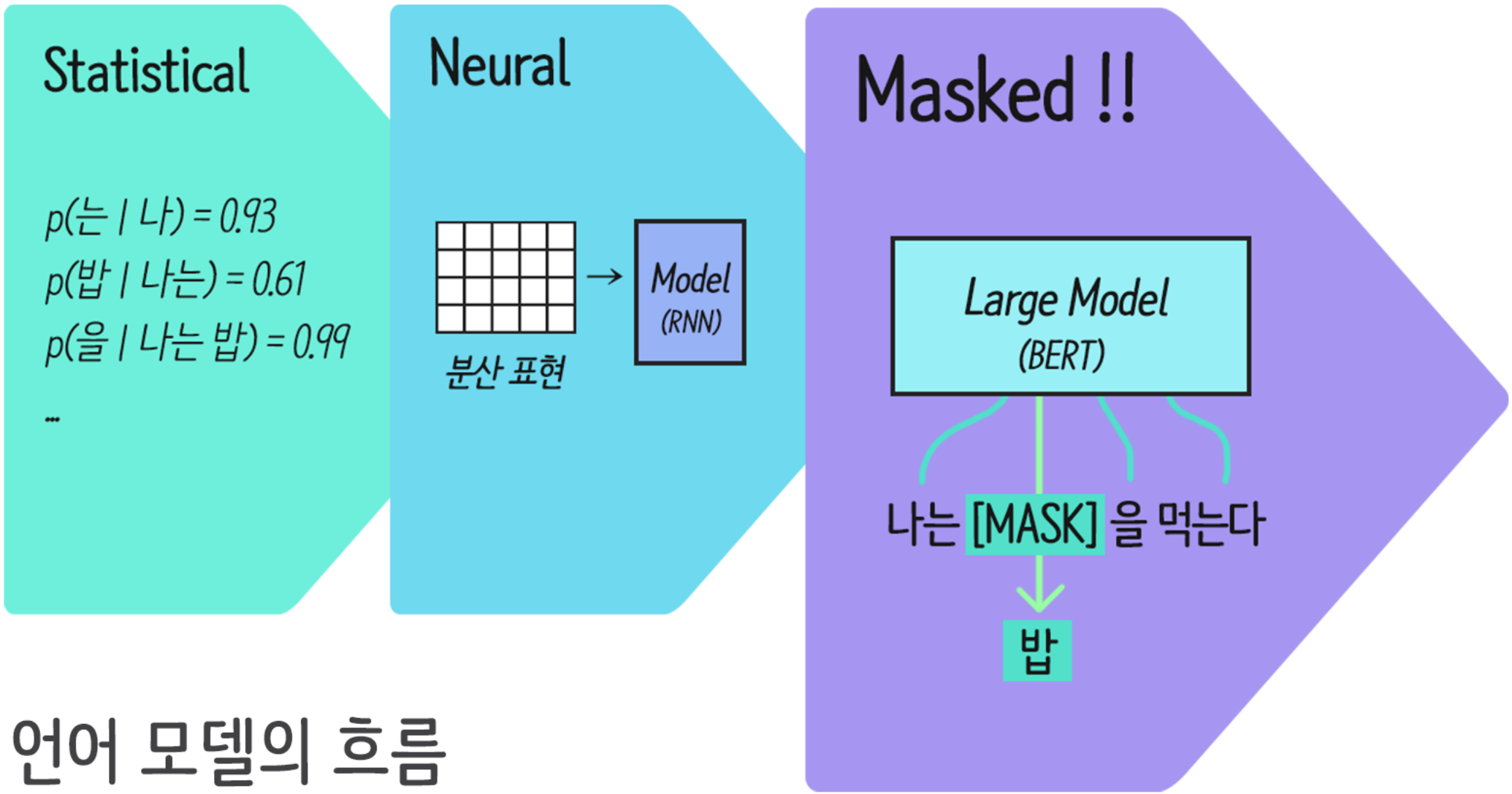
ELECTRA

: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS

2020. 06. 02

Moon Sungwon

INTRO

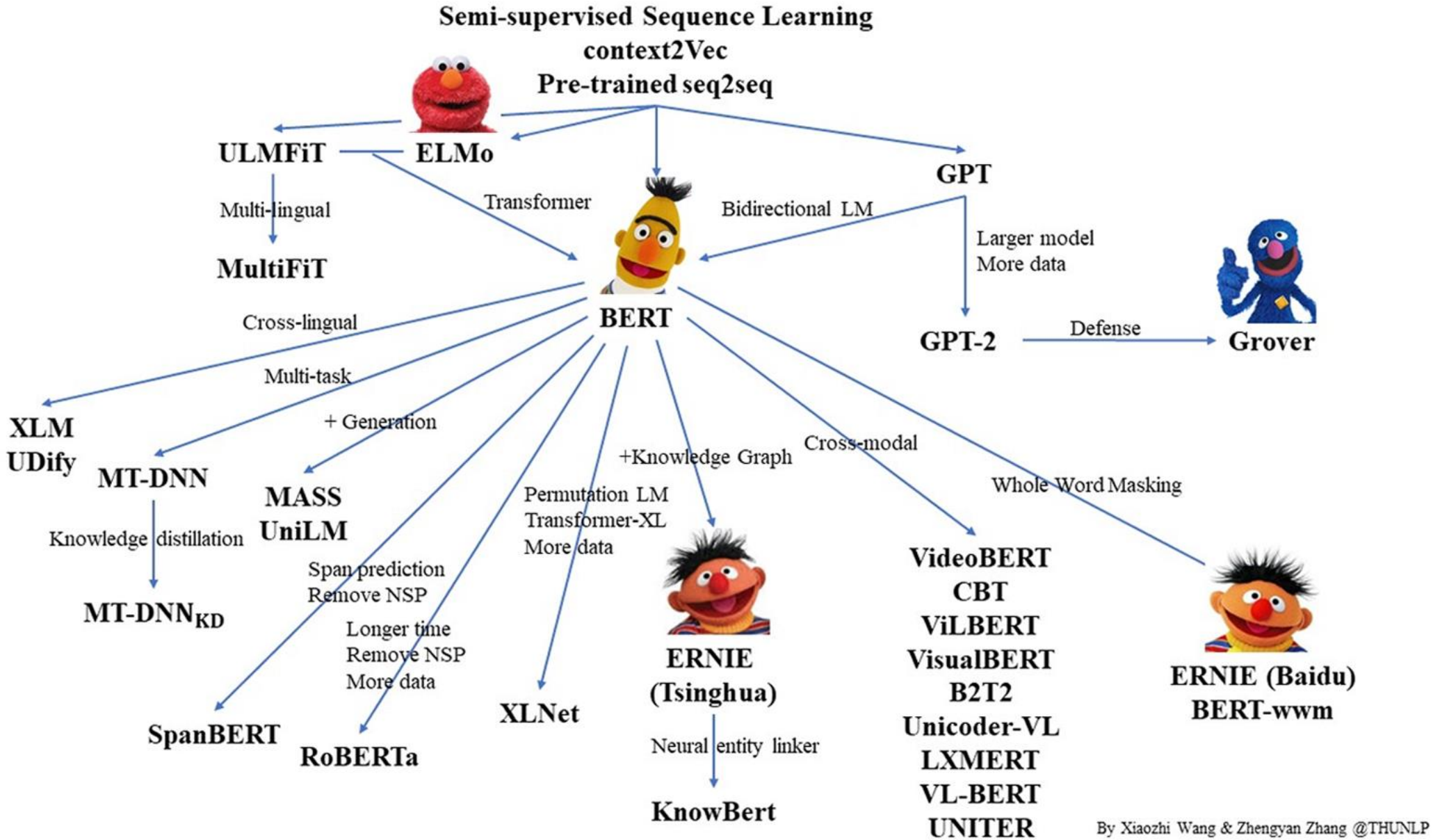


언어 모델의 흐름

MLM의 대표 주자

BERT

GPT-2


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ELMo
ULMFiT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
GPT
Larger model
More data
GPT-2
Defense
Grover
BERT
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

논문 기준, 학습에 필요한 가격

* TPU는 1 device -> 4 chips -> 8 cores 입니다.

Model	Size	TPU ($ per hour)	TPU Count (device)	Training Time	Cost (USD)	CO2 emissions (lbs)
BERT	24 Layers (340M)	v2 ($4.5)	16	4 days	$6,912 (약 850만원)	1428
GPT-2	48 Layers (1542M)	v3 ($8)	32	7 days	$43,008 (약 5,100만원)	2516
XLNet	24 Layers (365M)	v3 ($8)	128	2.5 days	$61,440 (약 7,300만원)	-

* NY ↔ SF Air Travel: 1924 (lbs)

참고: https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/,
Energy and Policy Considerations for Deep Learning in NLP, 2019 E Strubell

출처: Clova AI DEVIEW 2019 〈엄~청 큰 언어 모델 공장 가동기!〉

So... ELECTRA!

GPT-2

120 Days

Google AI

ELECTRA

4 Days

Masked Language Model?

