BERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.
- Use the Transformer Encoder architecture.
- Introduced in 2018 by Google AI.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: pretraining of deep bidirectional transformers for language understanding* (J. Burstein, C. Doran, & T. Solorio, Eds.).

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

Dataset:

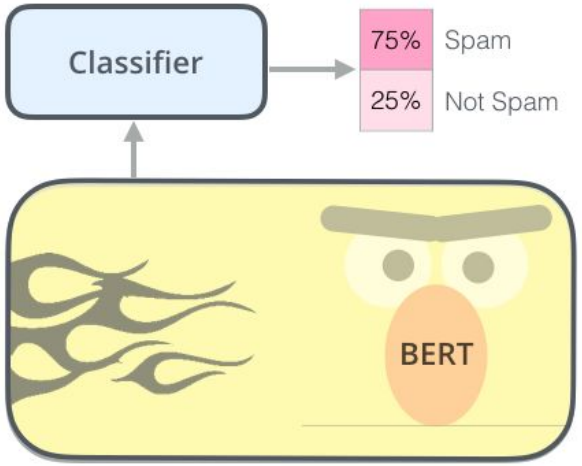
Objective:

Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

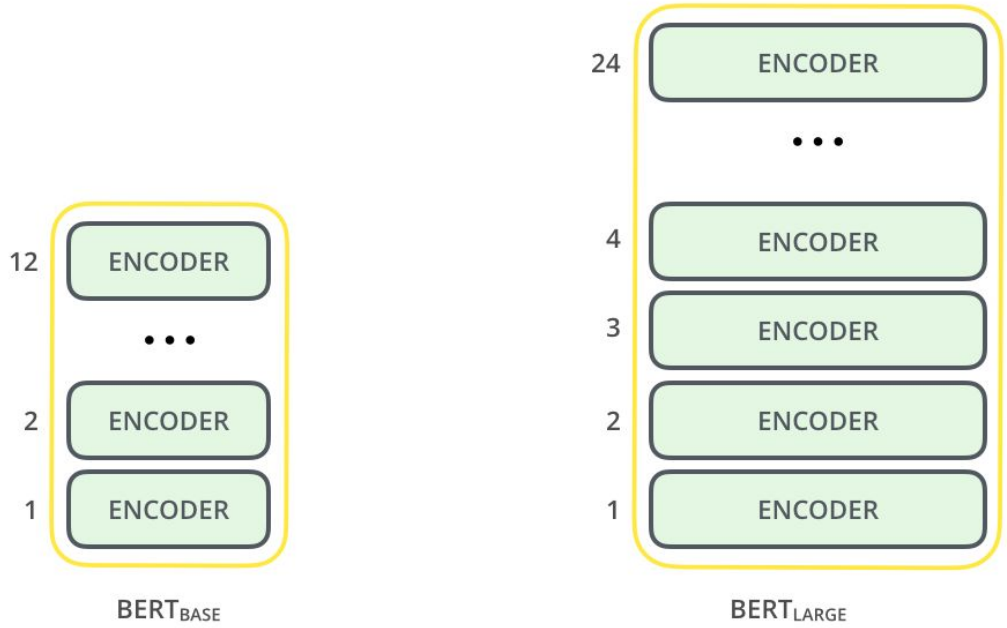
Model: (pre-trained in step #1)



Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

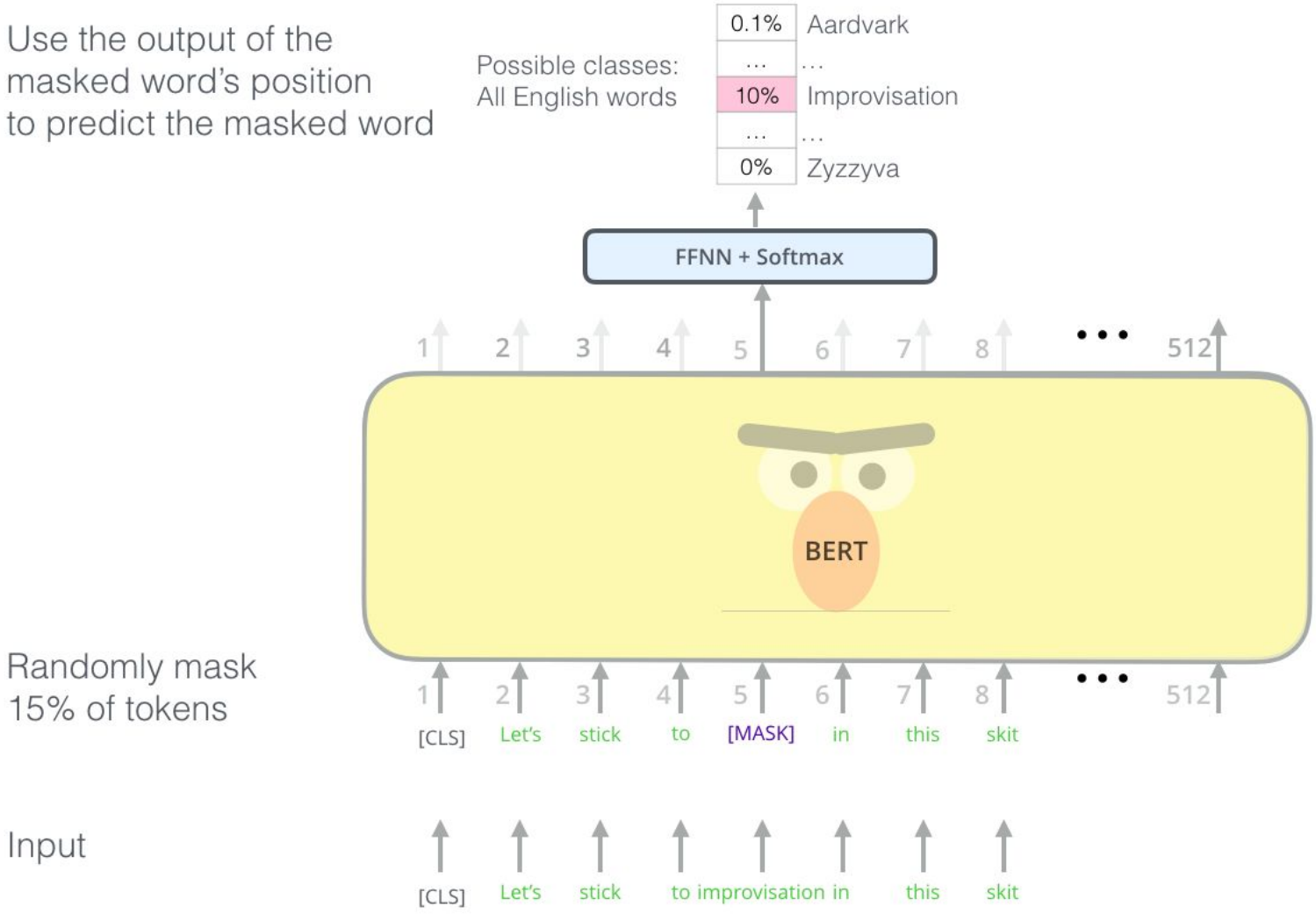
