Paper Review

Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision

Tan and Bansal, 2020, EMNLP

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

- Human Language Acquisition

<What This Seminar Does Not Cover>

- **Details of BERT**

  Devlin et al., 2019, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, NAACL

Introduction

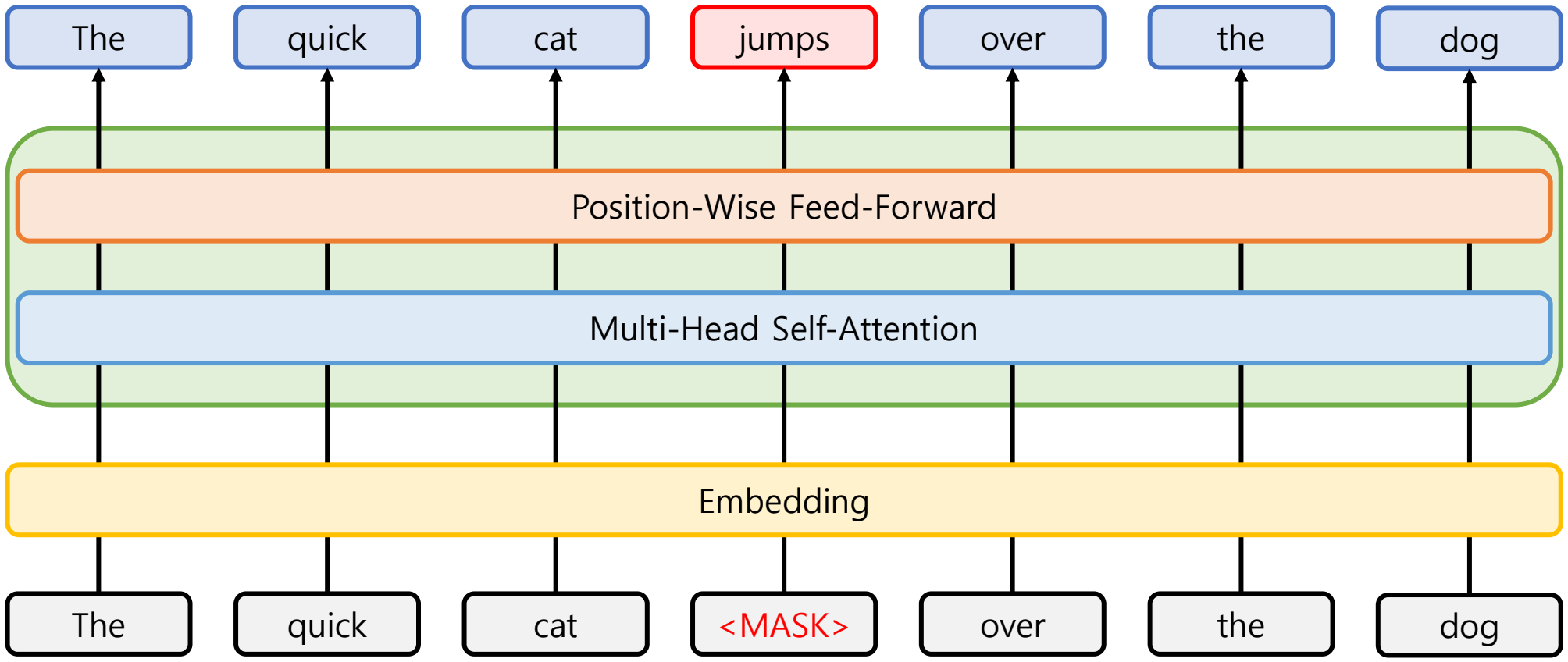
-Recap: Previous Seminar

<Recap: Previous Seminar>

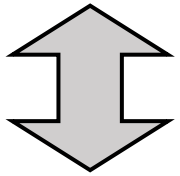
Human language acquisition

사람은 결국 언어의 형태(표현) 뿐만 아니라,
다양한 요소의 상호작용을 통해 언어를 습득함

<Model Language Acquisition>

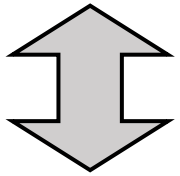


<Human Language Acquisition>



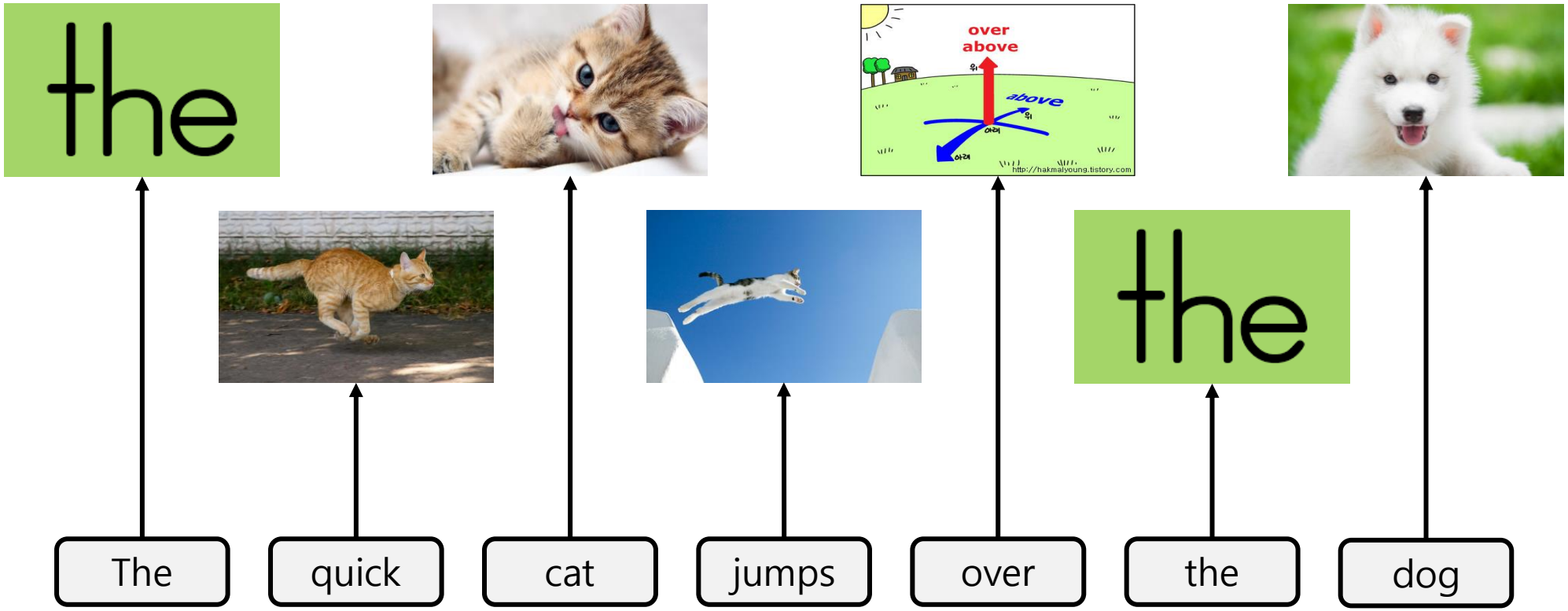
“The quick cat jumps over the dog.”

<Human Language Acquisition>

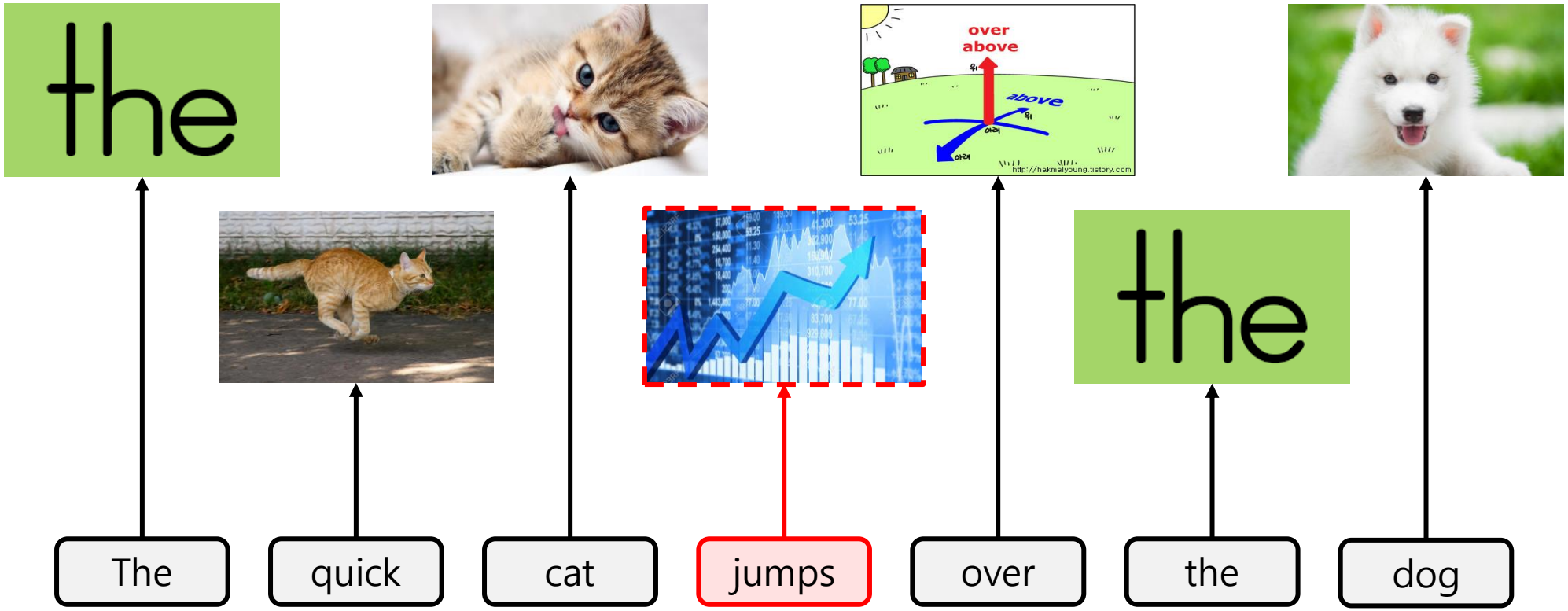


“The quick cat jumps over the dog.”