최근엔

마스크

토큰

을

사용하는

언어 모델

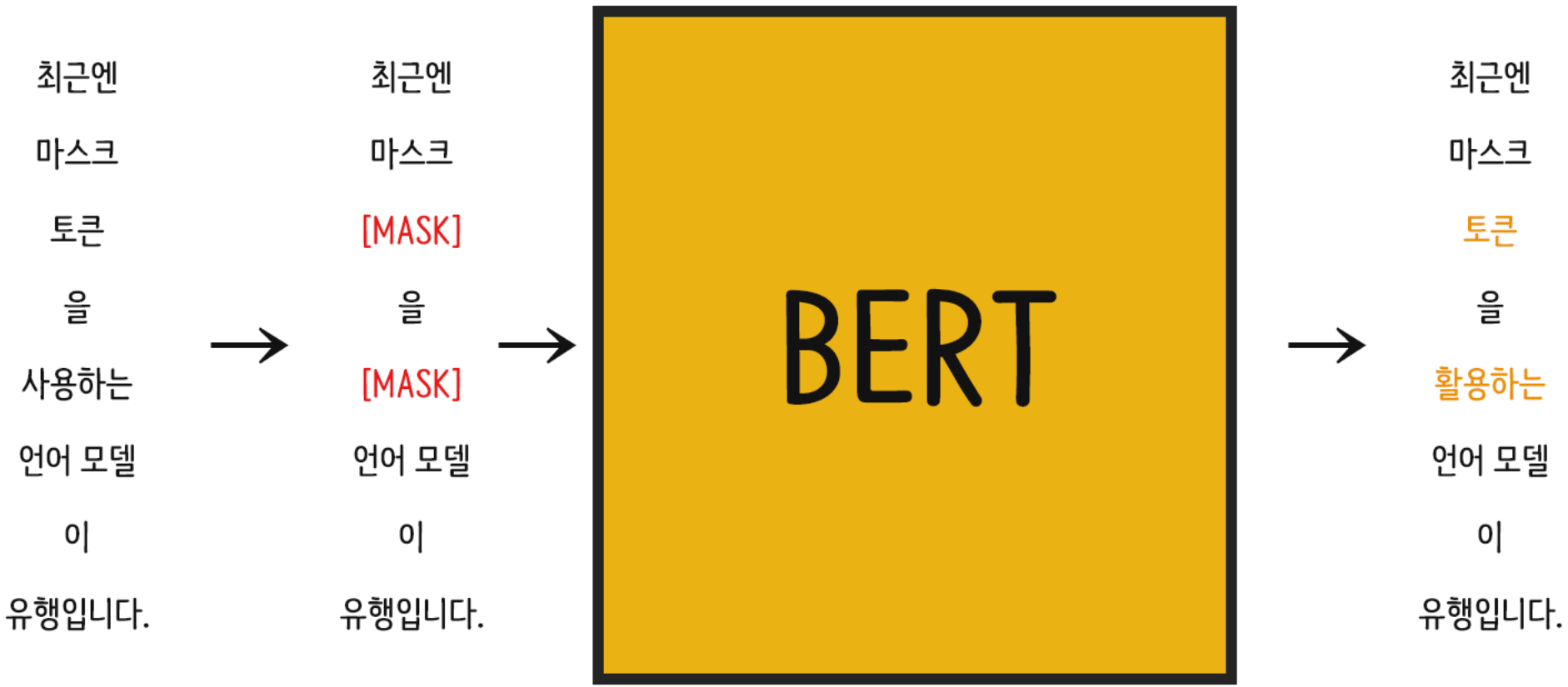
이

유행입니다.

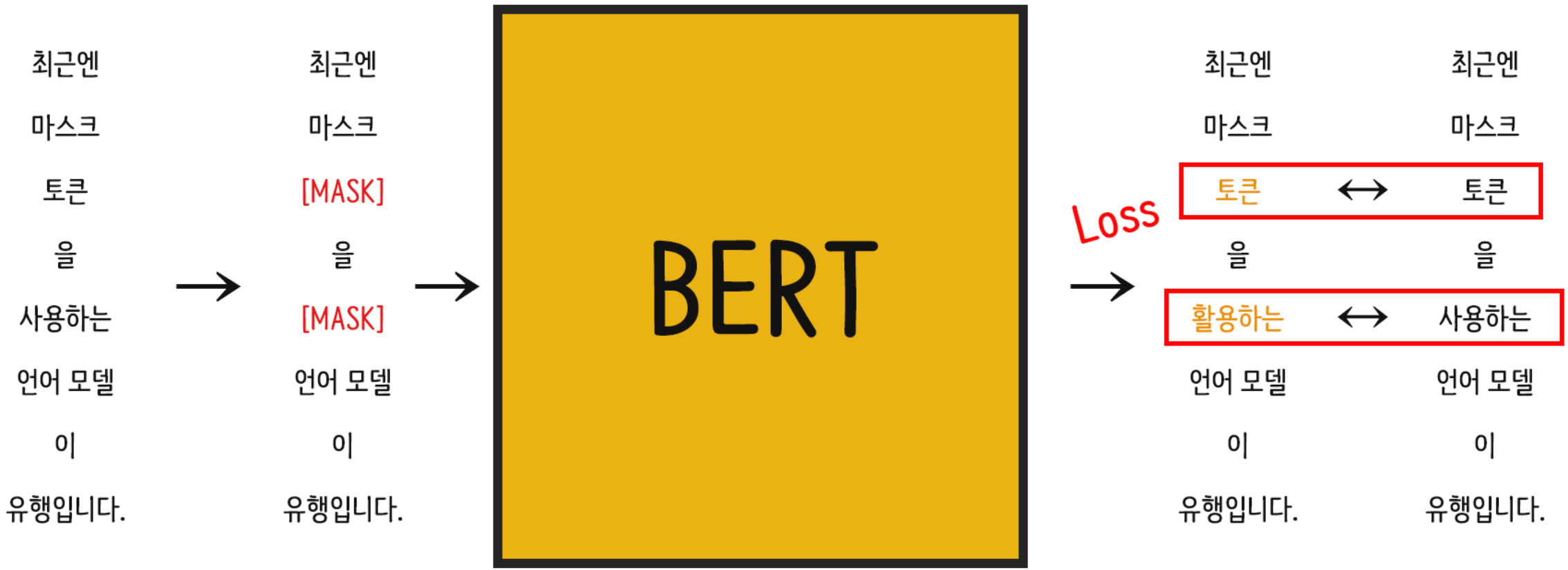
Masked Language Model?

최근엔	최근엔
마스크	마스크
토큰	[MASK]
을	을
사용하는	[MASK]
언어 모델	언어 모델
이	이
유행입니다.	유행입니다.

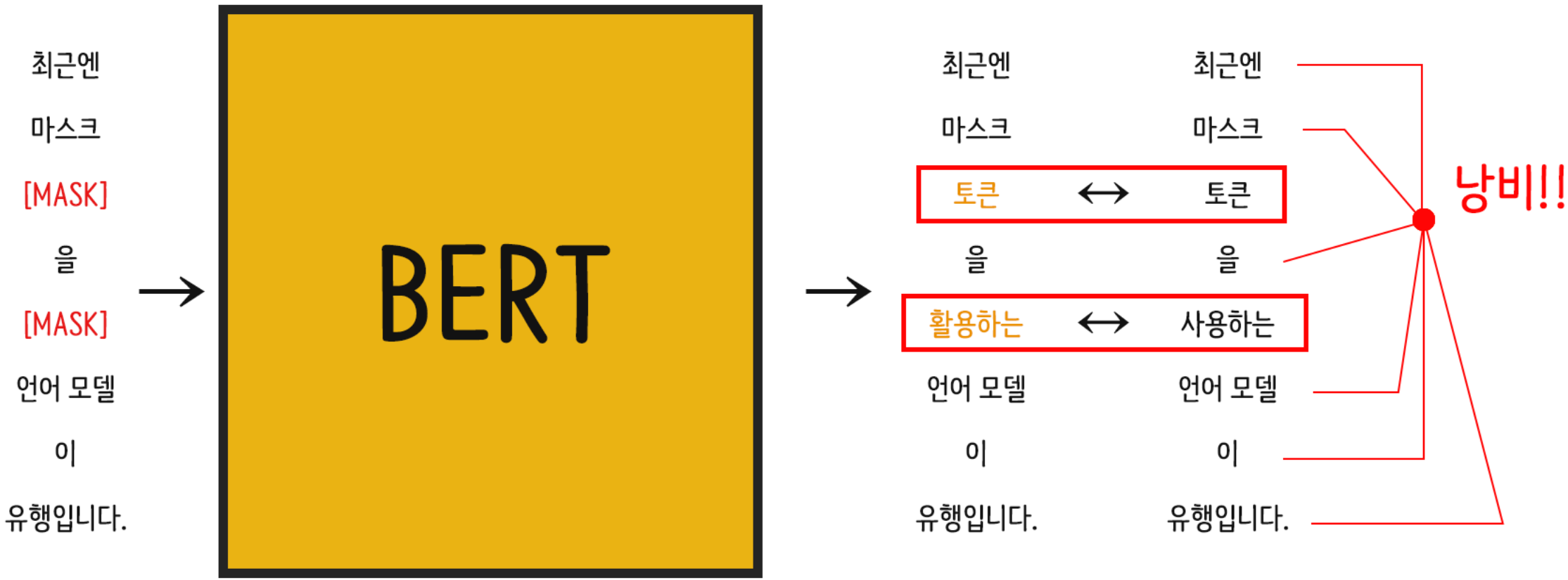
Masked Language Model?



Masked Language Model?

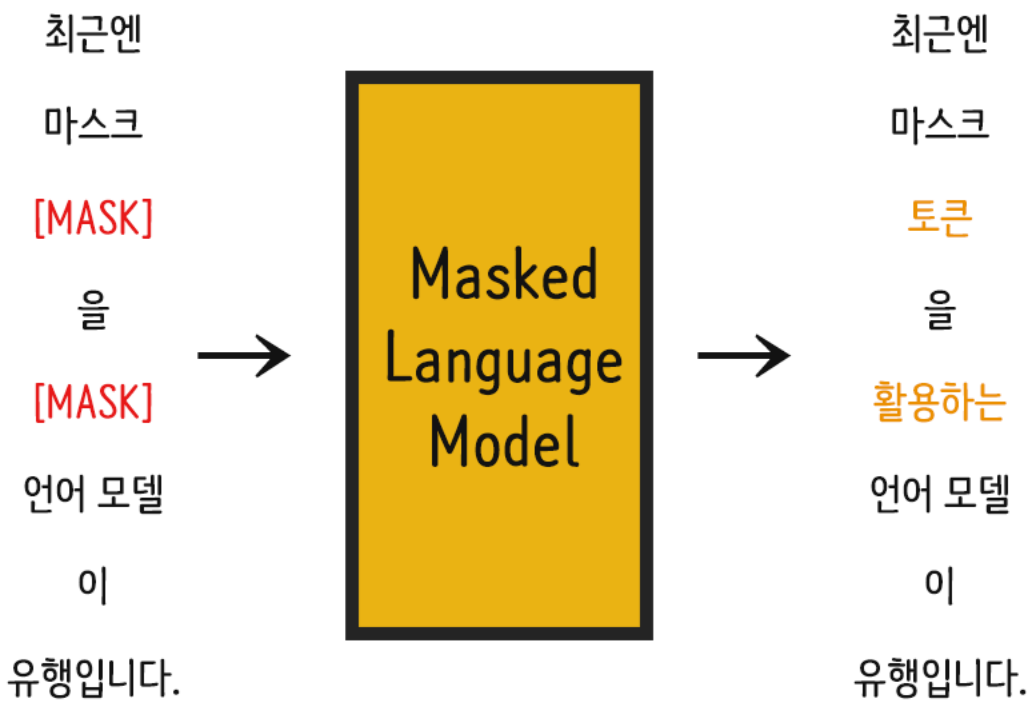


문제는 바로...

