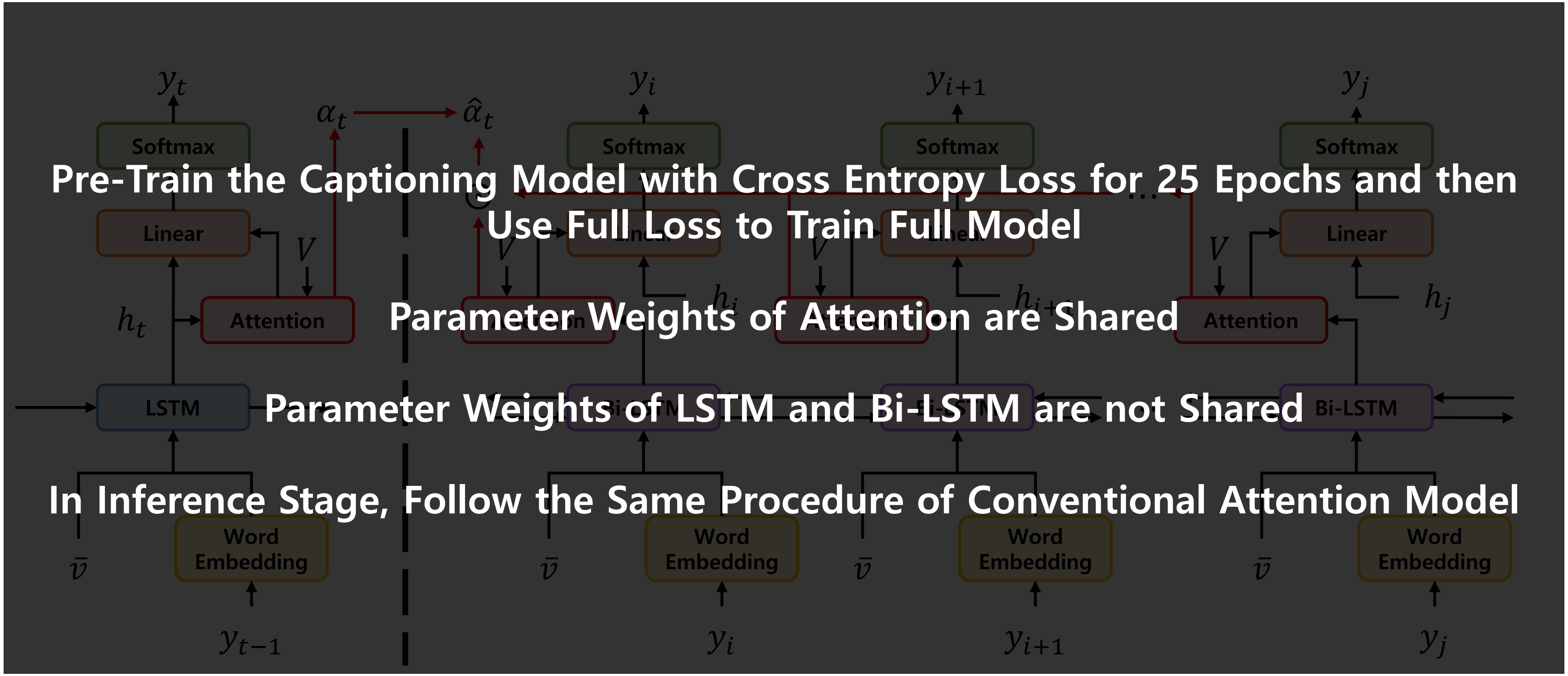
<Prophet Attention>

$$\mathcal{L}_{Full}(\theta) = \mathcal{L}_{CE}(\theta) + \hat{\mathcal{L}}_{CE}(\theta) + \lambda\mathcal{L}_{Att}(\theta)$$



<Prophet Attention>

$$\mathcal{L}_{Full}(\theta) = \mathcal{L}_{CE}(\theta) + \hat{\mathcal{L}}_{CE}(\theta) + \lambda\mathcal{L}_{Att}(\theta)$$



Pre-Train the Captioning Model with Cross Entropy Loss for 25 Epochs and then Use Full Loss to Train Full Model