<Human Language Acquisition>

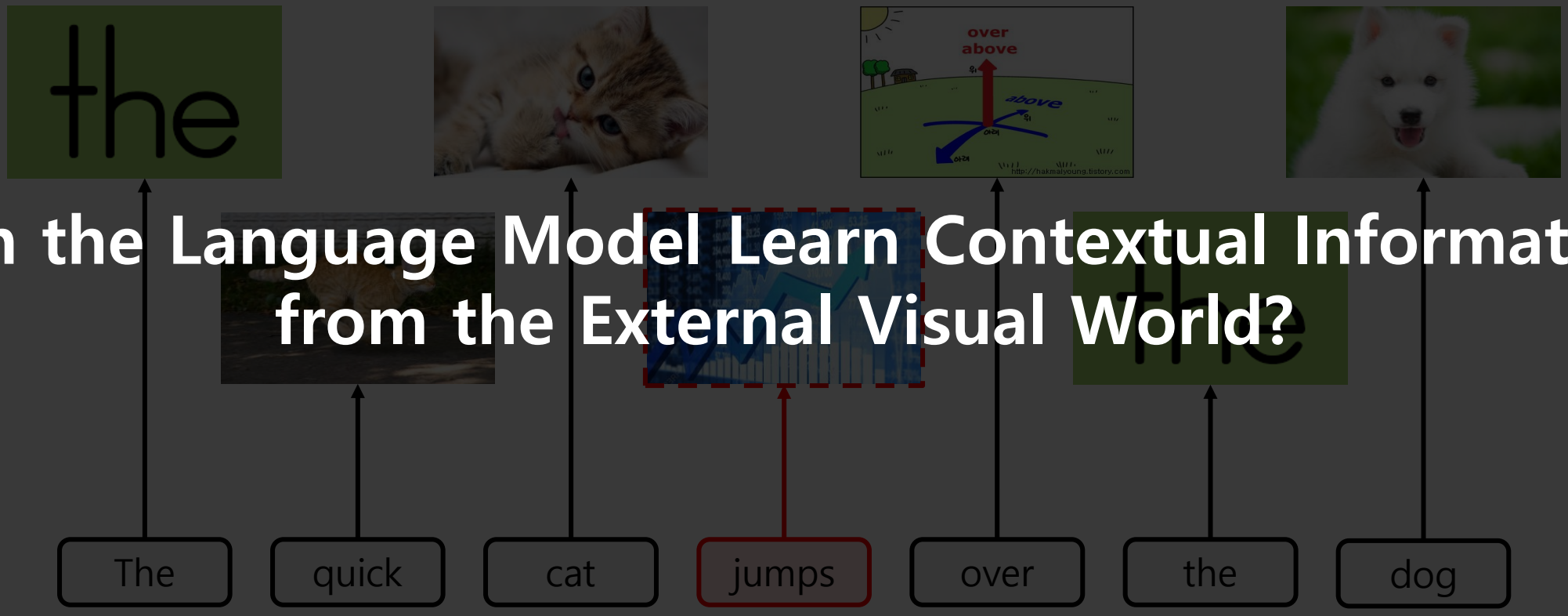


<Human Language Acquisition>



Introduction
-Human Language Acquisition

<Human Language Acquisition>



Can the Language Model Learn Contextual Information from the External Visual World?

Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision

Tan and Bansal, 2020, EMNLP

Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision

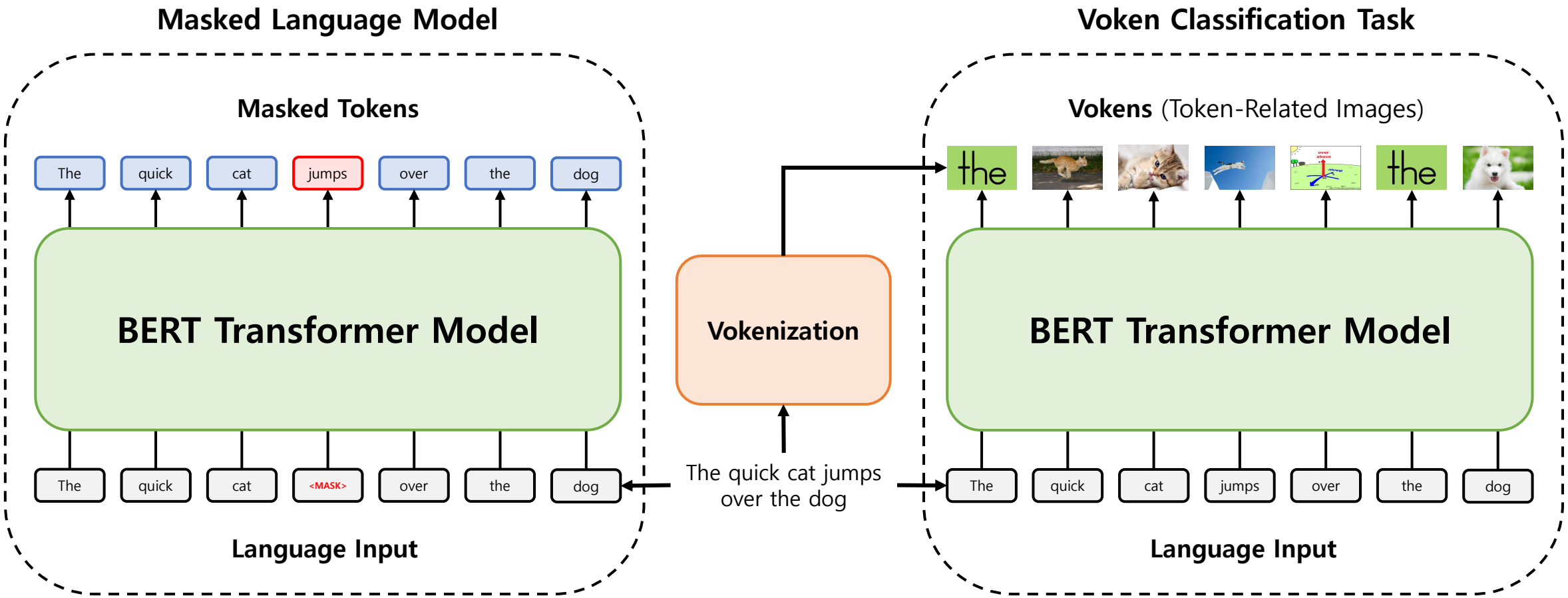
Tan and Bansal, 2020, EMNLP

Model

- The Voken-Classification Task
- Visually-Supervised Language Model
- Two Challenges in Creating Vokens

Model
-Overview

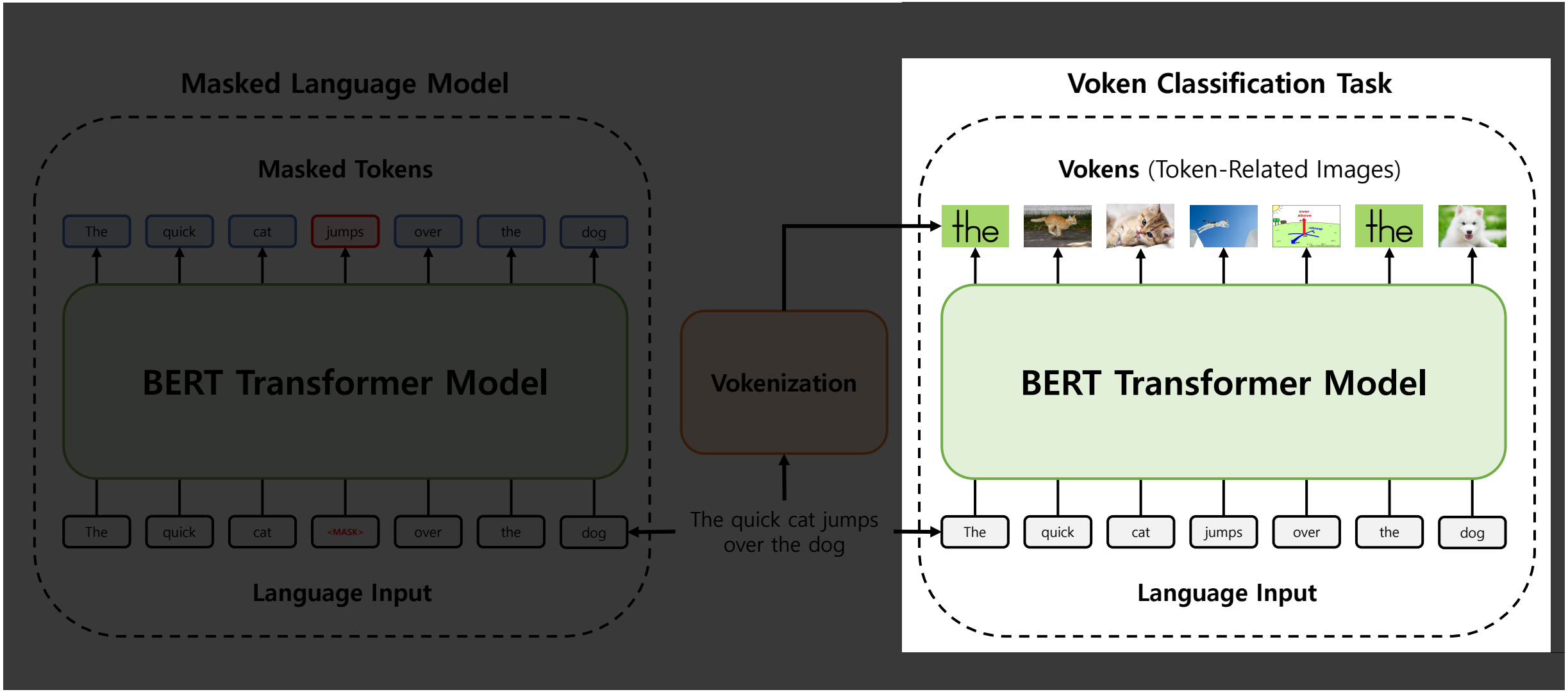
<Visually-Supervised Language Model>



Model

-The Voken Classification Task

<The Voken Classification Task>



<The Voken Classification Task>

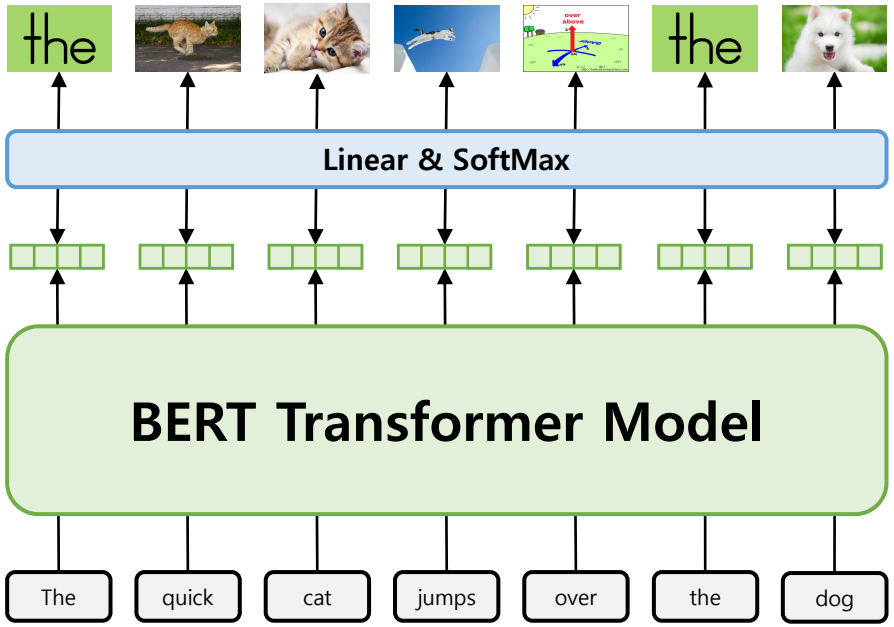
$$\{h_i\} : Language\ Model\ Representations$$