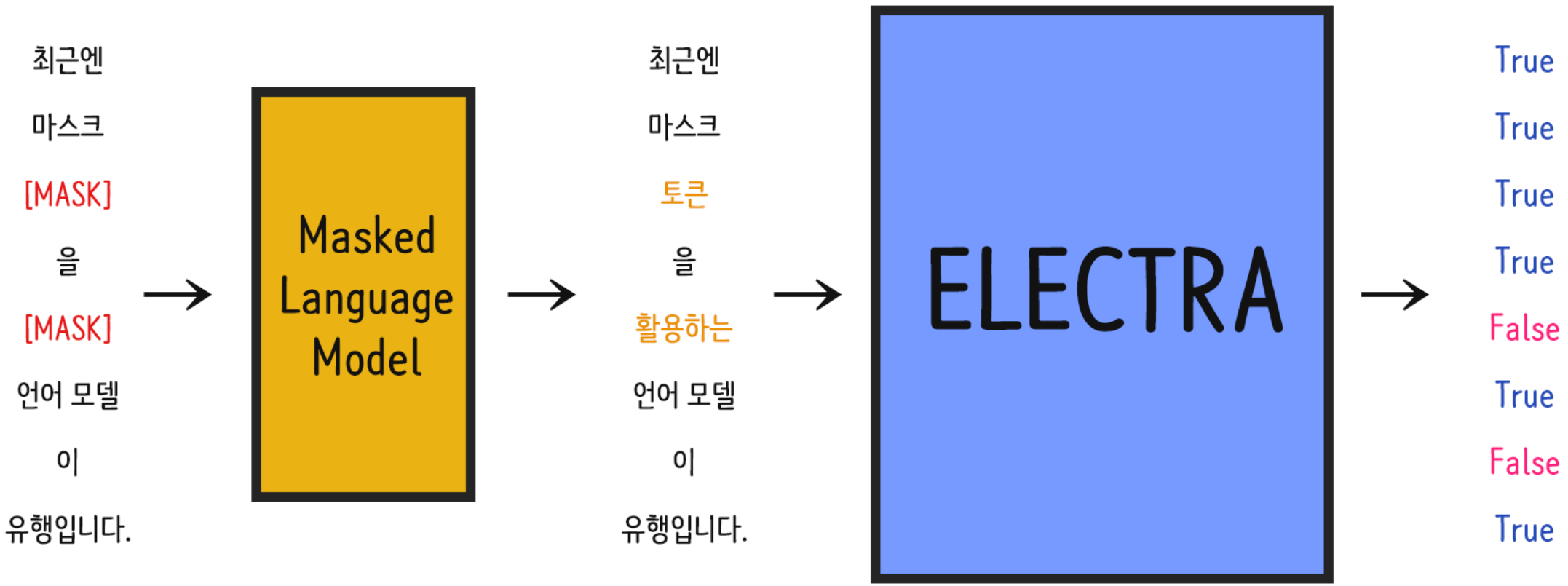


데이터의 15% 밖에 학습할 수 없다!

Replaced Token Detection

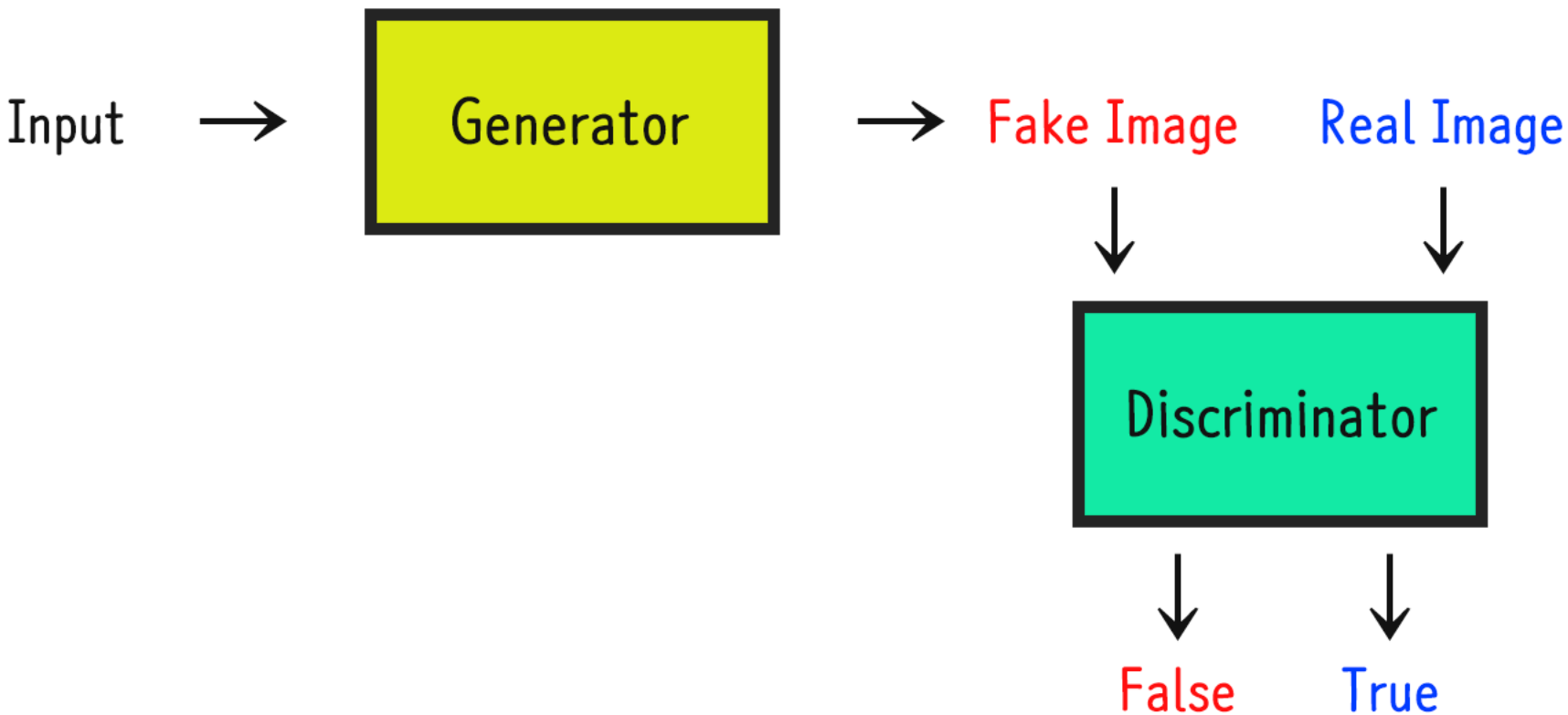


Replaced Token Detection

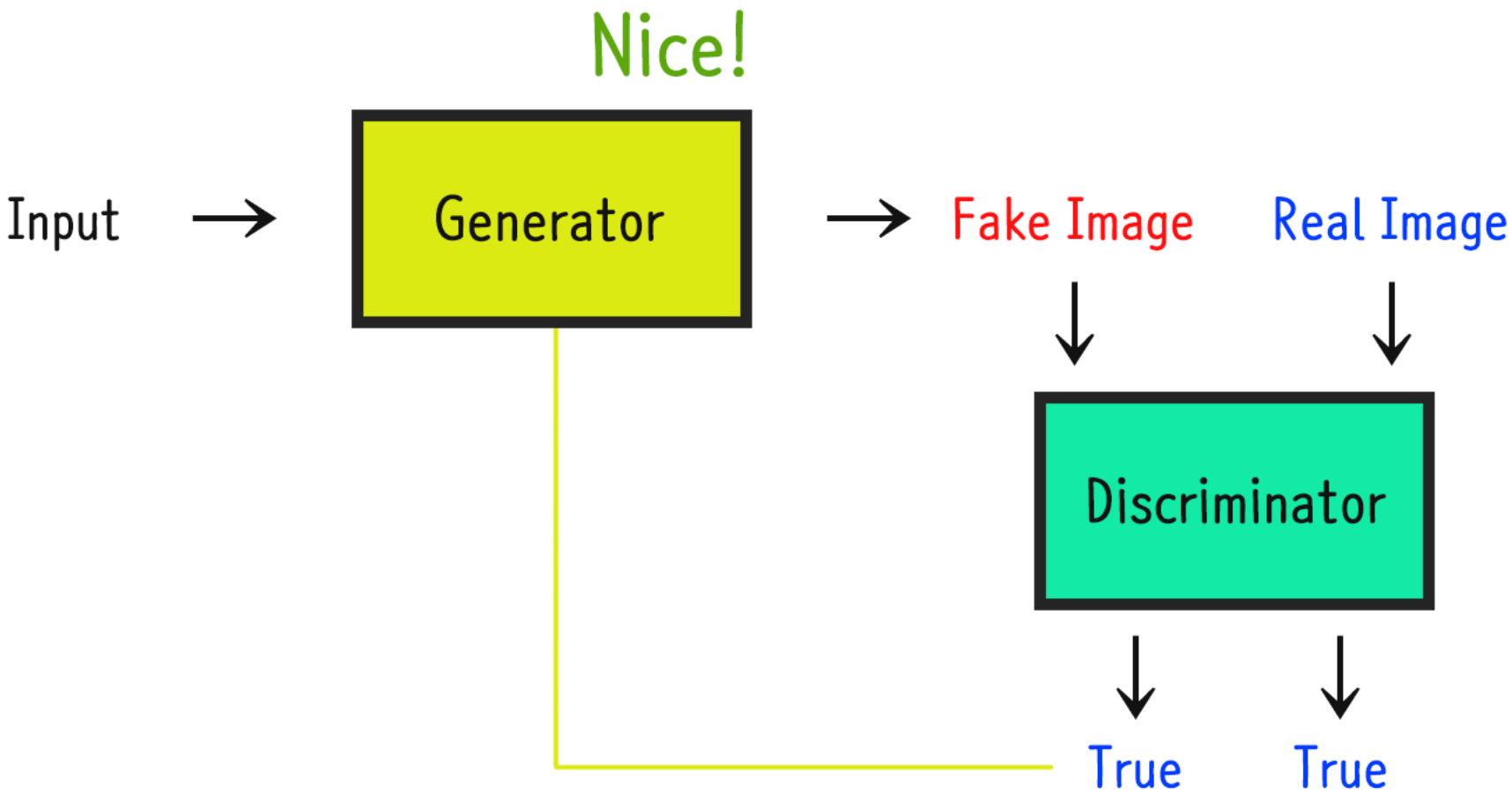


MLM이 생성한 토큰인지 여부를 판별!