Parameter Weights of Attention are Shared

Parameter Weights of LSTM and Bi-LSTM are not Shared

In Inference Stage, Follow the Same Procedure of Conventional Attention Model

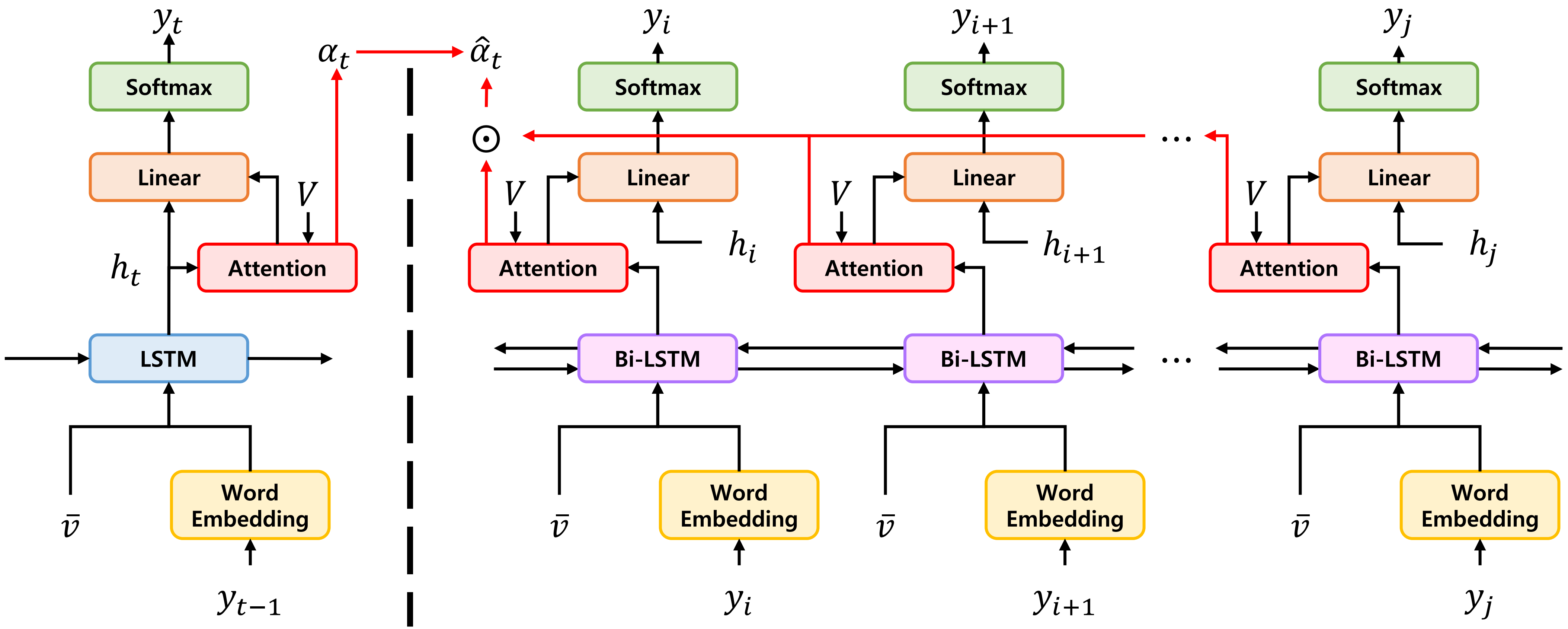
<Visual Attention> <Prophet Attention>

Model

- Constant Prophet Attention

<Constant Prophet Attention>