$$s = \{w_i\}, where\ S: Sentence, w_i : Tokens$$

$$h_1, h_2, \cdots, h_l = LM(w_1, w_2, \cdots, w_l)$$

$$p_i(v|s) = softmax_v\{Wh_i + b\}$$

$$\mathcal{L}_{VOKEN-CLS}(s) = -\sum_{i=1}^{l} \log p_i(v(w_i\ ; s) \mid s)$$



<The Voken Classification Task>

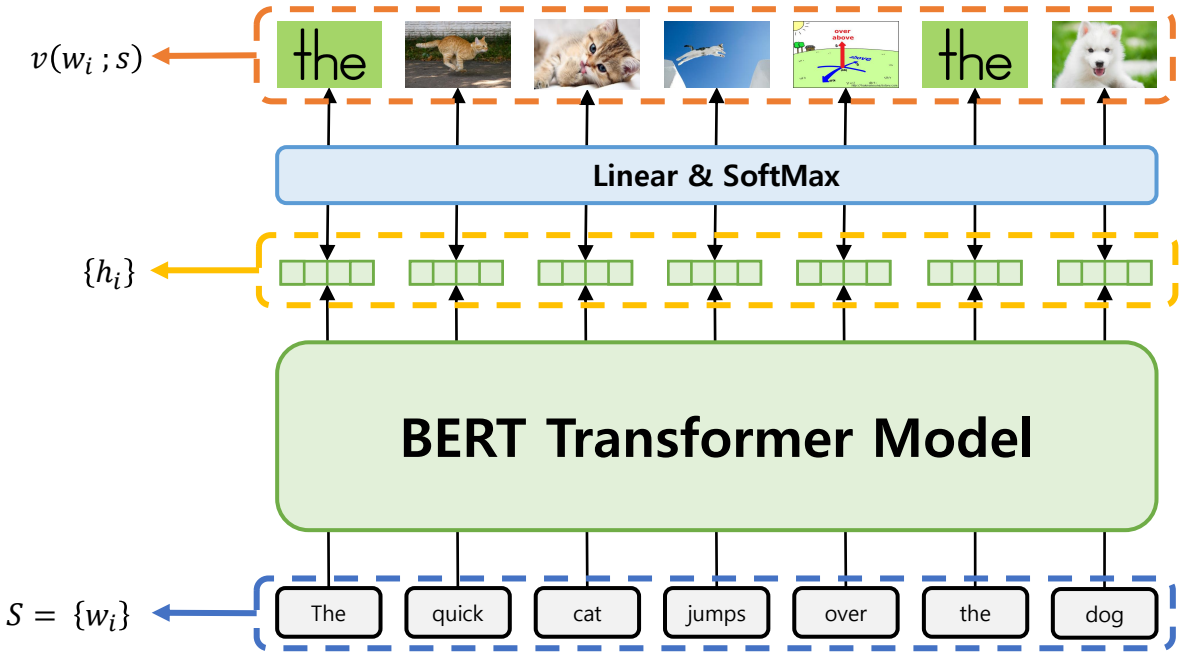
$\{h_i\}$: *Language Model Representations*

$s = \{w_i\}$, *where* S: *Sentence*, w_i : *Tokens*

$$h_1, h_2, \cdots, h_l = LM(w_1, w_2, \cdots, w_l)$$

$$p_i(v|s) = softmax_v\{Wh_i + b\}$$

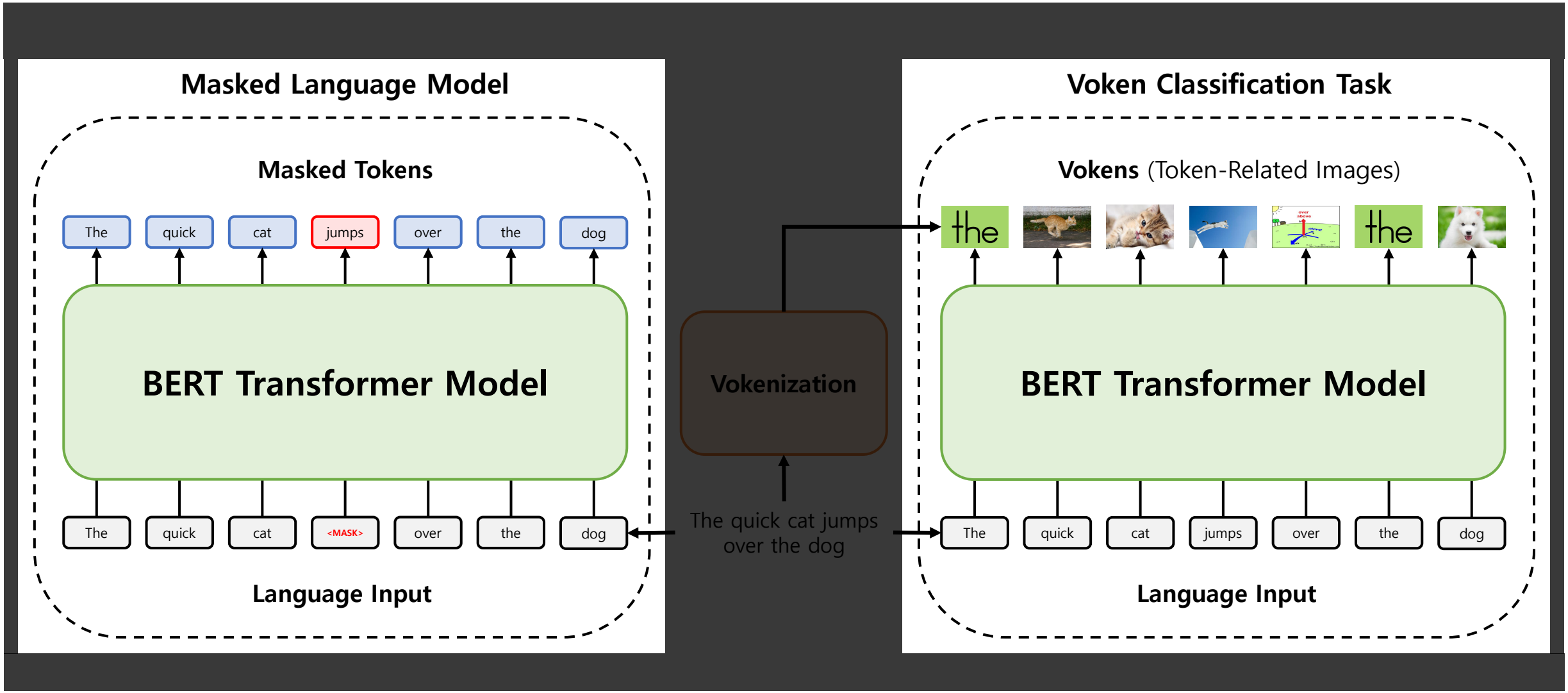
$$\mathcal{L}_{VOKEN\text{-}CLS}(s) = -\sum_{i=1}^{l} \log p_i(v(w_i\ ; s)\ |\ s)$$



Model

-Visually-Supervised Language Model

<Visually-Supervised Language Model>



<Visually-Supervised Language Model>

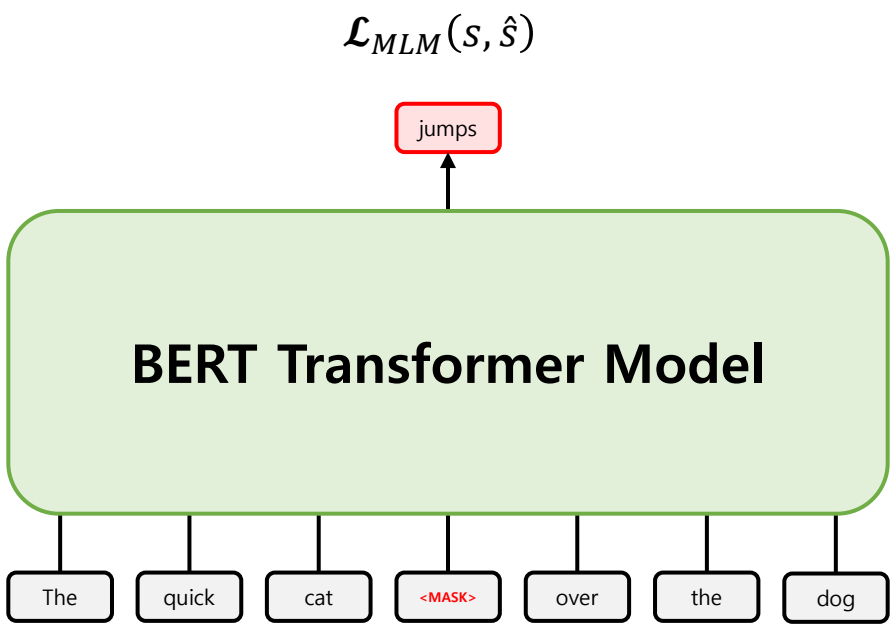
s , $\hat{s}$: $Set\ of\ Tokens, Masked\ Tokens$

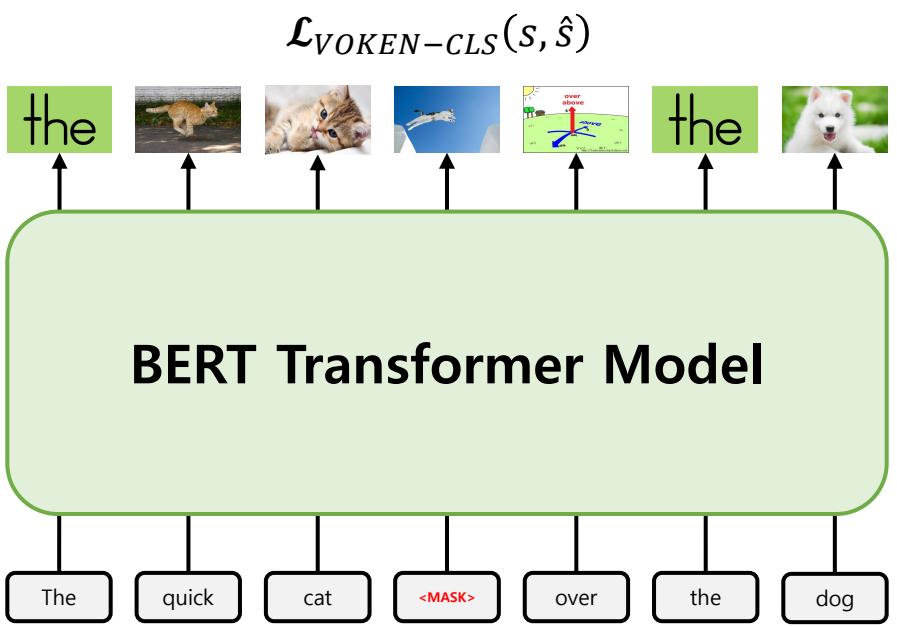
$s \setminus \hat{s}$: $Unmasked\ Tokens$

$$\mathcal{L}_{MLM}(s, \hat{s}) = -\sum_{w_i \in \hat{s}} \log q_i\ (w_i \mid s \setminus \hat{s})$$

$$\mathcal{L}_{VOKEN-CLS}(s, \hat{s}) = -\sum_{w_i \in s} \log p_i\ (v(w_i\ ; s) \mid s \setminus \hat{s})$$

$$\mathcal{L}_{VLM}(s, \hat{s}) = \mathcal{L}_{VOKEN-CLS}(s, \hat{s}) + \lambda \mathcal{L}_{MLM}(s, \hat{s}) \quad (1)$$





<Two Challenges in Creating Vokens>

It is Difficult to Apply **Image Captioning** to Large Corpus for the Following Two Reasons:

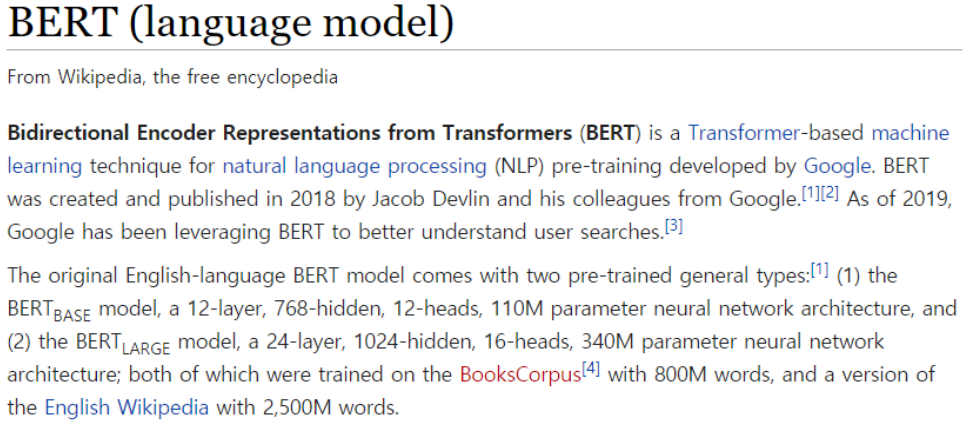
(1) Different Distributions between Grounded Language and Other Natural Language Corpora

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.