Architecture

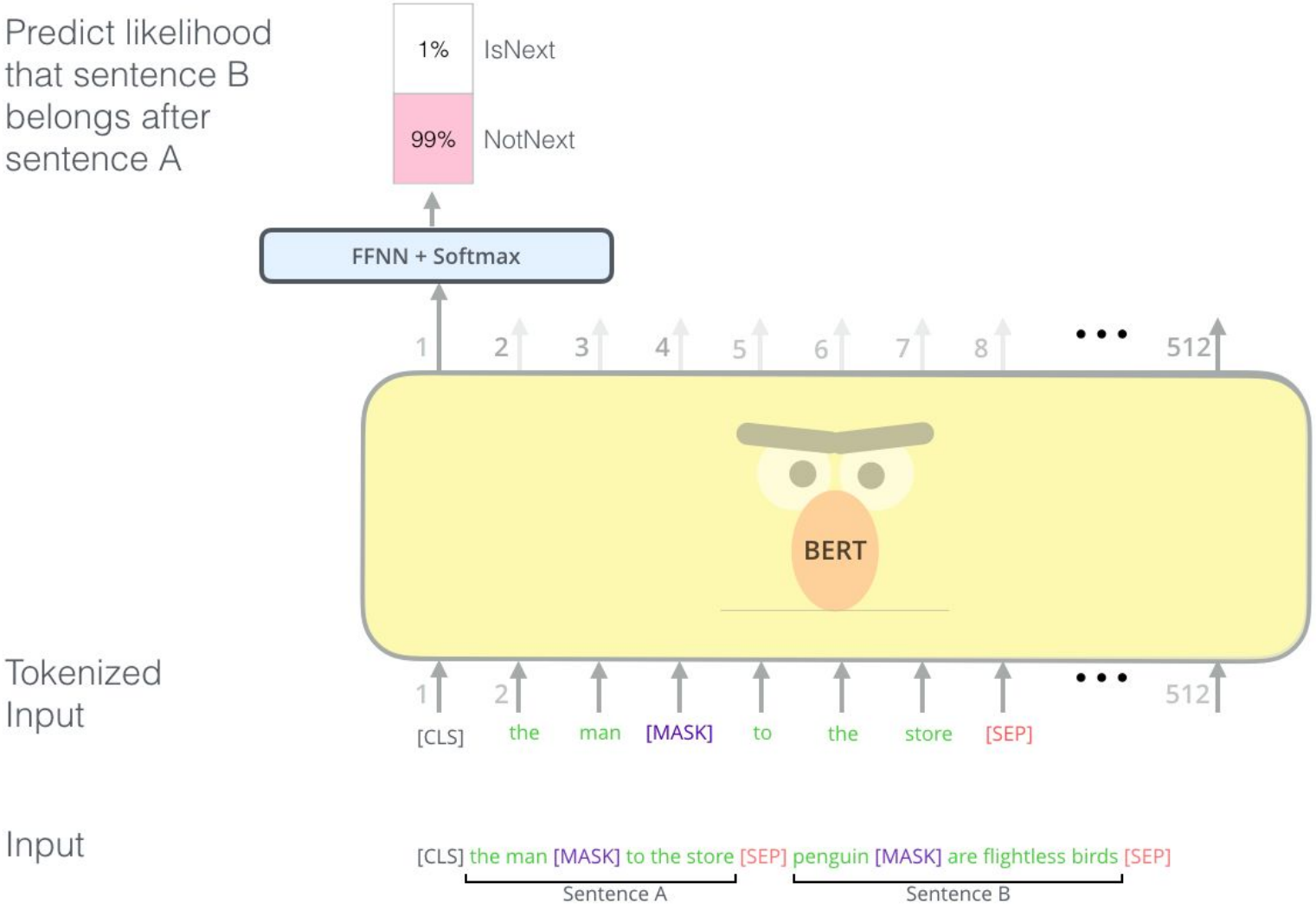


Pretraining

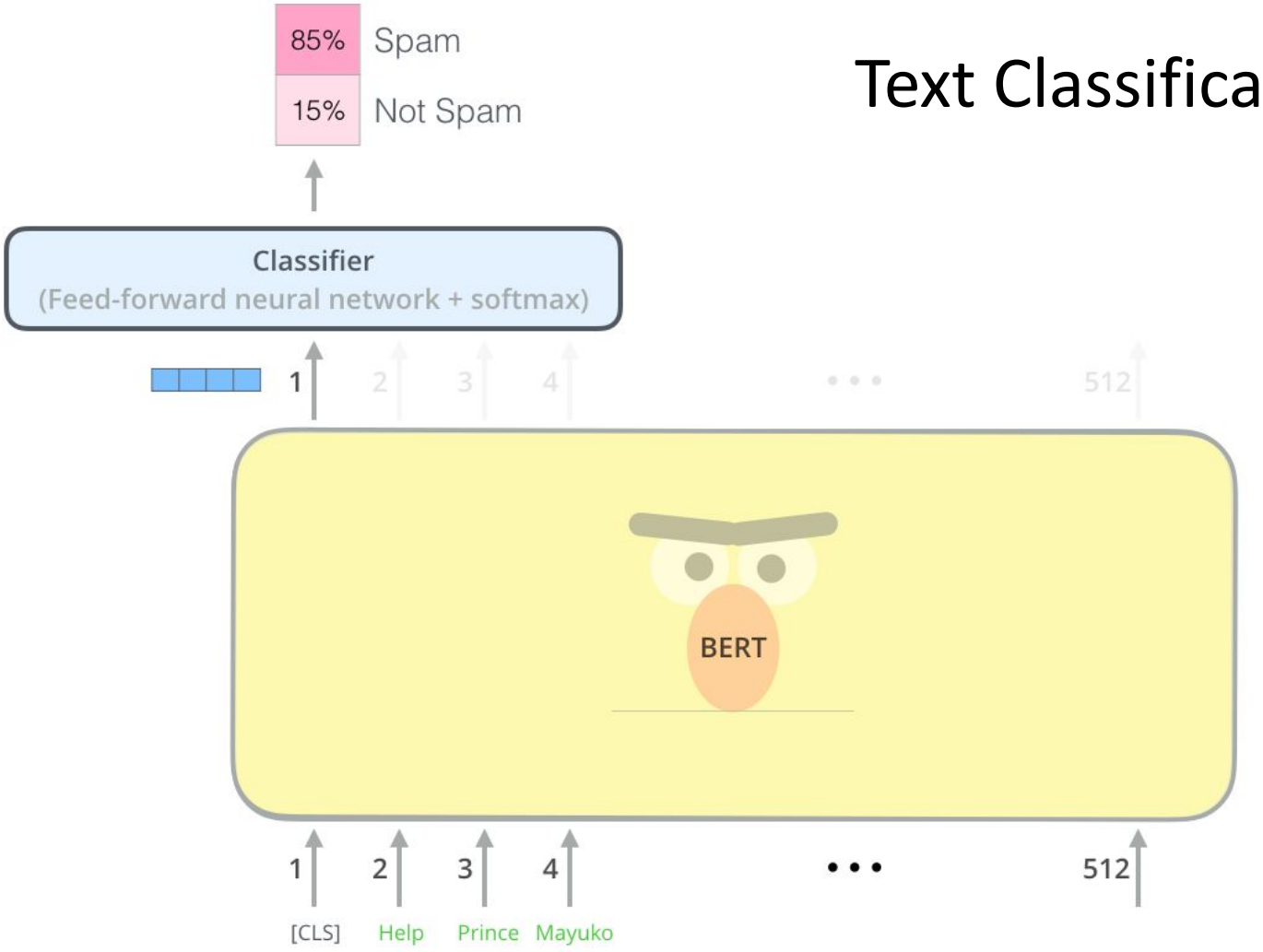
- Two unsupervised tasks:
  1. Masked Language Model
  2. Next Sentence Prediction