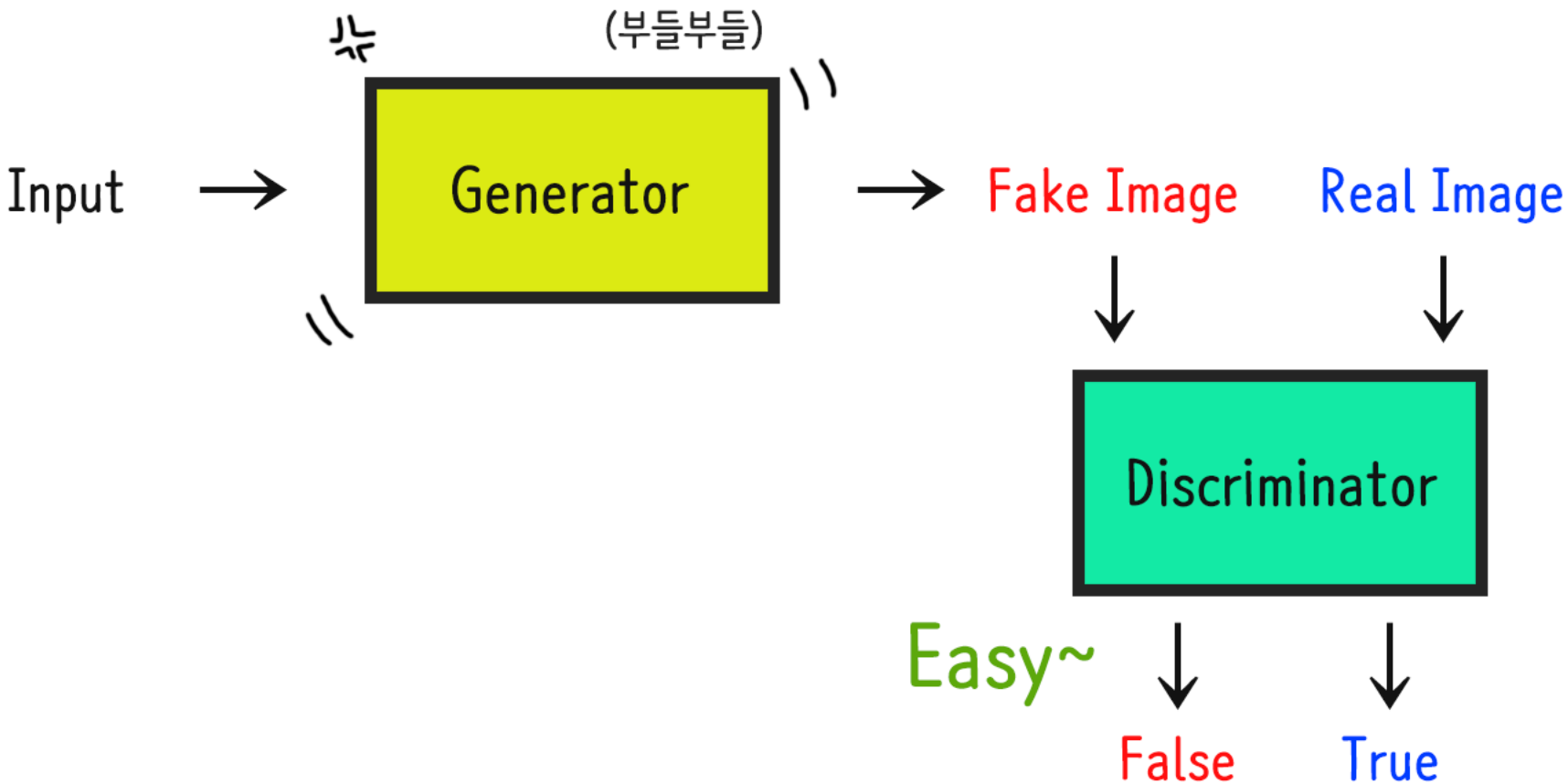
깨알상식 - GAN


Input
Generator
Fake Image
Real Image
Discriminator
False
True

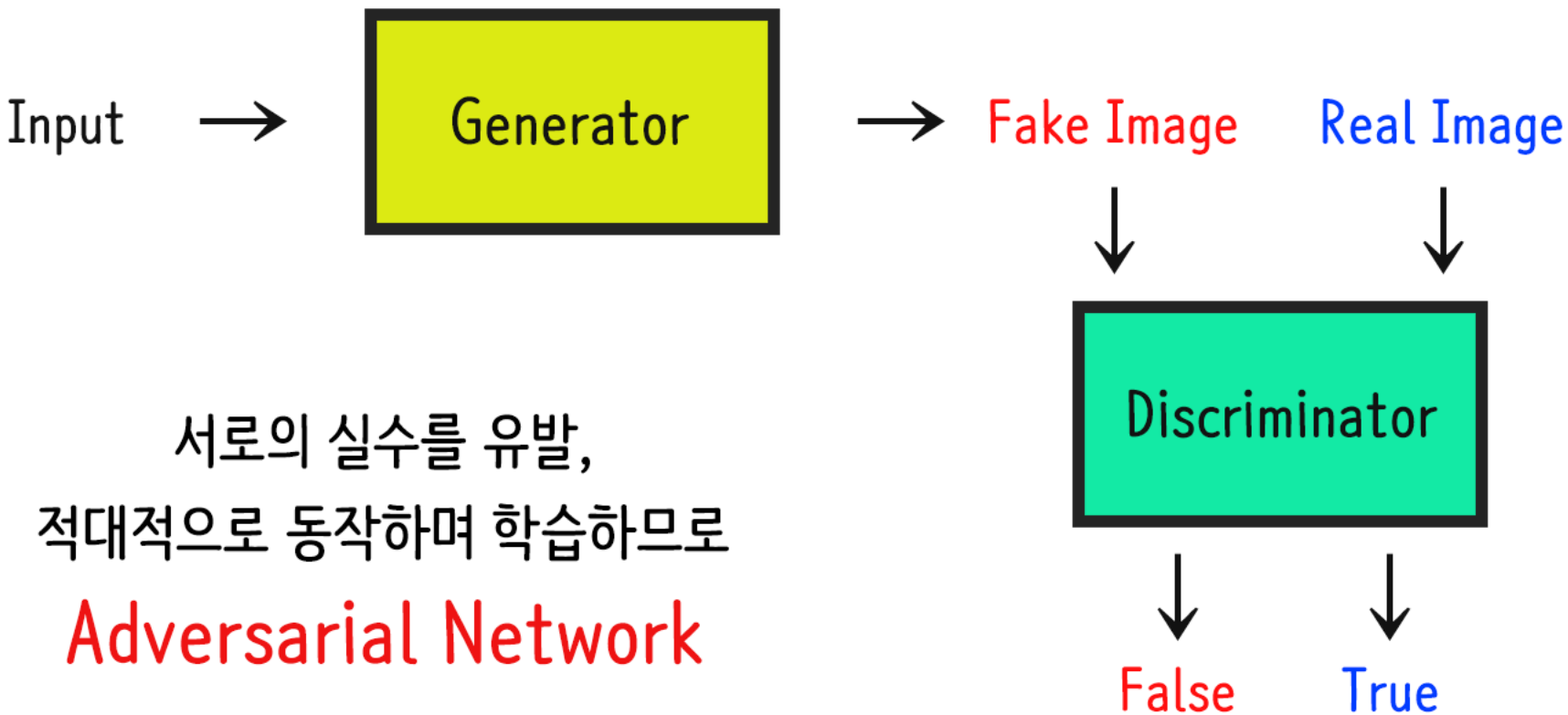
깨알상식 - GAN


Nice!
Input
Generator
Fake Image
Real Image
Discriminator
True
True

깨알상식 - GAN

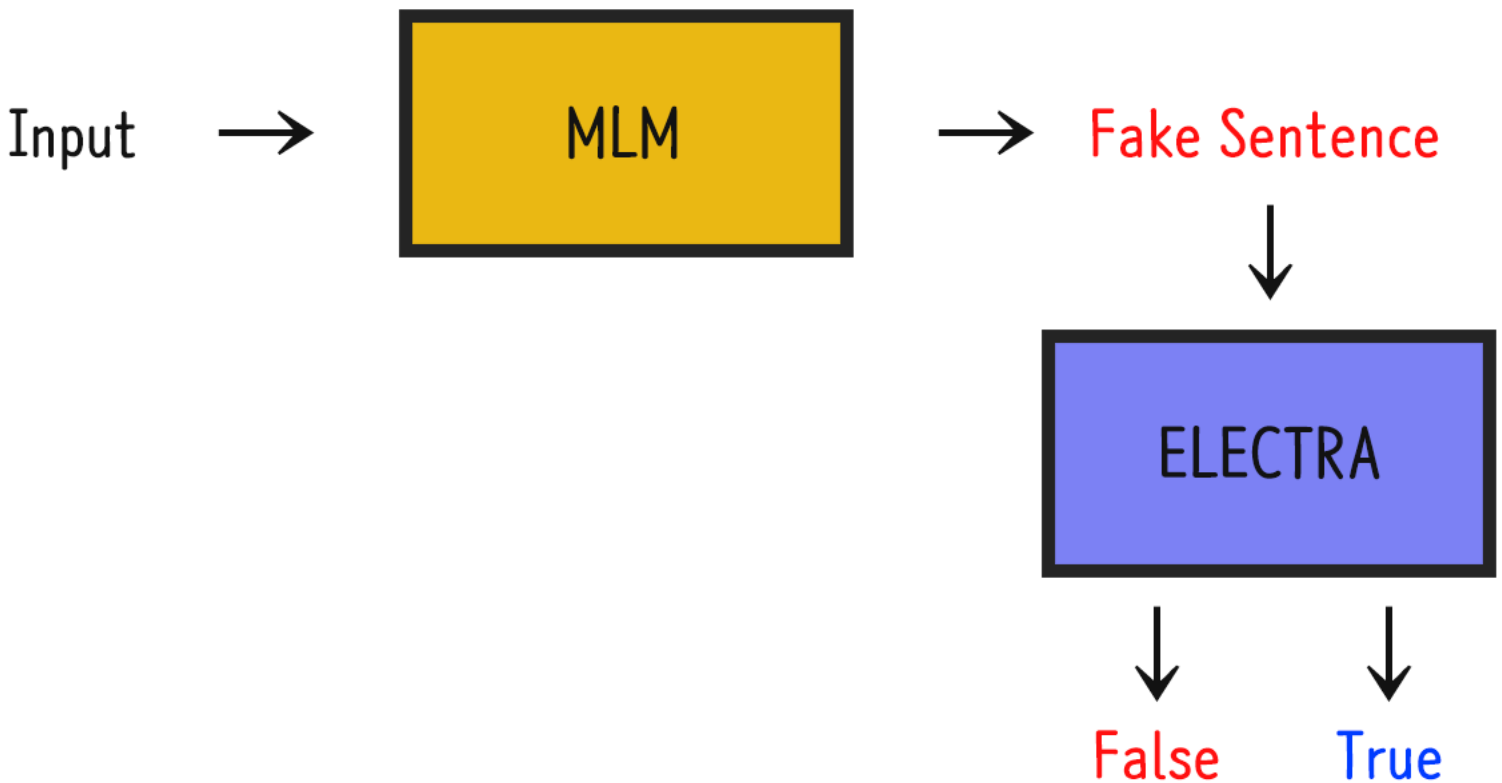

(부들부들)
Input
Generator
Fake Image
Real Image
Discriminator
Easy~
False
True

깨알상식 - GAN



서로의 실수를 유발,
적대적으로 동작하며 학습하므로
Adversarial Network

ELECTRA는 GAN?