Average Sentence Length = 11.8 words

<MS COCO 2015 Image Captioning Task>



BERT (language model)

From Wikipedia, the free encyclopedia

Bidirectional Encoder Representations from Transformers (**BERT**) is a Transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google.[1][2] As of 2019, Google has been leveraging BERT to better understand user searches.[3]

The original English-language BERT model comes with two pre-trained general types:[1] (1) the $BERT_{BASE}$ model, a 12-layer, 768-hidden, 12-heads, 110M parameter neural network architecture, and (2) the $BERT_{LARGE}$ model, a 24-layer, 1024-hidden, 16-heads, 340M parameter neural network architecture; both of which were trained on the BooksCorpus[4] with 800M words, and a version of the English Wikipedia with 2,500M words.

Average Sentence Length = 24.1 words

<Wikipedia Database>

<Two Challenges in Creating Vokens>

It is Difficult to Apply **Image Captioning** to Large Corpus for the Following Two Reasons:

(2) Low Grounding Ratio in Natural Language

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.

Num of Sentences : 410K
Num of Images : 82K

<MS COCO 2015 Image Captioning Task>

BERT (language model)

From Wikipedia, the free encyclopedia

Bidirectional Encoder Representations from Transformers (**BERT**) is a Transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google.[1][2] As of 2019, Google has been leveraging BERT to better understand user searches.[3]

The original English-language BERT model comes with two pre-trained general types:[1] (1) the $BERT_{BASE}$ model, a 12-layer, 768-hidden, 12-heads, 110M parameter neural network architecture, and (2) the $BERT_{LARGE}$ model, a 24-layer, 1024-hidden, 16-heads, 340M parameter neural network architecture; both of which were trained on the BooksCorpus[4] with 800M words, and a version of the English Wikipedia with 2,500M words.

Num of Articles : 5.5M
Image Grounding Ratio : 27.7%

<Wikipedia Database>

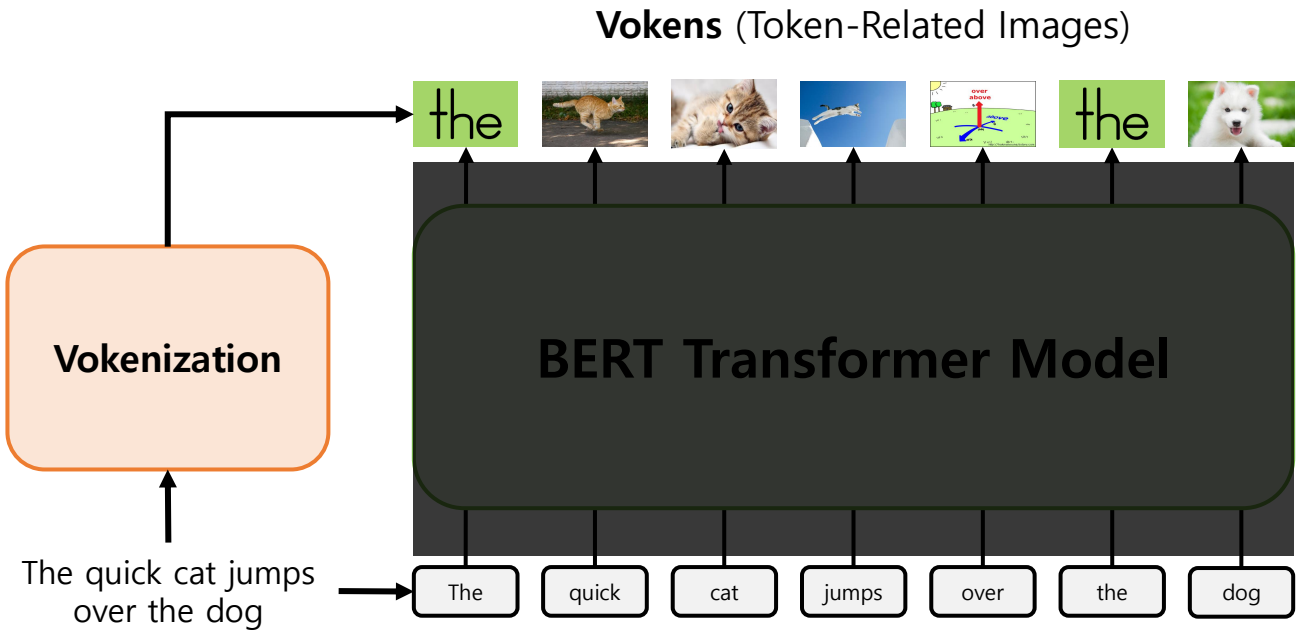
Vokenization

- The Vokenization Process
- Contextual Token-Image Matching Model
- Revokenization

<Vokenization>

Training **"Vokenizer"** from **Image-Captioning Dataset** and

Using it to **Annotate Large Language Corpora**



<The Vokenization Process>

$w_i : Token, s = (w_1, \cdots, w_l) : Sentence$

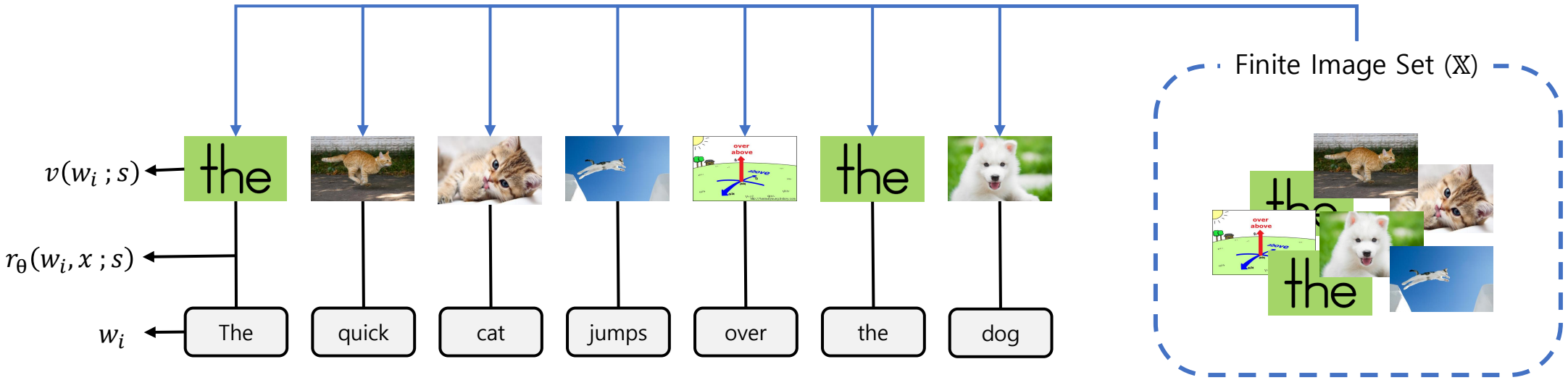
$v(w_i\ ; s) : Image\ Relevant\ to\ Token\ (Voken)$

$\mathbb{X} = (x_1, \cdots, x_n) : Finte\ Imate\ Set$

$r_\theta(w_i, x\ ; s) : Token\text{–}Image\ Relevance\ Scoring\ function$

$\theta^* : Optimal\ Parameter$

$$v(w_i\ ; s) = argmax_{x \in \mathbb{X}}\, r_{\theta^*}(w_i, x\ ; s)$$



Vokenization

-The Vokenization Process

Xie et al., Aggregated Residual Transformations for Deep Neural Network, CVPR, 2017

<Modeling>

$$r_\theta(w_i, x; s) = f_\theta(w_i\,; s)^T g_\theta(x) : Inner\ Product\ Factorization$$

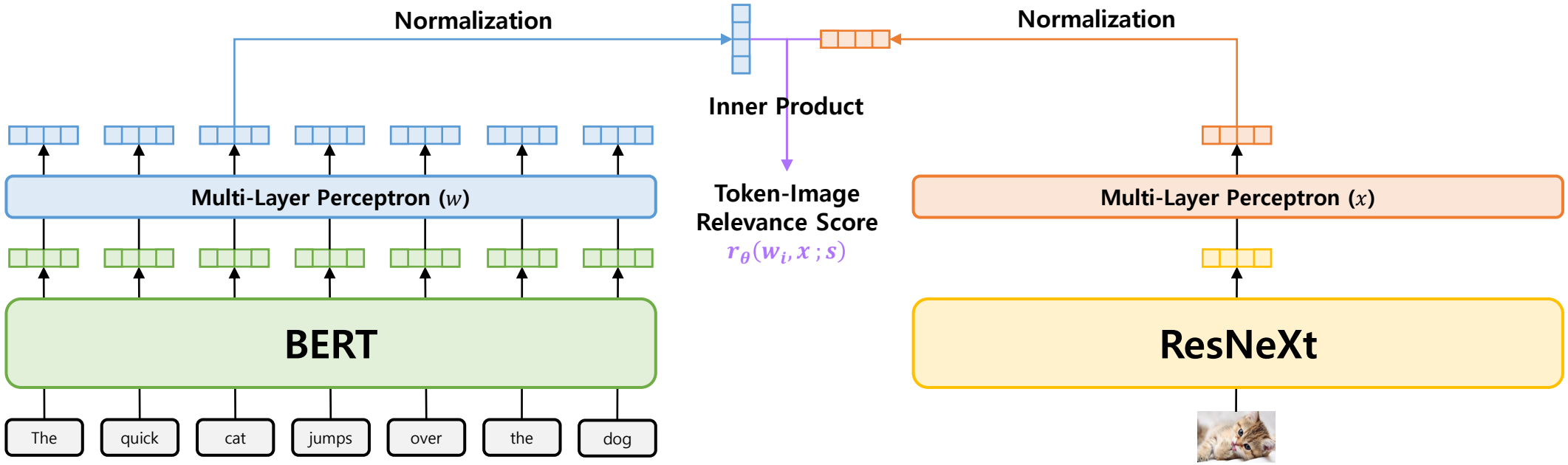
$$f_\theta(w_i\,; s) : language\ Feature\ Representation$$

$$g_\theta(x) : Image\ Feature\ Representation$$

$$h_1, \cdots, h_l = BERT(w_1, \cdots, w_l), \qquad e = ResNeXt(x)$$

$$w\_mlp_\theta, x\_mlp_\theta : Multi\text{-}Layer\ Perceptron$$

$$f_\theta(w_i\,; s) = \frac{w\_mlp_\theta(h_i)}{||w\_mlp_\theta(h_i)||}, \qquad g_\theta(x) = \frac{x\_mlp_\theta(e)}{||x\_mlp_\theta(e)||}$$



Vokenization

-The Vokenization Process

<Modeling>

$$r_\theta(w_i, x; s) = f_\theta(w_i\,; s)^T g_\theta(x) : Inner\ Product\ Factorization$$