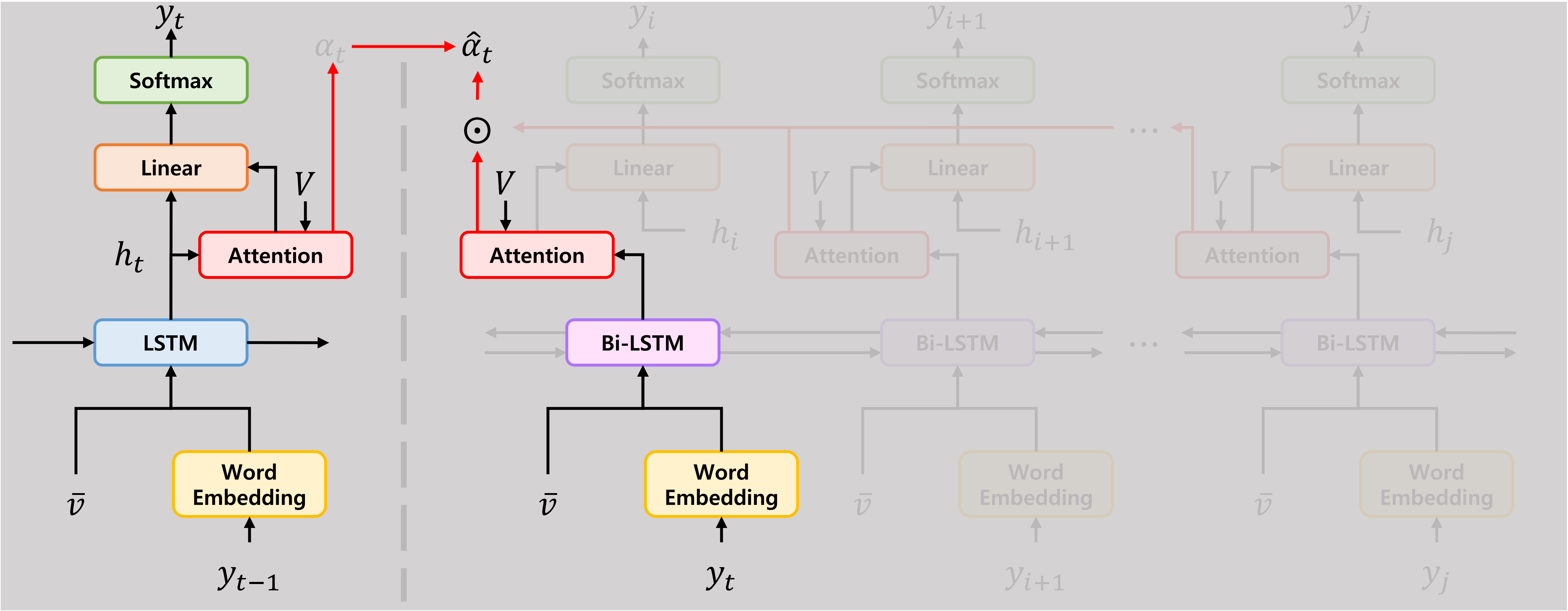
$$\hat{a}_t = f_{Prophet}\left(h'_{i:j}, V\right) = f_{Att}(h'_t, V), where\ i = j = t$$



<Visual Attention> <Prophet Attention>

<Constant Prophet Attention>

$$\hat{a}_t = f_{Prophet}\left(h'_{i:j}, V\right) = f_{Att}(h'_t, V), where\ i = j = t$$