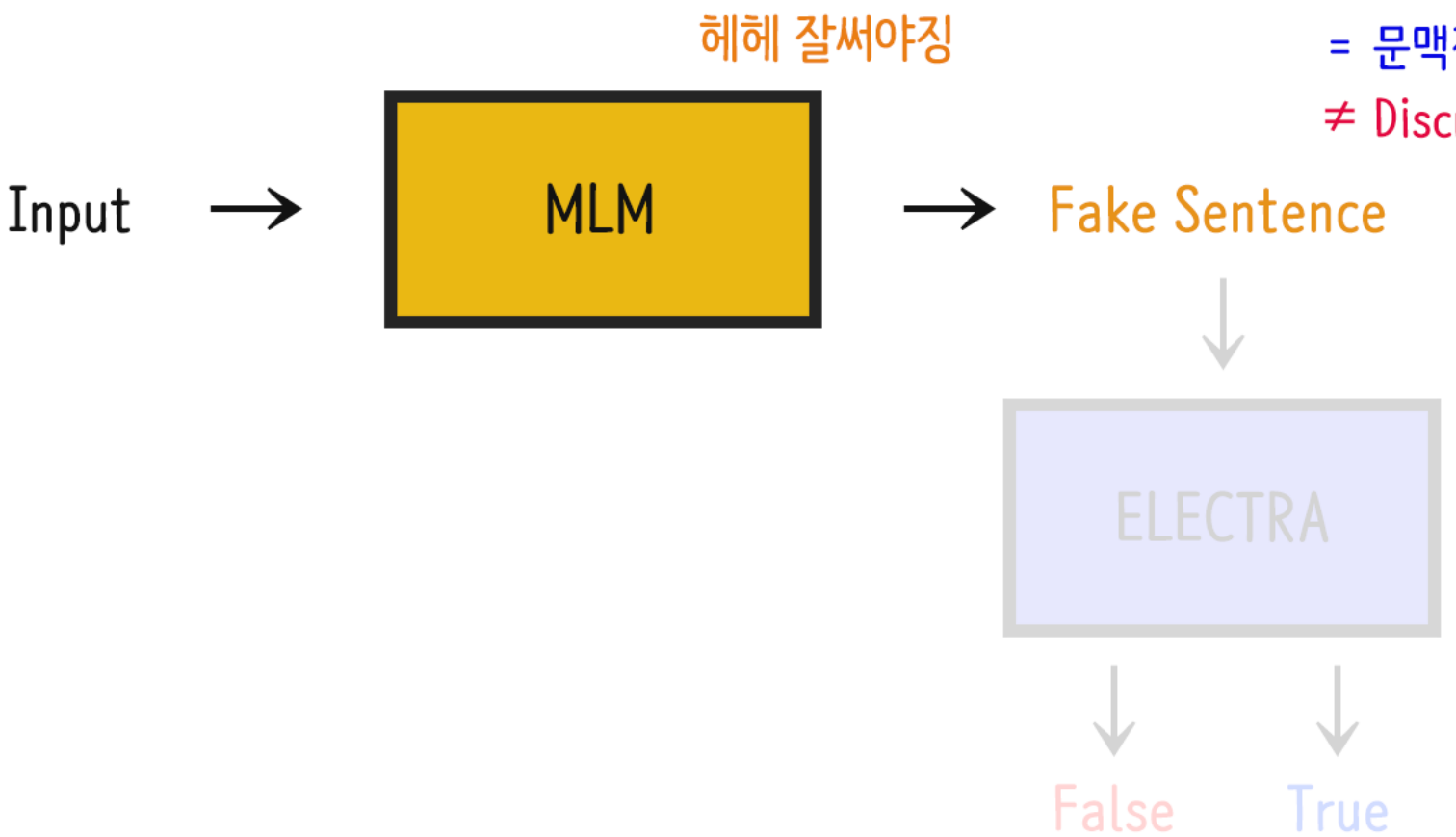


ELECTRA는 GAN?

Likelihood를 Maximize !!

= 문맥적으로 말이 되게

≠ Discriminator가 틀리게



ELECTRA는 GAN?

Input →

CAT 100%

Likelihood를 Maximize !!

= 문맥적으로 말이 되게

≠ Discriminator가 틀리게

헤헤 잘써야징

Fake Sentence

ELECTRA

False True

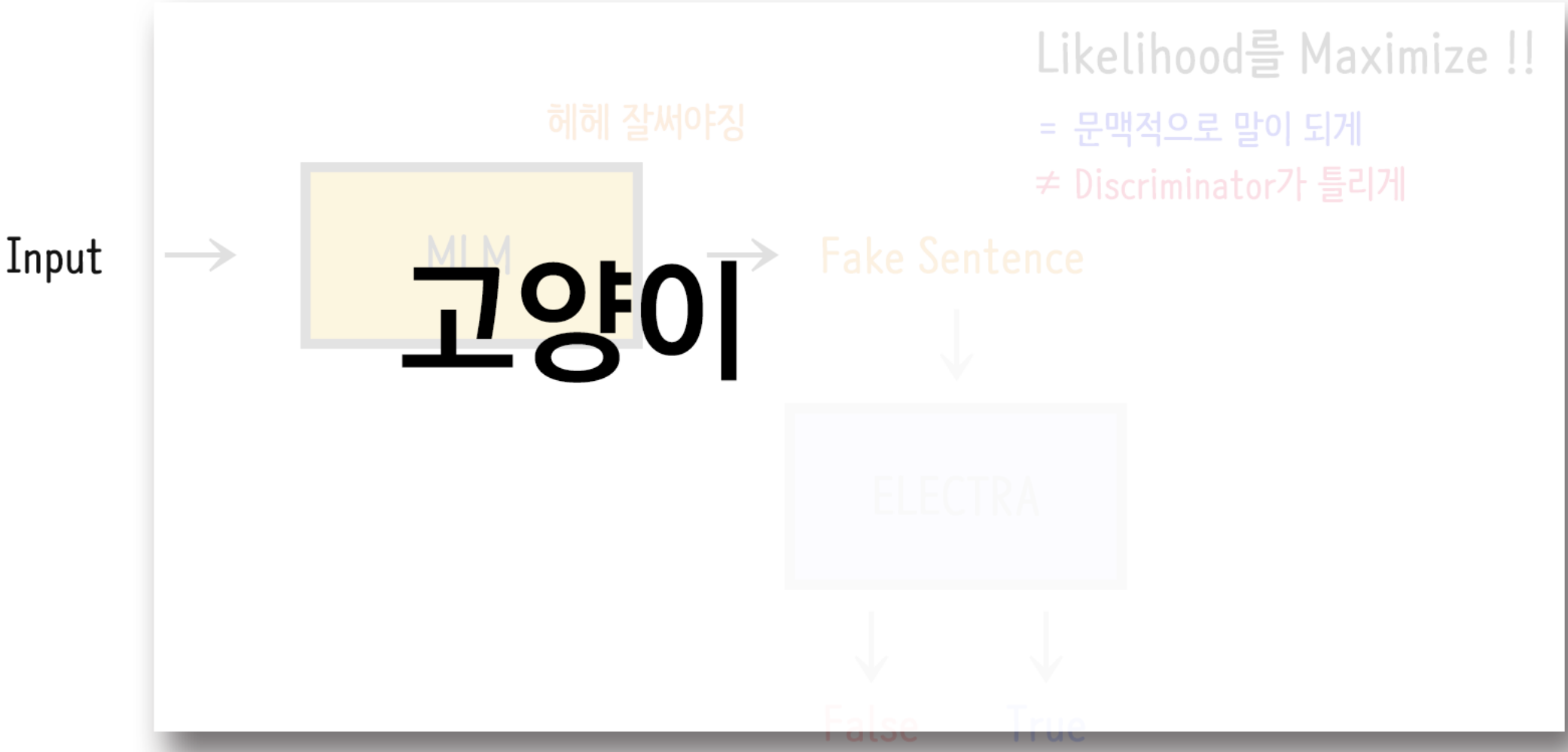
ELECTRA는 GAN?