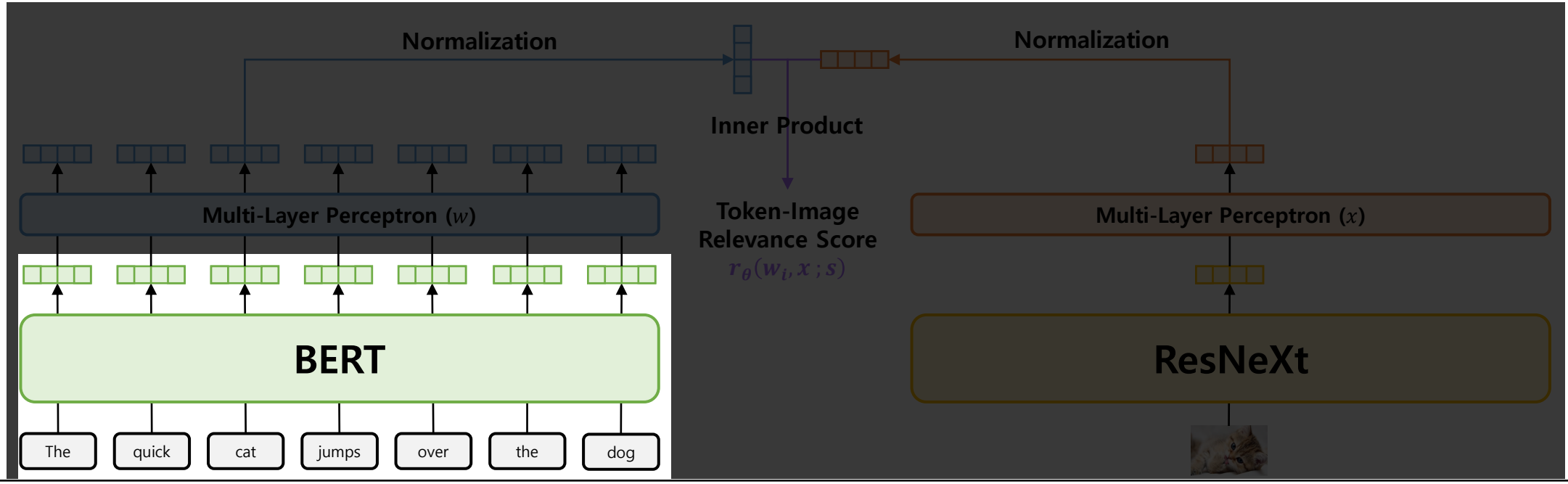
$$f_\theta(w_i\,; s) : language\ Feature\ Representation$$

$$g_\theta(x) : Image\ Feature\ Representation$$

$$\boldsymbol{h_1, \cdots, h_l = BERT(w_1, \cdots, w_l)}, \qquad e = ResNeXt(x)$$

$$w\_mlp_\theta, x\_mlp_\theta : Multi\text{-}Layer\ Perceptron$$

$$f_\theta(w_i\,; s) = \frac{w\_mlp_\theta(h_i)}{||w\_mlp_\theta(h_i)||}, \qquad g_\theta(x) = \frac{x\_mlp_\theta(e)}{||x\_mlp_\theta(e)||}$$



Vokenization

-The Vokenization Process

<Modeling>

$$r_\theta(w_i, x; s) = f_\theta(w_i; s)^T g_\theta(x) : Inner\ Product\ Factorization$$

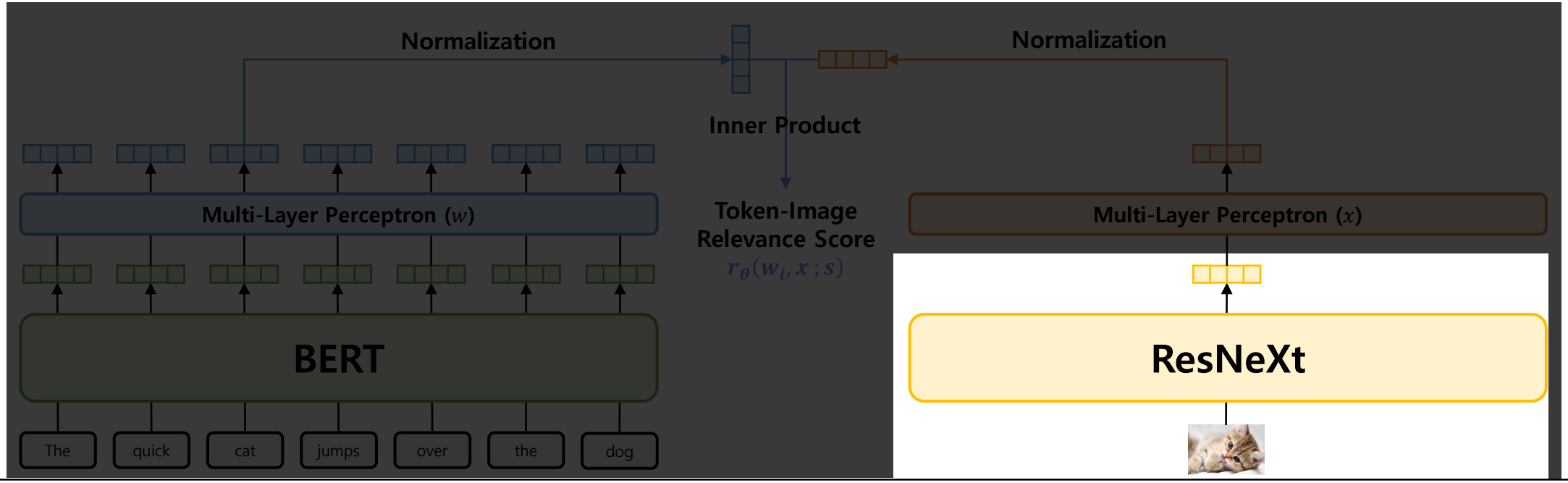
$$f_\theta(w_i; s) : language\ Feature\ Representation$$

$$g_\theta(x) : Image\ Feature\ Representation$$

$$h_1, \cdots, h_l = BERT(w_1, \cdots, w_l), \qquad \boldsymbol{e = ResNeXt(x)}$$

$$w\_mlp_\theta, x\_mlp_\theta : Multi\text{-}Layer\ Perceptron$$

$$f_\theta(w_i; s) = \frac{w\_mlp_\theta(h_i)}{||w\_mlp_\theta(h_i)||}, \qquad g_\theta(x) = \frac{x\_mlp_\theta(e)}{||x\_mlp_\theta(e)||}$$



Vokenization

-The Vokenization Process

<Modeling>

$$r_\theta(w_i, x; s) = f_\theta(w_i\,; s)^T g_\theta(x) : Inner\ Product\ Factorization$$

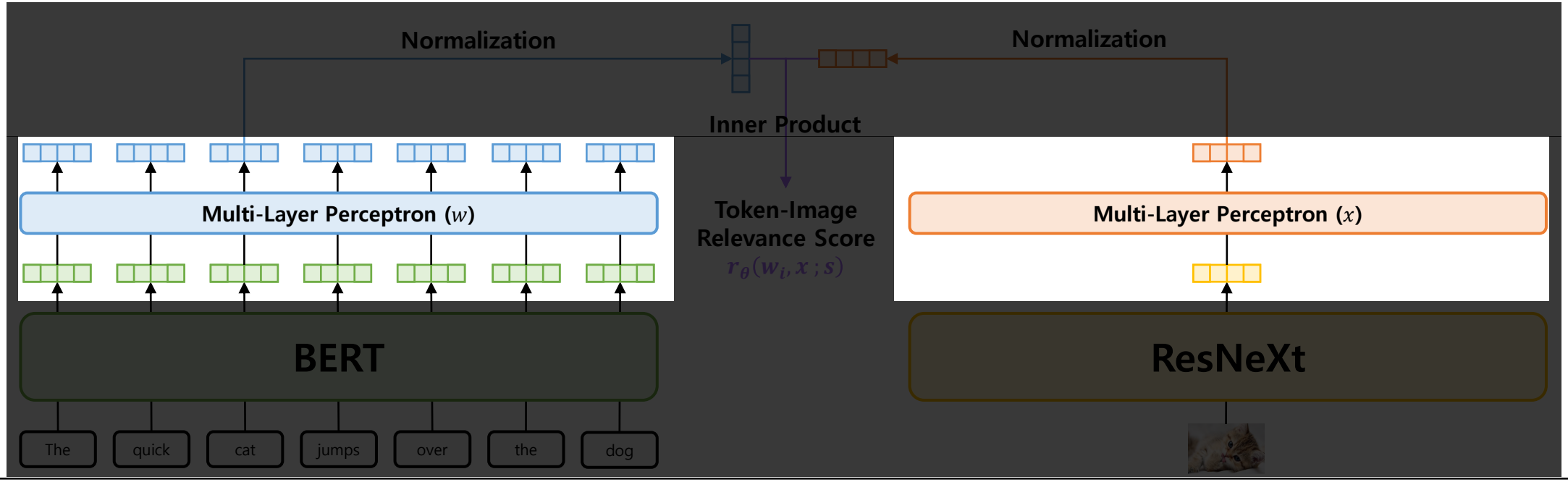
$$f_\theta(w_i\,; s) : language\ Feature\ Representation$$

$$g_\theta(x) : Image\ Feature\ Representation$$

$$h_1, \cdots, h_l = BERT(w_1, \cdots, w_l), \qquad e = ResNeXt(x)$$

$$\boldsymbol{w\_mlp_\theta, x\_mlp_\theta : Multi\text{-}Layer\ Perceptron}$$

$$f_\theta(w_i\,; s) = \frac{w\_mlp_\theta(h_i)}{||w\_mlp_\theta(h_i)||}, \qquad g_\theta(x) = \frac{x\_mlp_\theta(e)}{||x\_mlp_\theta(e)||}$$



Vokenization

-The Vokenization Process

<Modeling>

$$r_\theta(w_i, x; s) = f_\theta(w_i; s)^T g_\theta(x) : \textbf{Inner Product Factorization}$$

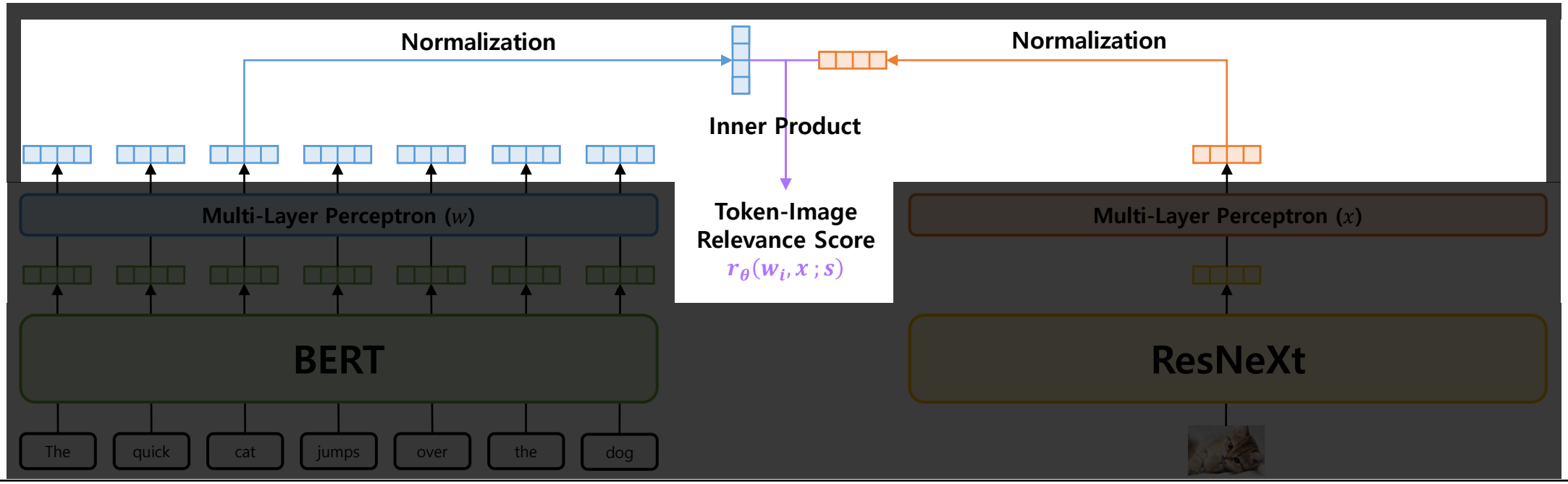
$f_\theta(w_i; s)$: language Feature Representation

$g_\theta(x)$: Image Feature Representation

$h_1, \cdots, h_l = BERT(w_1, \cdots, w_l), \qquad e = ResNeXt(x)$

$w\_mlp_\theta, x\_mlp_\theta$: Multi-Layer Perceptron

$$f_\theta(w_i; s) = \frac{w\_mlp_\theta(h_i)}{||w\_mlp_\theta(h_i)||}, \qquad g_\theta(x) = \frac{x\_mlp_\theta(e)}{||x\_mlp_\theta(e)||}$$



<MS COCO Dataset>

"The quick cat jumps over the dog."