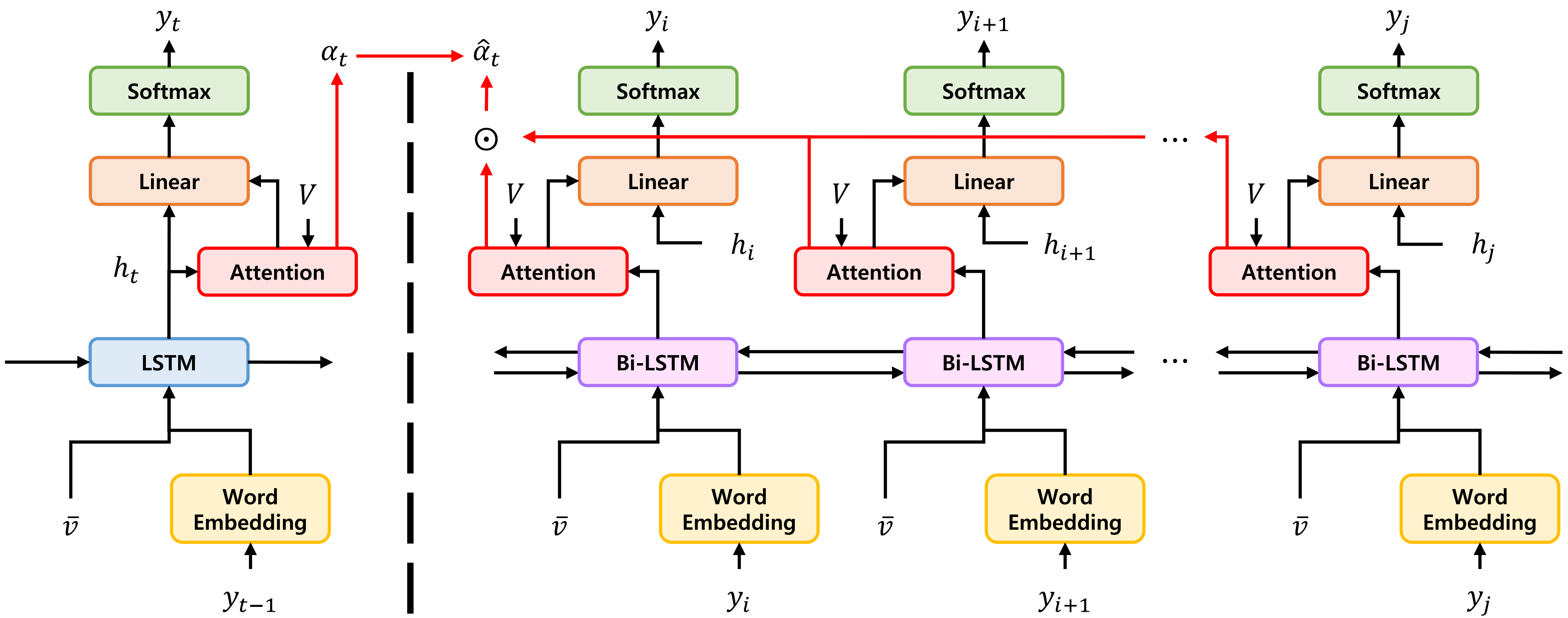
<Visual Attention>　　<Prophet Attention>

Model
- Dynamic Prophet Attention

<Dynamic Prophet Attention>

$$\hat{a}_t = f_{Prophet}(h'_{i:j}, V) = \begin{cases} \frac{1}{n-m}\sum_{k=m}^{n} f_{Att}(h'_k, V), & \text{if } y_t \in NP: y_{m:n} \\ MASK, & \text{if } y_t \in NV: \{y_{NV}\} \\ f_{Att}(h'_t, V), & \text{otherwise} \end{cases}$$

NP : Noun Phrase, NV : Non-Visual



<Visual Attention> <Prophet Attention>

Experiments

- Results

<Result>

Methods	$F1_{all}$	$F1_{loc}$	B-4	M	C	S
	Flickr30k Entities					
NBT [33]	-	-	27.1	21.7	57.5	15.6
Up-Down [2]	4.53	13.0	27.3	21.7	56.6	16.0
GVD [59]	3.88	11.7	26.9	22.1	60.1	16.1
Cyclical [35]‡	4.98	13.53	27.4	22.3	61.4	16.6
Up-Down*	4.19	12.1	26.4	21.5	57.0	15.6
w/ DPA	**5.45**†	**15.3**†	**27.2**†	**22.3**†	**60.8**†	**16.3**†
GVD*	3.97	11.8	26.6	22.1	59.9	16.3
w/ DPA	**4.79**†	**15.5**†	**27.6**†	**22.6**†	**62.7**†	**16.7**†

Methods	B-4	M	R-L	C	S
	MSCOCO				
Up-Down [2]	36.3	27.7	56.9	120.1	21.4
ORT [17]	38.6	28.7	58.4	128.3	22.6
AoANet [20]	38.9	29.2	58.8	129.8	22.4
X-Trans. [38]‡	39.7	29.5	59.1	132.8	23.4
Up-Down*	36.7	27.9	57.1	123.5	21.3
w/ DPA	**38.6**†	**29.1**†	**58.3**†	**129.0**†	**22.2**†
AoANet*	38.8	29.0	58.7	129.6	22.6
w/ DPA	**40.5**†	**29.6**†	**59.2**†	**133.4**†	**23.3**†

<Performance of Offline Evaluation on the Flickr30k Entities and the MSCOCO Image Captioning Datasets>

<MSCOCO Benchmark>

Methods	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down [2]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
GLIED [28]	80.1	94.6	64.7	88.9	50.2	80.4	38.5	70.3	28.6	37.9	58.3	73.8	123.3	125.6
SGAE [54]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
GCN-LSTM [55]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
AoANet [20]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
$\mathcal{M}^2$ Trans. [10]‡	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Trans. [38]‡	**81.9**	95.7	**66.9**	90.5	**52.4**	82.5	**40.3**	72.4	**29.6**	39.2	**59.5**	75.0	**131.1**	133.5
Ours	81.8	**96.3**	66.5	**91.2**	51.9	**83.2**	39.8	**73.3**	**29.6**	**39.3**	59.4	**75.1**	130.4	**133.7**

<Highest Ranking Published Image Captioning Results on the Online MSCOCO Test Server>

<Grounding Performance>

Datasets	vs. Models	Baseline wins (%)	Tie (%)	w/ DPA wins (%)
Flickr30k Entities	Up-Down	19.6	46.8	**33.6**
	GVD	23.6	44.4	**32.0**
MSCOCO	Up-Down	22.0	40.4	**37.6**
	AoANet	26.4	38.8	**34.8**