Input →

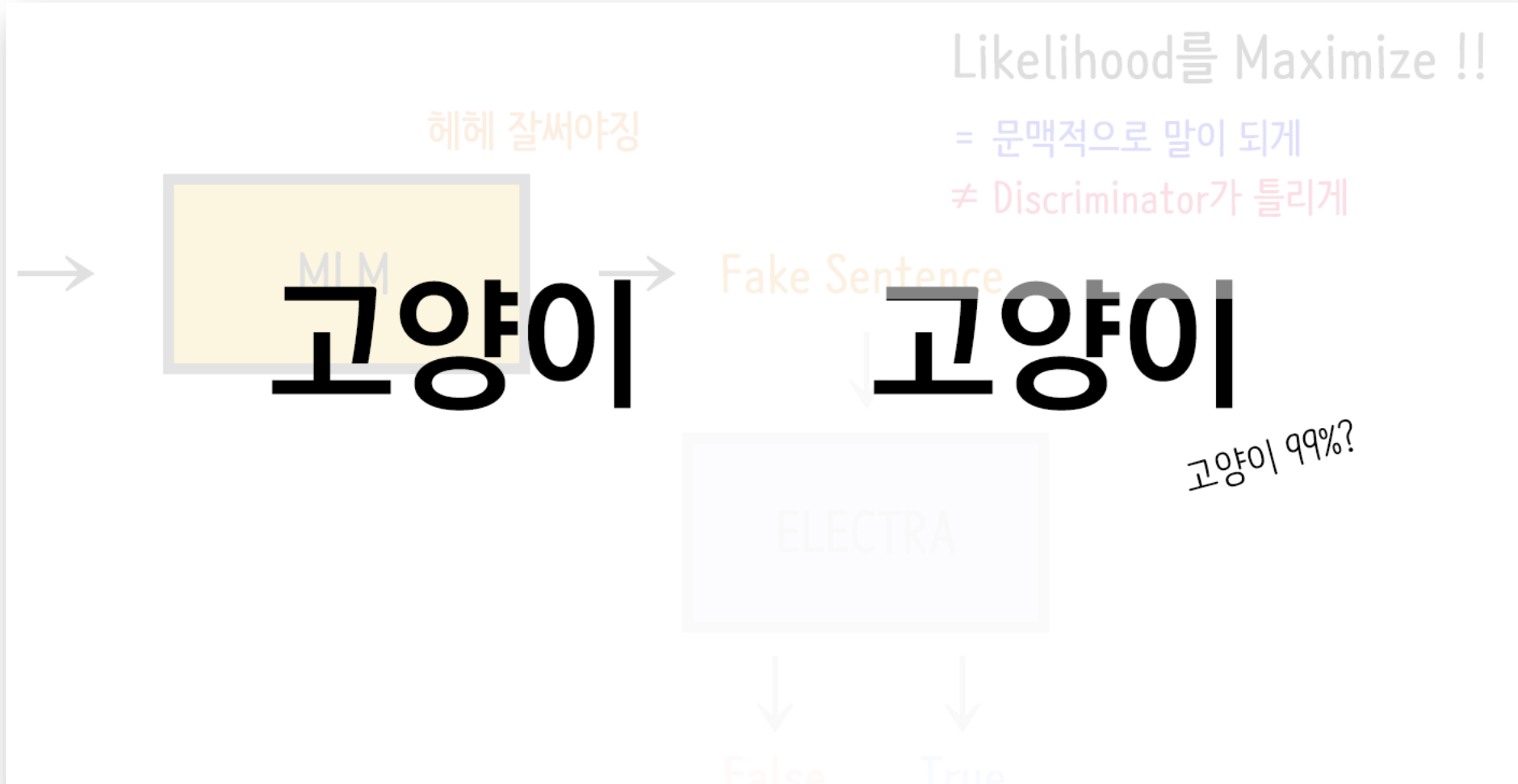
CAT 100%

CAT 35%...?

ELECTRA는 GAN?



ELECTRA는 GAN?



ELECTRA는 GAN?

Likelihood를 Maximize !!

= 문맥적으로 말이 되게

≠ Discriminator가 틀리게

헤헤 잘써야징

Input → MLM → 고양이 → Fake Sentence → 고양이

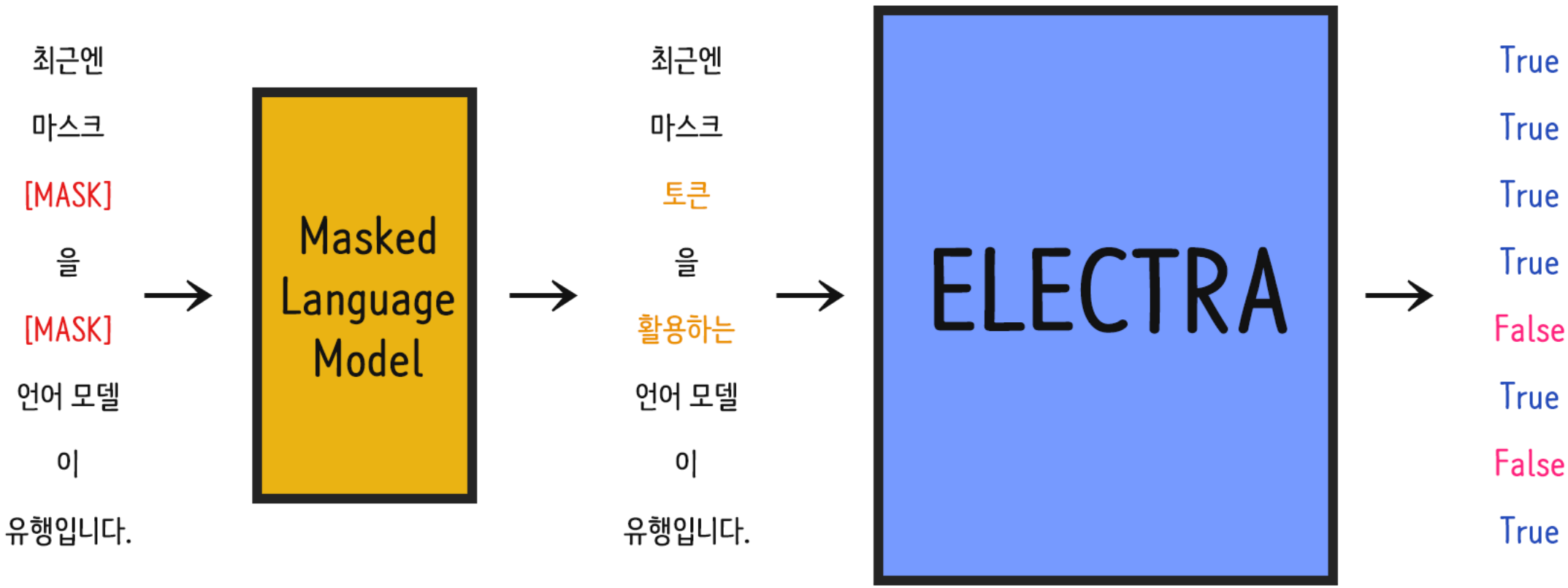
고양이 99%?

ELECTRA

고양이 + 0.1?

False True

Detail


최근엔
마스크
[MASK]
을
[MASK]
언어 모델
이
유행입니다.
Masked Language Model
최근엔
마스크
토큰
을
활용하는
언어 모델
이
유행입니다.
ELECTRA
True
True
True
True
False
True
False
True

Detail

최근엔

마스크

[MASK]

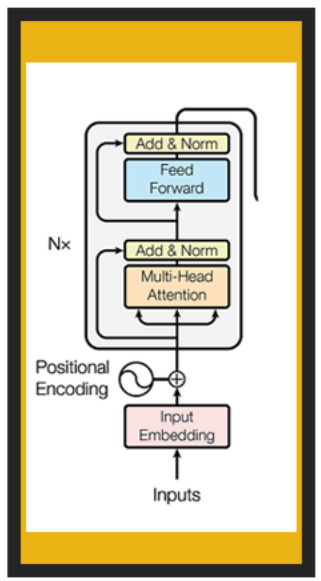
을

[MASK]

언어 모델

이

유행입니다.



최근엔

마스크

토큰

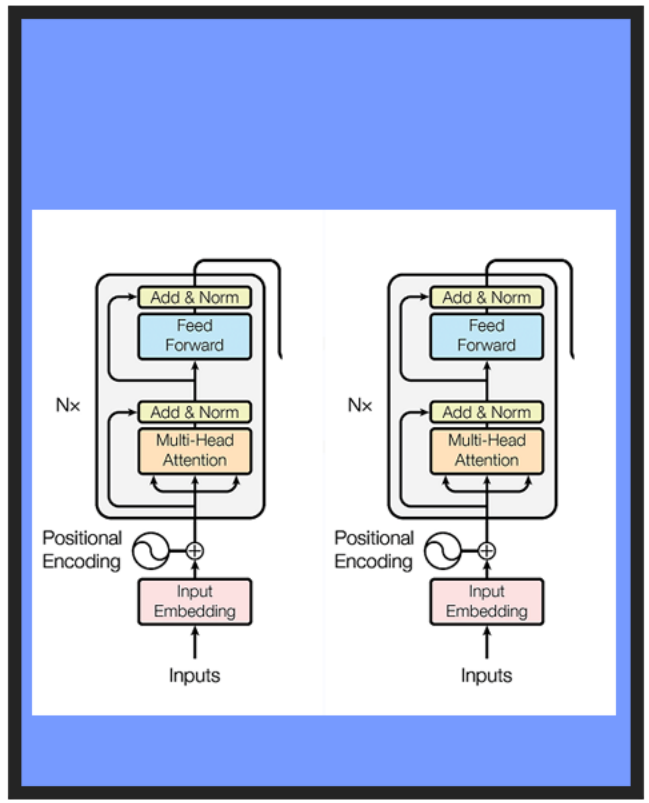
을

활용하는

언어 모델

이

유행입니다.