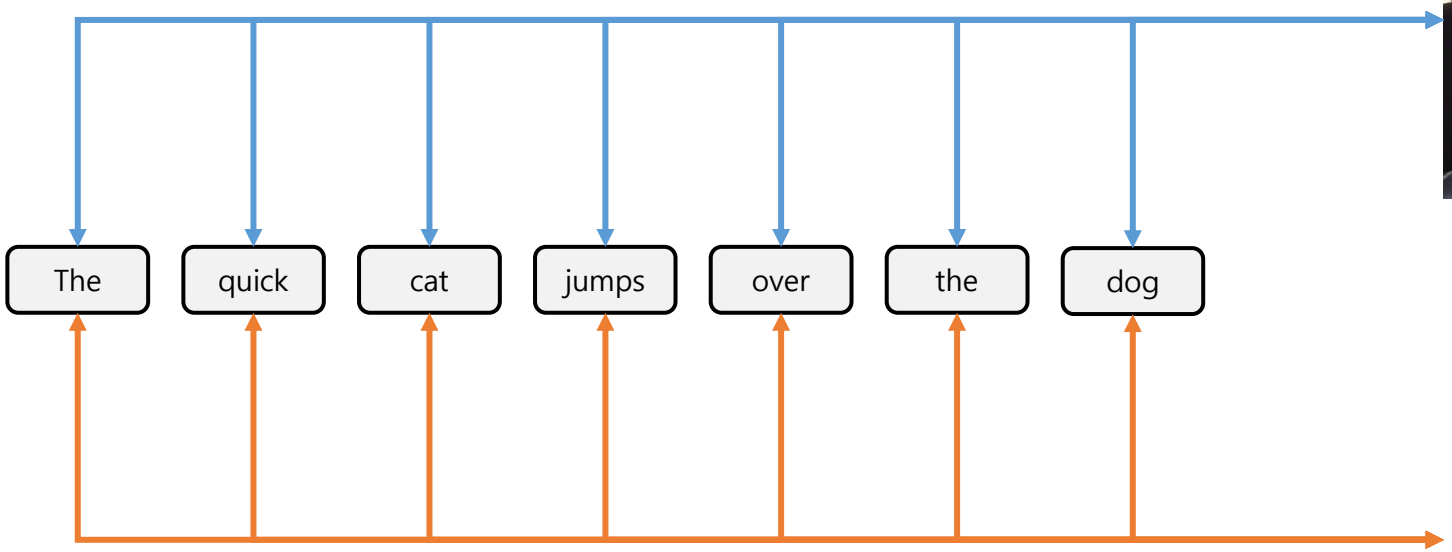
<Sentence>

<Image That Semantically Matches Given Sentence>

Vokenization

-The Vokenization Process

<Training>



<Tokens in Given Sentence>

<Given Image>

<Randomly Sampled Image>

Vokenization

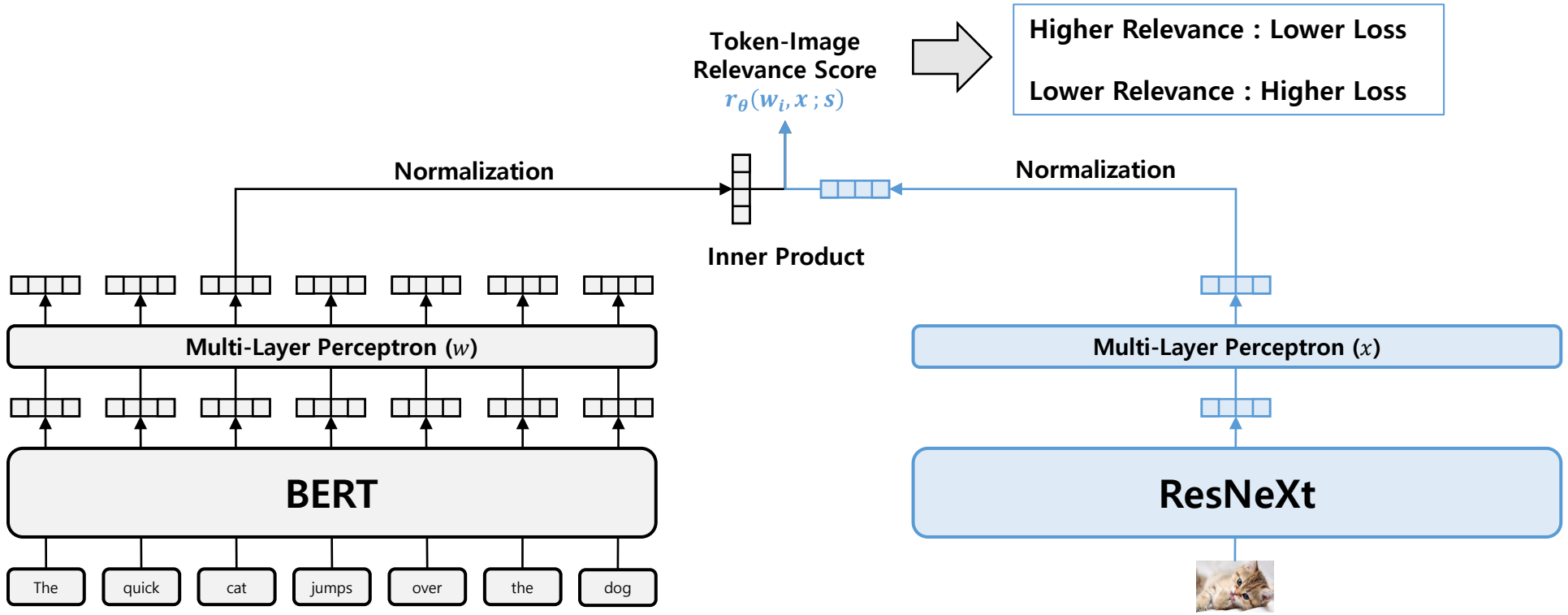
-The Vokenization Process

<Training>

(s, x): *Ground Truth Image*– *Captioning Data Point*

x': *Randomly Sampled Image, where* $x' \neq x$

$$\mathcal{L}_\theta(s, x, x') = \sum_{i=1}^{l} \max\{0, M - r_\theta(w_i, x; s) + r_\theta(w_i, x'; s)\}$$



Vokenization

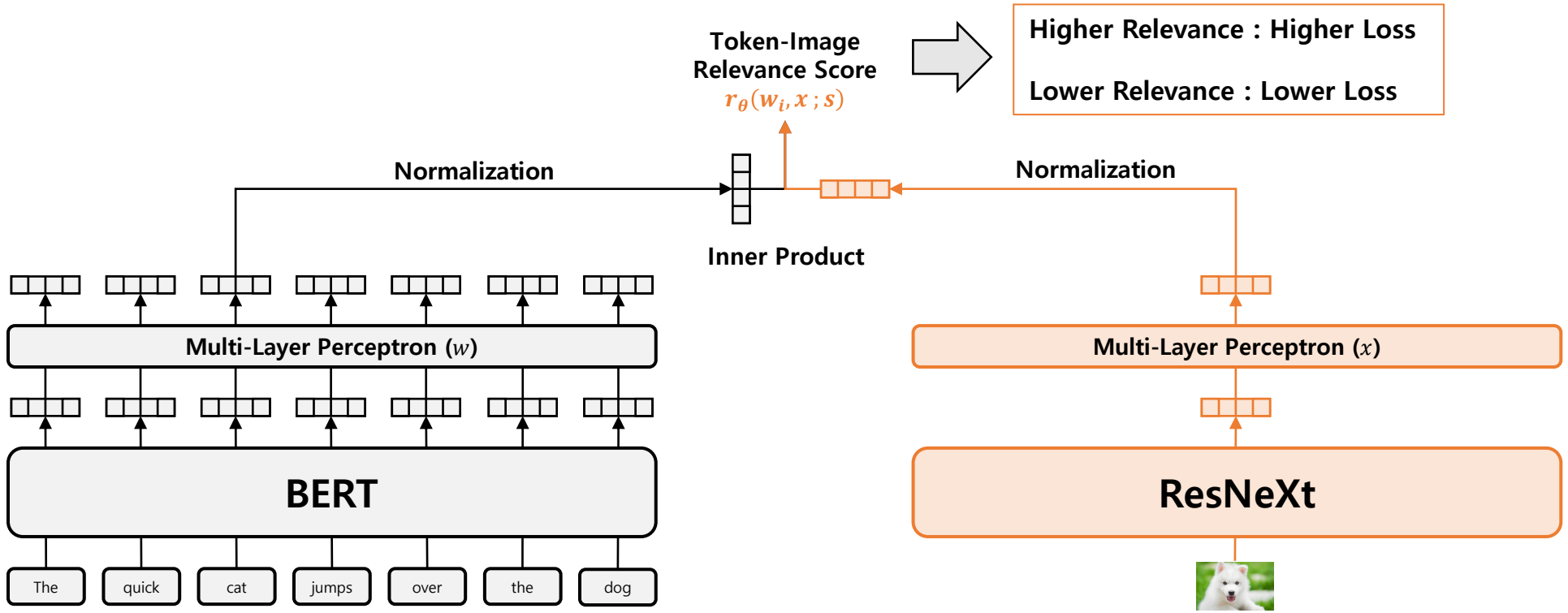
-The Vokenization Process

<Training>

(s, x): *Ground Truth Image–Captioning Data Point*

x': *Randomly Sampled Image*, *where* $x' \neq x$

$$\mathcal{L}_\theta(s, x, x') = \sum_{i=1}^{l} \max\{0, M - r_\theta(w_i, x; s) + r_\theta(w_i, x'; s)\}$$