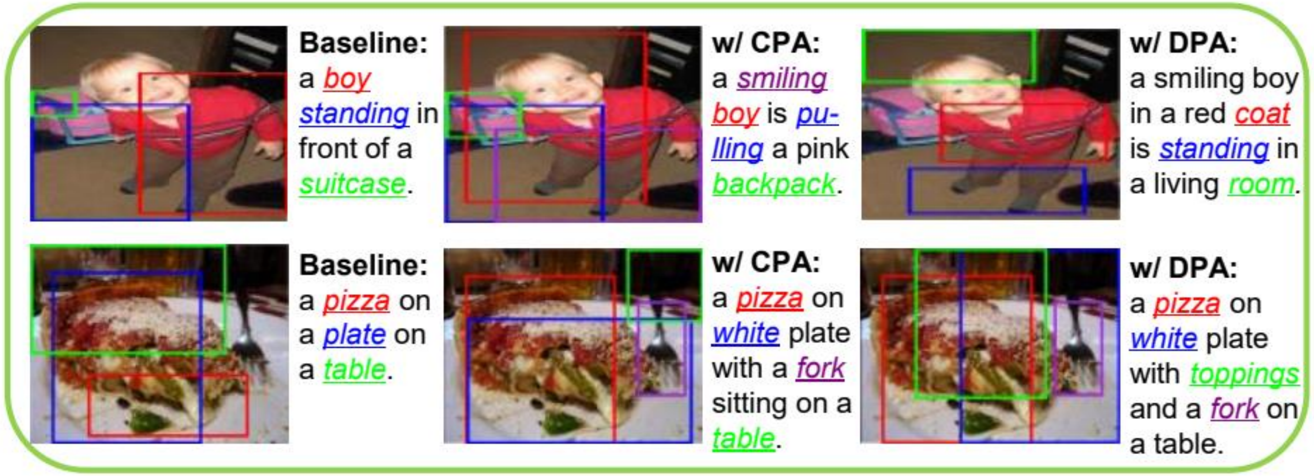
<Grounding Performance of Human Evaluation>

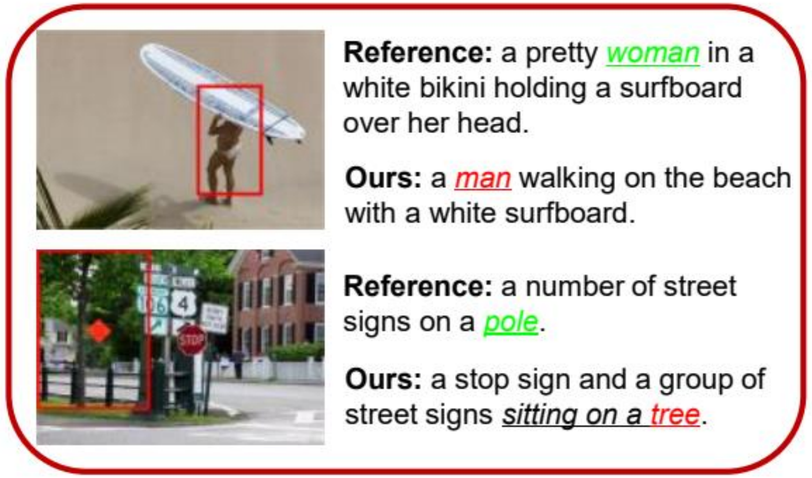
Experiments

- Results

<Grounding Performance>

Categories	"w/ CPA" wins (%)	Tie (%)	"w/ DPA" wins (%)
Object	25.8	44.6	**29.6**
Relationship	25.0	46.6	**28.4**
Attribute	21.2	43.0	**35.8**





<Results of Human Evaluation on the MSCOCO Dataset in terms of Object>

<Application in Other Tasks>

Methods	Paraphrase		Video Captioning
	BLEU	METEOR	CIDEr
Baseline	29.2	23.5	48.9
w/ DPA	**36.5 (+7.3)**	**26.8 (+3.3)**	**52.2 (+3.3)**

<Results of Paraphrase and Video Captioning Task>

Conclusion

<Conclusion>

- Proposed Prophet Attention to enable attention models to correctly ground words that are to be generated to proper image regions.

- Evaluated Prophet Attention for image captioning on the Flickr30k Entities and the MSCOCO datasets and Achieved the 1st place on the leaderboard.

- Attempted to adapt Prophet Attention to other language generation task and obtained positive experimental results on paraphrase generation and video captioning tasks.

Any Questions?

Thank You