True

True

True

True

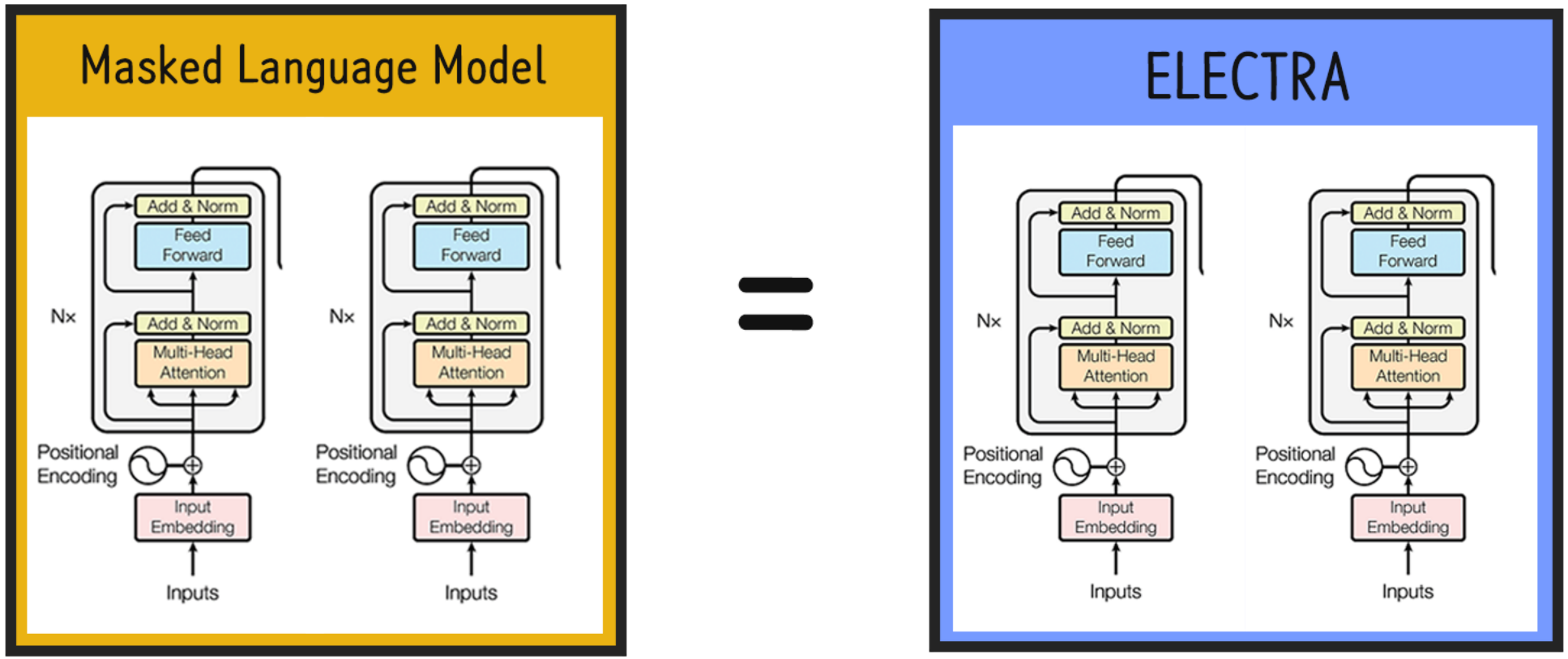
False

True

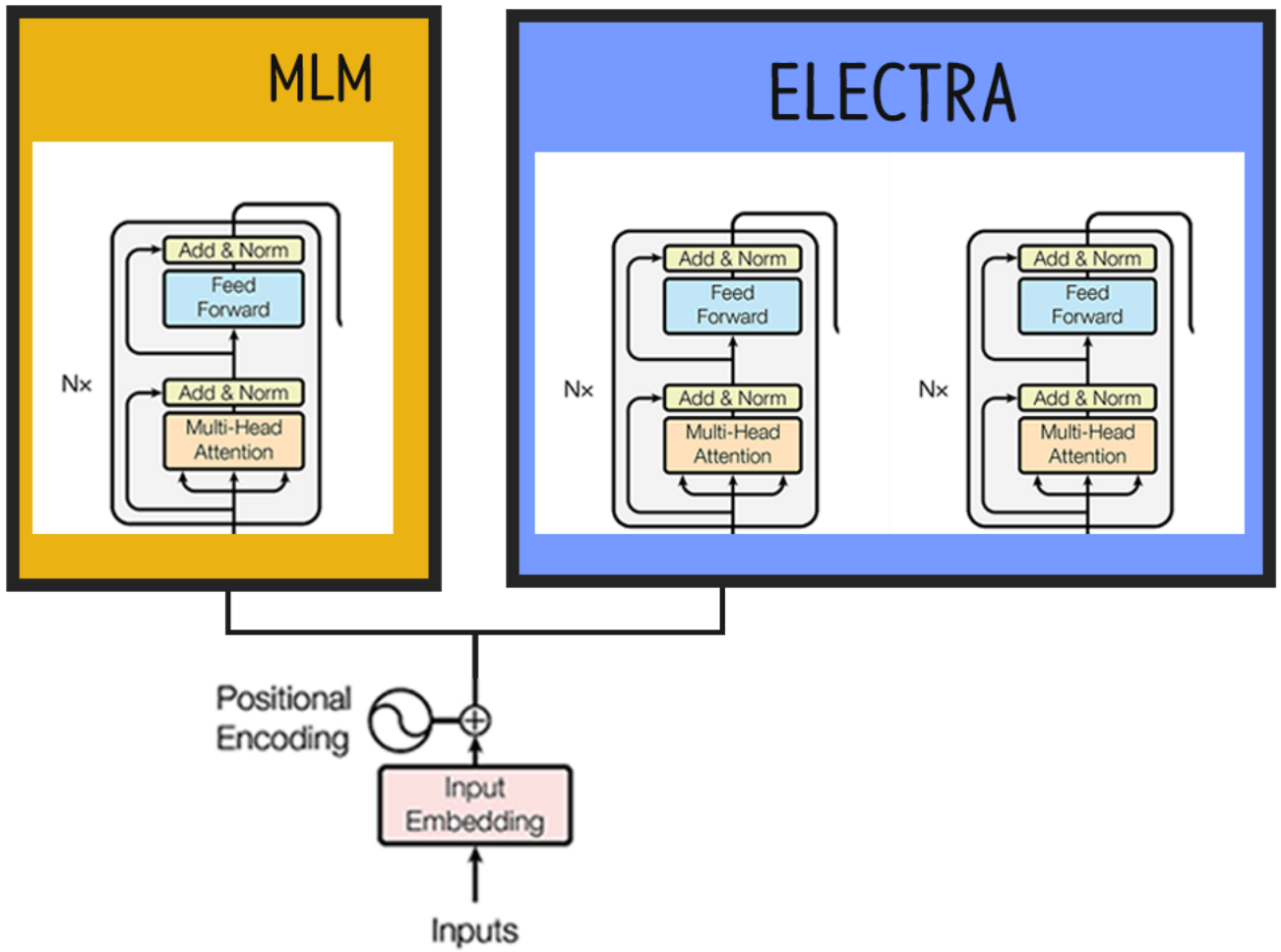
False

True

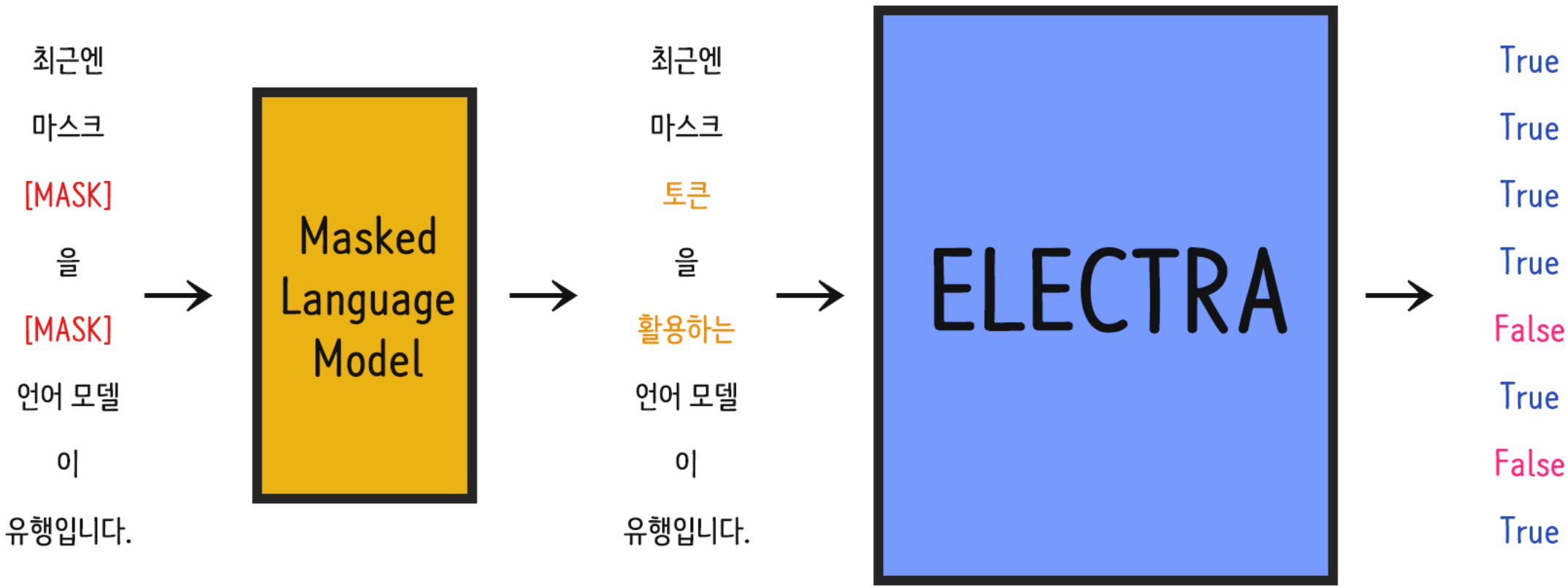
Weight Sharing



Weight Sharing (Only Embedding)