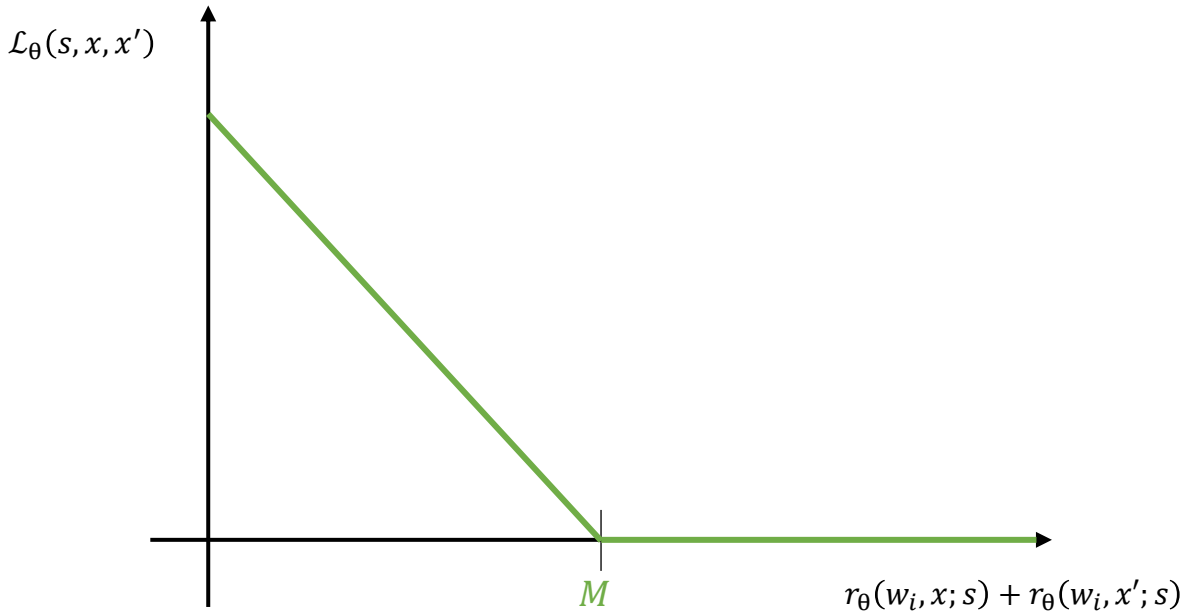


<Training>

(s, x): *Ground Truth Image* – *Captioning Data Point*

x': *Randomly Sampled Image*, *where* $x' \neq x$

$$\mathcal{L}_\theta(s, x, x') = \sum_{i=1}^{l} \max\{0, M - r_\theta(w_i, x; s) + r_\theta(w_i, x'; s)\}$$



Vokenization

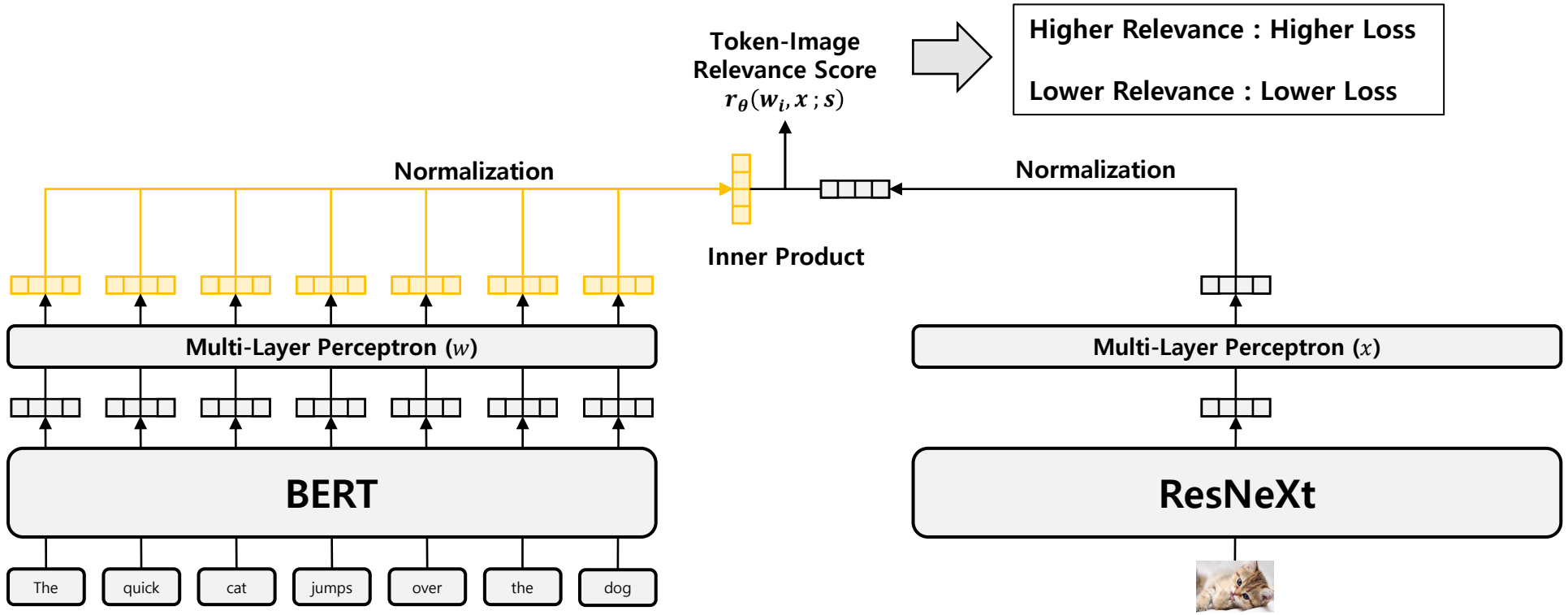
-The Vokenization Process

<Training>

(s, x): *Ground Truth Image–Captioning Data Point*

x': *Randomly Sampled Image, where* $x' \neq x$

$$\mathcal{L}_\theta(s, x, x') = \sum_{i=1}^{l} \max\{0, M - r_\theta(w_i, x; s) + r_\theta(w_i, x'; s)\}$$



Vokenization

-The Vokenization Process

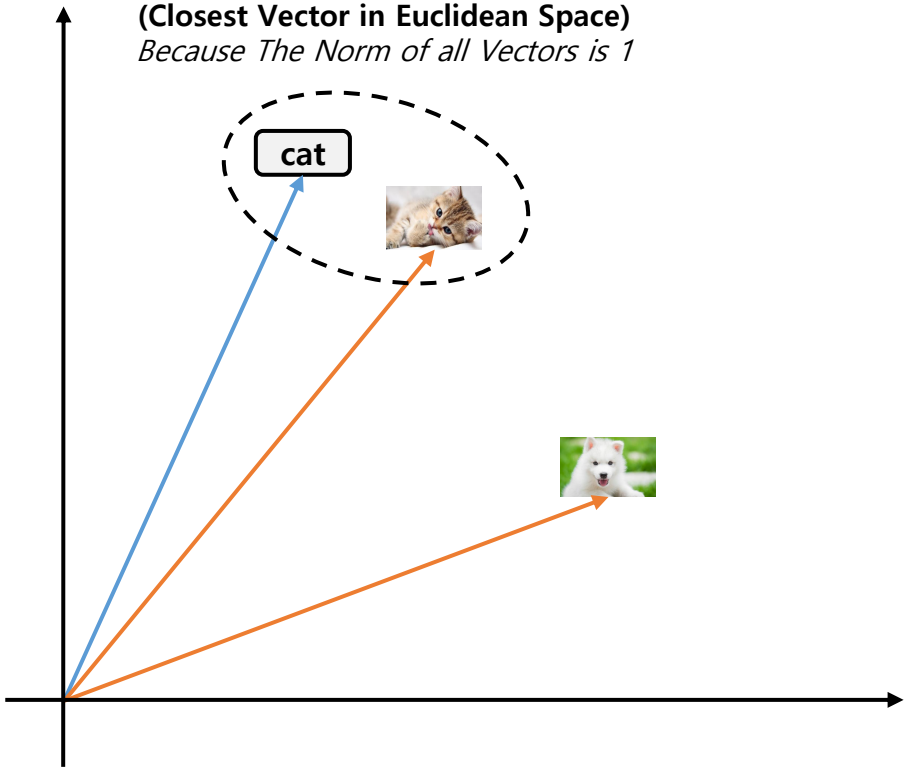
<Inference>

$f_\theta(w_i\,; s)$: *language Feature Representation*

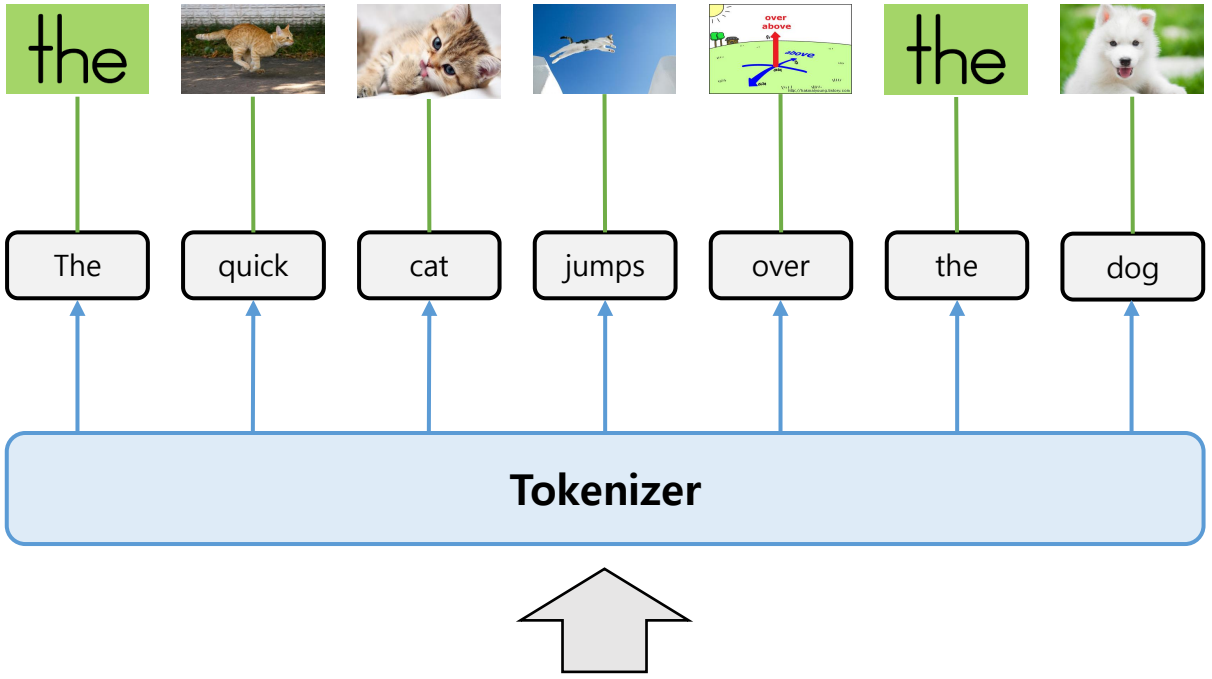
$g_\theta(x)$: *Image Feature Representation*

$$f_\theta(w_i\,; s) = \frac{w\_mlp_\theta(h_i)}{\left||w\_mlp_\theta(h_i)|\right|}, \qquad g_\theta(x) = \frac{x\_mlp_\theta(e)}{\left||x\_mlp_\theta(e)|\right|}$$

Maximum Inner product Search
(Closest Vector in Euclidean Space)
Because The Norm of all Vectors is 1

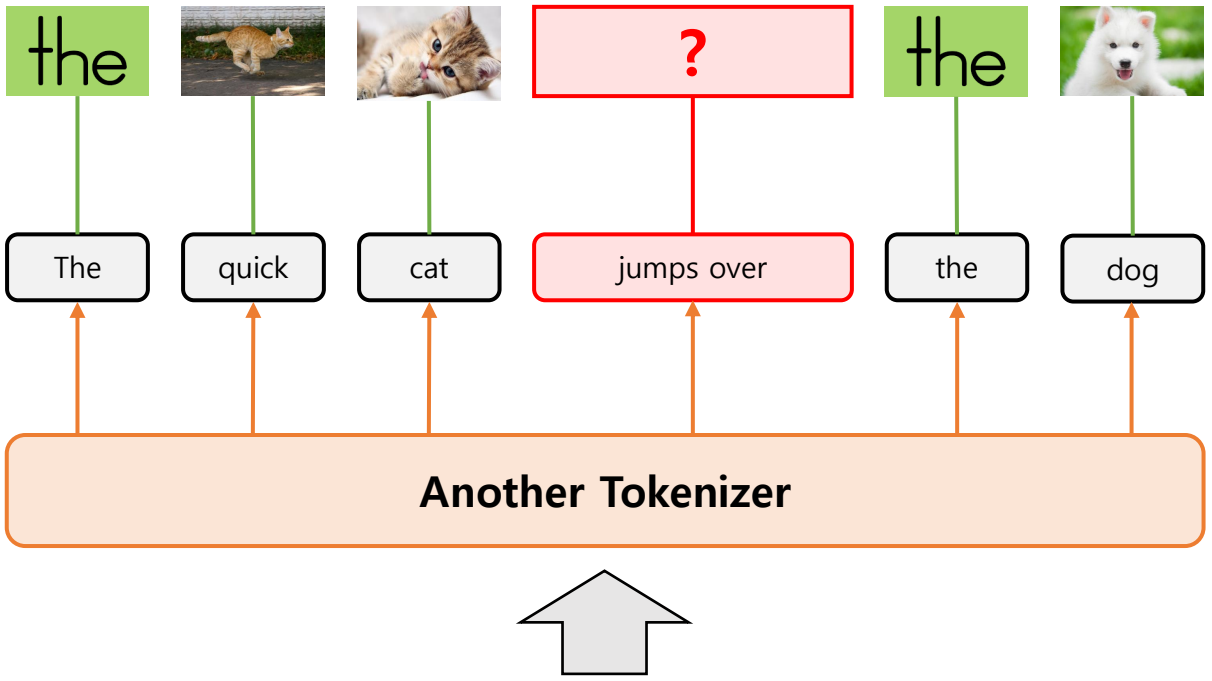


<Revokenization>



“The quick cat jumps over the dog.”

<Revokenization>



"The quick cat jumps over the dog."

<Revokenization>

$$T_1, T_2 : Tokenizer$$

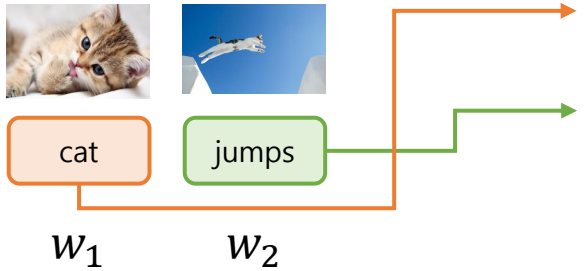
$$T_1(s) = (w_1, \cdots, w_l), T_2(s) = (u_1, \cdots, u_m)$$

$$\{v(w_i; s)\}_{i=1}^{l} : Vokens\ Aligned\ to\ Tokens\ \{w_i\}_{i=1}^{l}$$

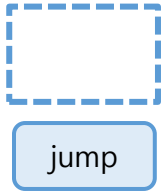
$$v(u_j; s) := v(w_{ind(j)}; s)$$

$$ind(j) = argmax_{i=1}^{l}\ overlap(w_i, u_j)$$

$$overlap(w_i, u_j) = \frac{|w_i \cap u_j|}{|w_i \cup u_j|}$$



w_1 w_2



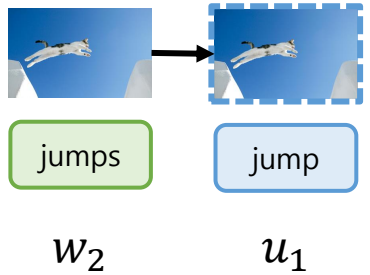
u_1

$$overlap(w_1, u_1) = \frac{|w_1 \cap u_1|}{|w_1 \cup u_1|} = \frac{|\emptyset|}{|\text{catjump}|} = \frac{0}{7} = 0$$

$$overlap(w_2, u_1) = \frac{|w_2 \cap u_1|}{|w_2 \cup u_1|} = \frac{|\text{jump}|}{|\text{jumps}|} = \frac{4}{5} = 0.8$$

$$ind(1) = argmax_{i=1}^{2}\ overlap(w_i, u_1) = 2$$

$$v(u_1; s) := v(w_2; s)$$



w_2 u_1

Experiments

- Data, Task & Details
- Result

<Data, Task & Details>

<Pre-Training Data>

English Wikipedia, Wiki103

<Fine-Tuning Task>

GLUE, SQuAD, SWAG, SST-2, QNLI, QQP, MNLI

<Implementation Details>

Concatenation of last 4 layers of BERT

ResNeXt-101-32x8d

Two Fully-Connected Multi-Layer Perceptron, 256-dim ($w\_mlp_\theta, x\_mlp_\theta$)

ReLU, 64-dim output

M = 0.5, λ = 1, vocab_size = 50,000, batch_size = 256, epochs = 3

Removing Next-Sentence Prediction

12-layer $BERT_{BASE}$, 768-dim, **From Scratch**

Experiments

-Result

<Result>

Method	STT-2	QNLI	QQP	MNLI	SQuAD v1.1	SQuAD v2.0	SWAG	Avg.
BERT $_{6L/512H}$	88.0	85.2	87.1	77.9	71.3/80.2	57.2/60.8	56.2	75.6
BERT $_{6L/512H}$ + Voken-cls	89.7	85.0	87.3	78.6	71.5/80.2	61.3/64.6	58.2	76.8
BERT $_{12L/768H}$	89.3	87.9	83.2	79.4	77.0/85.3	67.7/71.1	65.7	79.4
BERT $_{12L/768H}$ + Voken_cls	**92.2**	**88.6**	**88.6**	**82.6**	**78.8/86.7**	68.1/71.2	**70.6**	**82.1**
RoBERTa $_{6L/512H}$	87.8	82.4	85.2	73.1	50.9/61.9	49.6/52.7	55.1	70.2
RoBERTa $_{6L/512H}$ + Voken_cls	87.8	85.1	85.3	76.5	55.0/66.4	50.9/54.1	60.0	72.6
RoBERTa $_{12L/768H}$	89.2	87.5	86.2	79.0	70.2/79.9	59.2/63.1	65.2	77.6
RoBERTa $_{12L/768H}$ + Voken_cls	**90.5**	**89.2**	**87.8**	**81.0**	**73.0/82.5**	**65.9/69.3**	**70.4**	**80.6**

Experiments

-Result

<Result>

Method	STT-2	QNLI	QQP	MNLI	SQuAD v1.1	SQuAD v2.0	SWAG	Avg.
BERT $_{6L/512H}$	88.0	85.2	87.1	77.9	71.3/80.2	57.2/60.8	56.2	75.6
BERT $_{6L/512H}$ + Voken-cls	**89.7**	**85.0**	**87.3**	**78.6**	**71.5/80.2**	**61.3/64.6**	**58.2**	**76.8**
BERT $_{12L/768H}$	89.3	87.9	83.2	79.4	77.0/85.3	67.7/71.1	65.7	79.4
BERT $_{12L/768H}$ + Voken_cls	**92.2**	**88.6**	**88.6**	**82.6**	**78.8/86.7**	**68.1/71.2**	**70.6**	**82.1**
RoBERTa $_{6L/512H}$	87.8	82.4	85.2	73.1	50.9/61.9	49.6/52.7	55.1	70.2
RoBERTa $_{6L/512H}$ + Voken_cls	**87.8**	**85.1**	**85.3**	**76.5**	**55.0/66.4**	**50.9/54.1**	**60.0**	**72.6**
RoBERTa $_{12L/768H}$	89.2	87.5	86.2	79.0	70.2/79.9	59.2/63.1	65.2	77.6
RoBERTa $_{12L/768H}$ + Voken_cls	**90.5**	**89.2**	**87.8**	**81.0**	**73.0/82.5**	**65.9/69.3**	**70.4**	**80.6**

Analysis

- Visualization of Vokens

<Visualization of Vokens>

"Humans learn language by listening, speaking, writing, reading"

"Down by the salley gardens my lobe and I did meet"

<Visualization of Vokens>

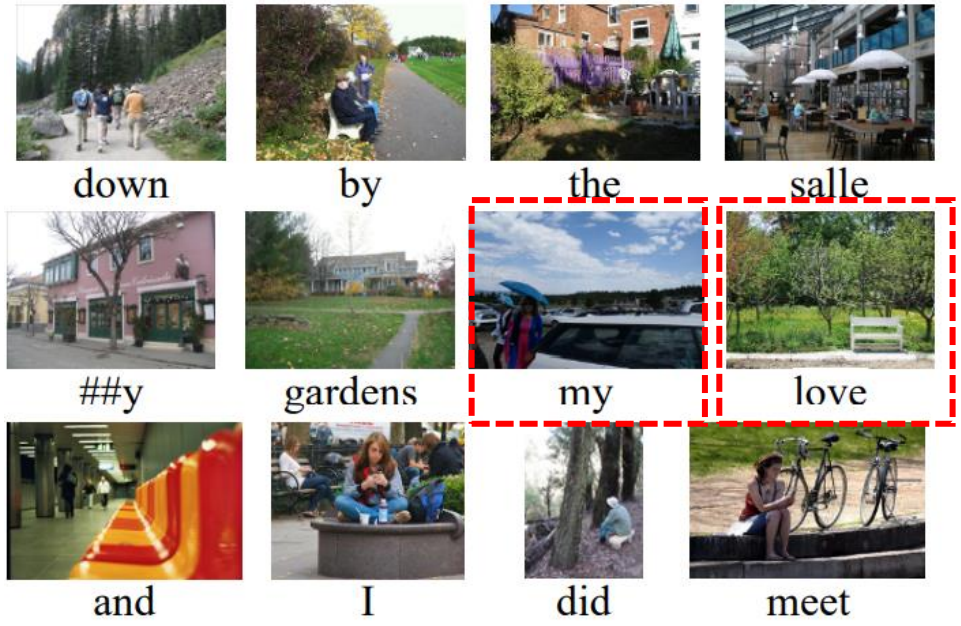
"Humans learn language by listening, speaking, writing, reading"

"Down by the salley gardens my lobe and I did meet"

The Vokenizer Maps the Voken in Consideration of the Context and Can Map Vokens to Non-Concrete Tokens

<Visualization of Vokens>

"Humans learn language by listening, speaking, writing, reading"



"Down by the salley gardens my lobe and I did meet"

Some Vokens Has a Shift in Alignment Which May be Caused by Limitations of Sentence-Image Weak Supervision

<Visualization of Vokens>

"Humans learn language by listening, speaking, writing, reading"

"Down by the salley gardens my lobe and I did meet"

The Related Visual Information Can Help Understand the Language and Lead to the Improvement

Conclusion

<Conclusion>

- This Paper Explored the Possibility of Utilizing Visual Supervision to Language Encoder.
- Vokenizer is Contextual Token-Image Matching Model and Can be Used to Vokenize the Language Corpus.
- Supervised by These Generated Vokens, The Language Model Has a Significant Improvement Over the Purely Self-Supervised Language Model on Multiple Language Tasks.

Any Questions?

Thank You