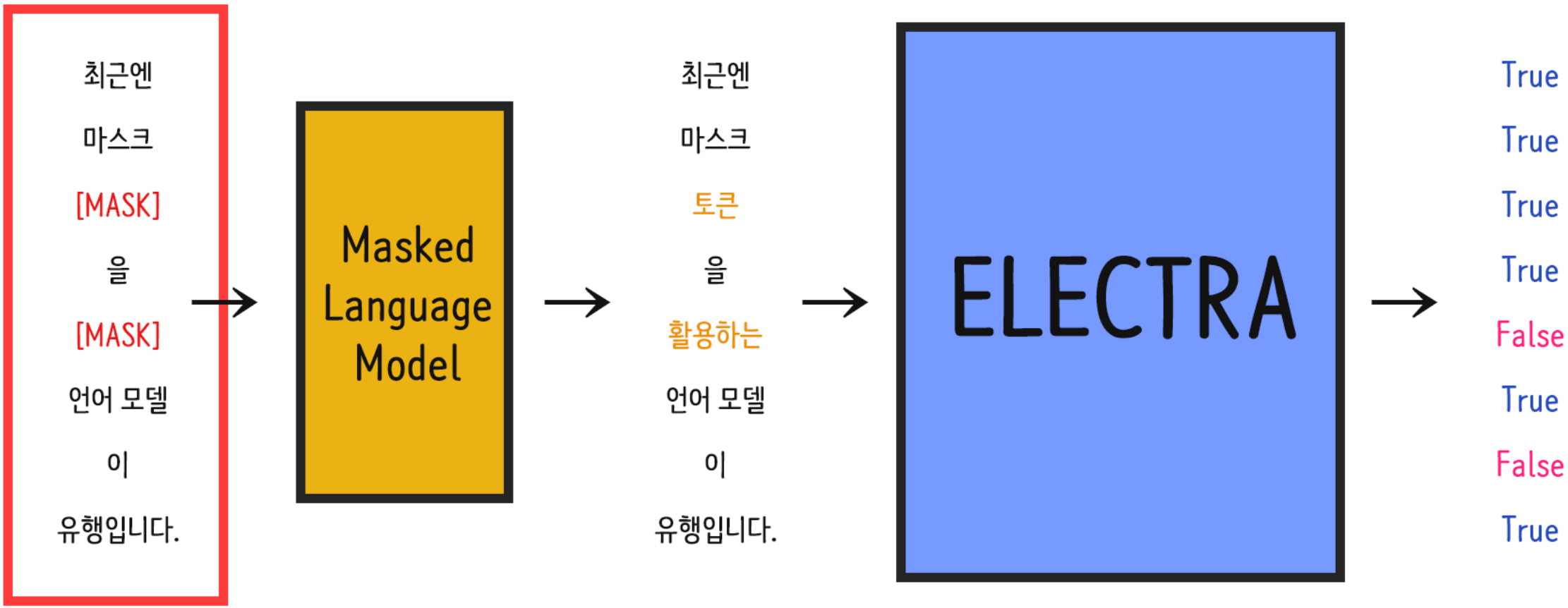


Detail

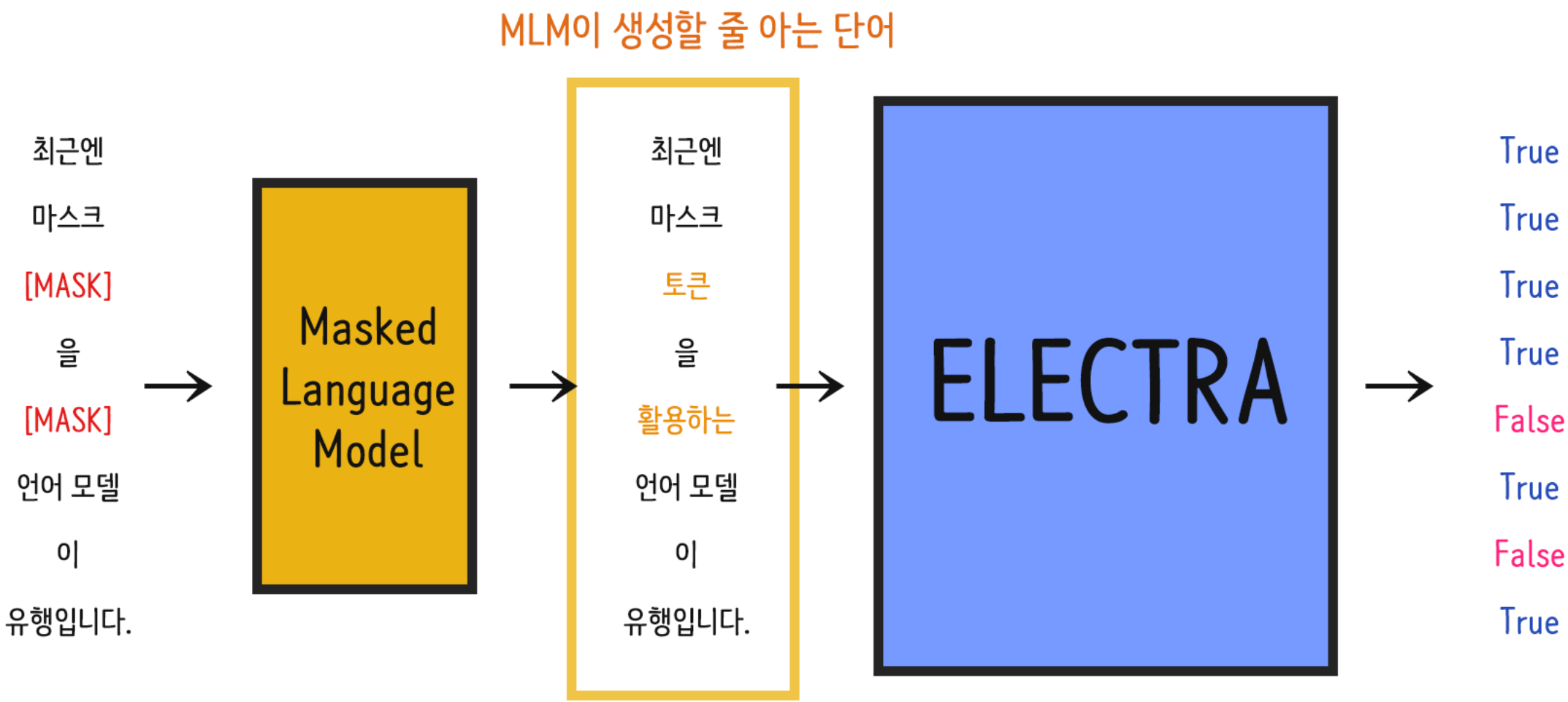

최근엔
마스크
[MASK]
을
[MASK]
언어 모델
이
유행입니다.
Masked Language Model
최근엔
마스크
토큰
을
활용하는
언어 모델
이
유행입니다.
ELECTRA
True
True
True
True
False
True
False
True

Detail

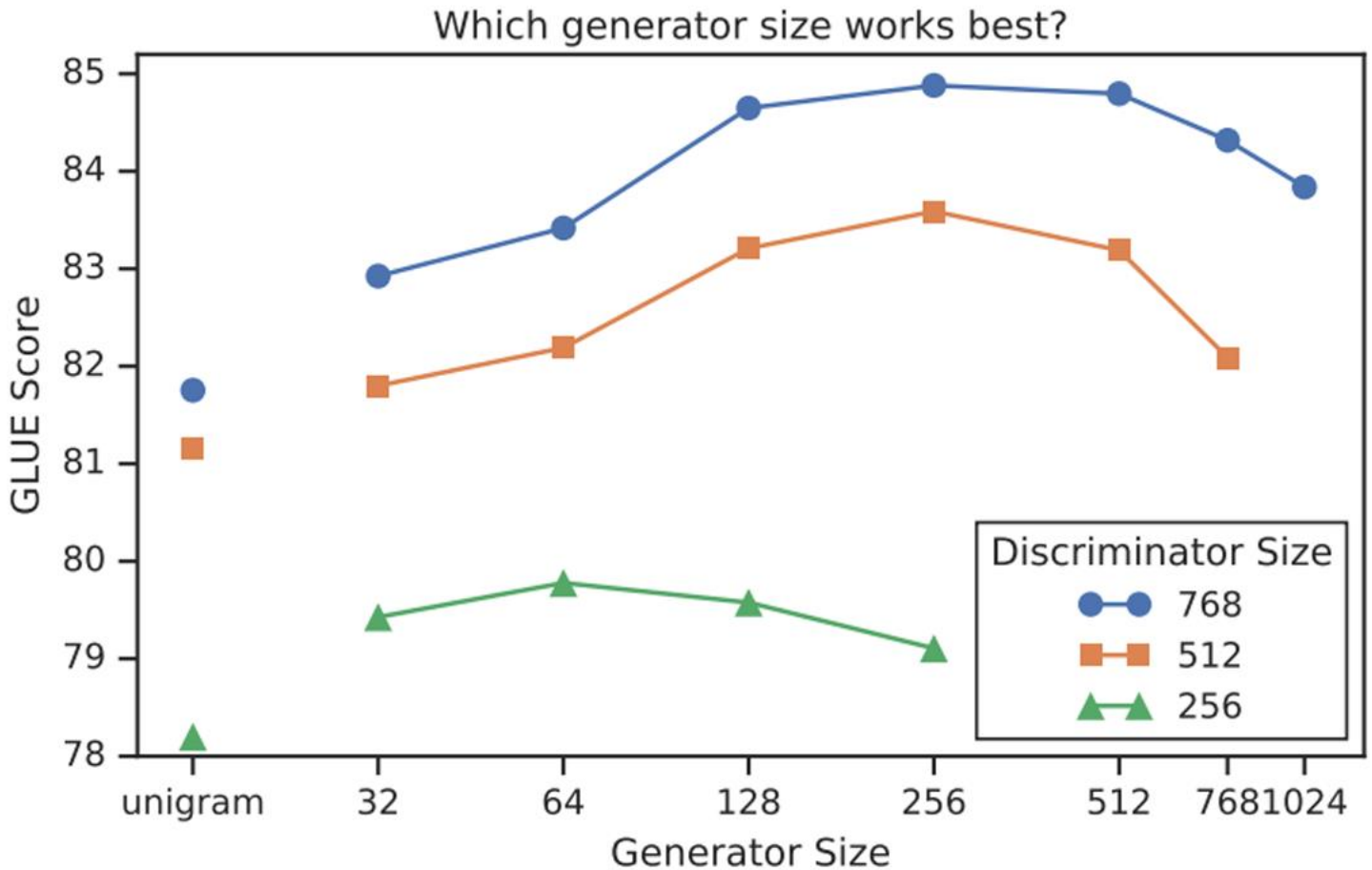
자연 (데이터셋) 에 존재하는 모든 단어



Detail


MLM이 생성할 줄 아는 단어
최근엔
마스크
[MASK]
을
[MASK]
언어 모델
이
유행입니다.
Masked Language Model
최근엔
마스크
토큰
을
활용하는
언어 모델
이
유행입니다.
ELECTRA
True
True
True
True
False
True
False
True

Smaller Generators



Result

Model	Train FLOPs	Params	SQuAD 1.1 dev		SQuAD 2.0 dev		SQuAD 2.0 test	
			EM	F1	EM	F1	EM	F1
BERT-Base	6.4e19 (0.09x)	110M	80.8	88.5	–	–	–	–
BERT	1.9e20 (0.27x)	335M	84.1	90.9	79.0	81.8	80.0	83.0
SpanBERT	7.1e20 (1x)	335M	88.8	94.6	85.7	88.7	85.7	88.7
XLNet-Base	6.6e19 (0.09x)	117M	81.3	–	78.5	–	–	–
XLNet	3.9e21 (5.4x)	360M	**89.7**	**95.1**	87.9	**90.6**	87.9	90.7
RoBERTa-100K	6.4e20 (0.90x)	356M	–	94.0	–	87.7	–	–
RoBERTa-500K	3.2e21 (4.5x)	356M	88.9	94.6	86.5	89.4	86.8	89.8
ALBERT	3.1e22 (44x)	235M	89.3	94.8	87.4	90.2	88.1	90.9
BERT (ours)	7.1e20 (1x)	335M	88.0	93.7	84.7	87.5	–	–
ELECTRA-Base	6.4e19 (0.09x)	110M	84.5	90.8	80.5	83.3	–	–
ELECTRA-400K	7.1e20 (1x)	335M	88.7	94.2	86.9	89.6	–	–
ELECTRA-1.75M	3.1e21 (4.4x)	335M	**89.7**	94.9	**88.0**	**90.6**	**88.7**	**91.4**

Table 4: Results on the SQuAD for non-ensemble models.

THANK YOU