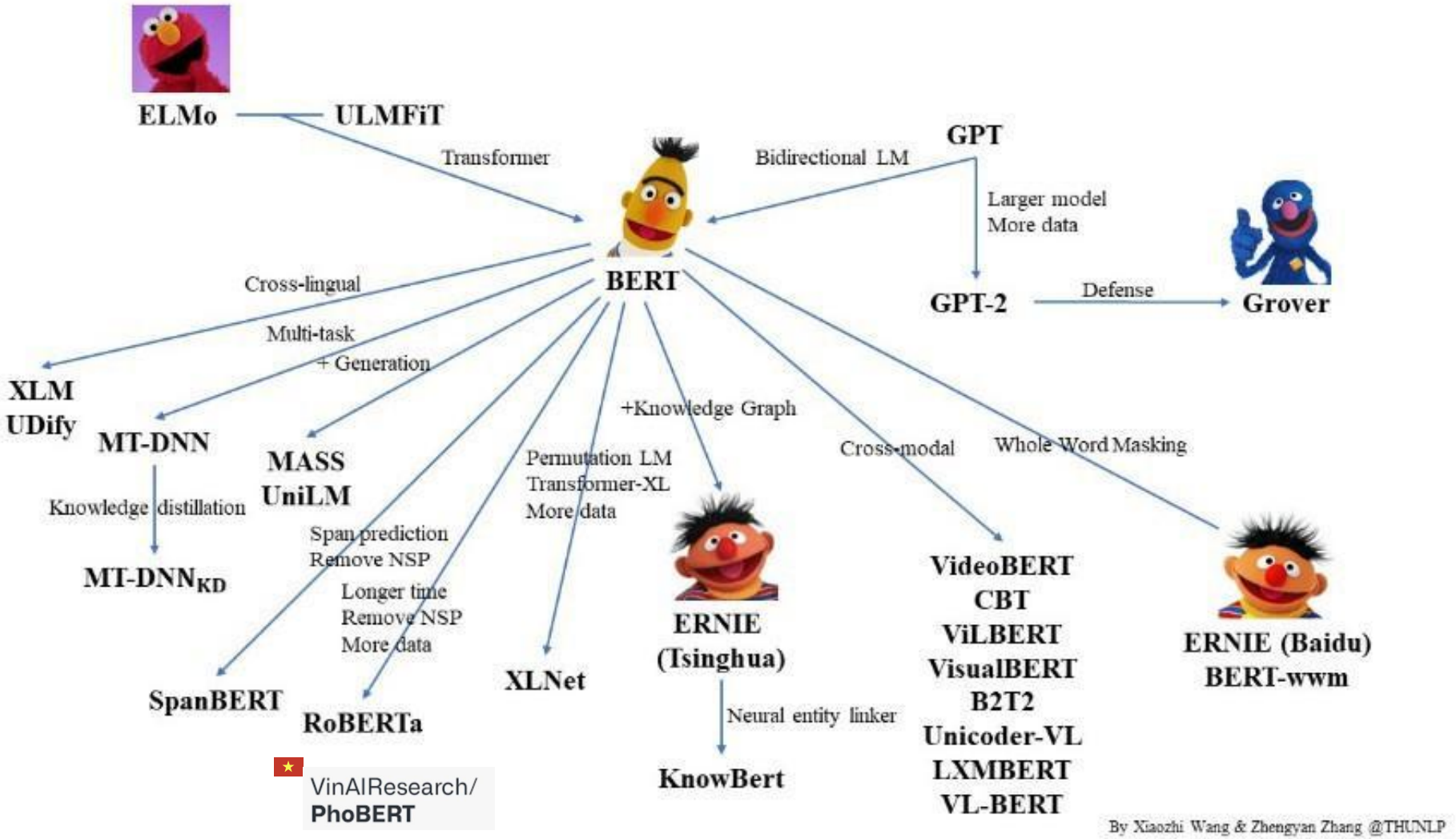
Use the output of the masked word's position to predict the masked word




Predict likelihood that sentence B belongs after sentence A
1%
IsNext
99%
NotNext
FFNN + Softmax
1
2
3
4
5
6
7
8
512
BERT
Tokenized Input
[CLS]
the
man
[MASK]
to
the
store
[SEP]
Input
[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]
Sentence A
Sentence B

Text Classification




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

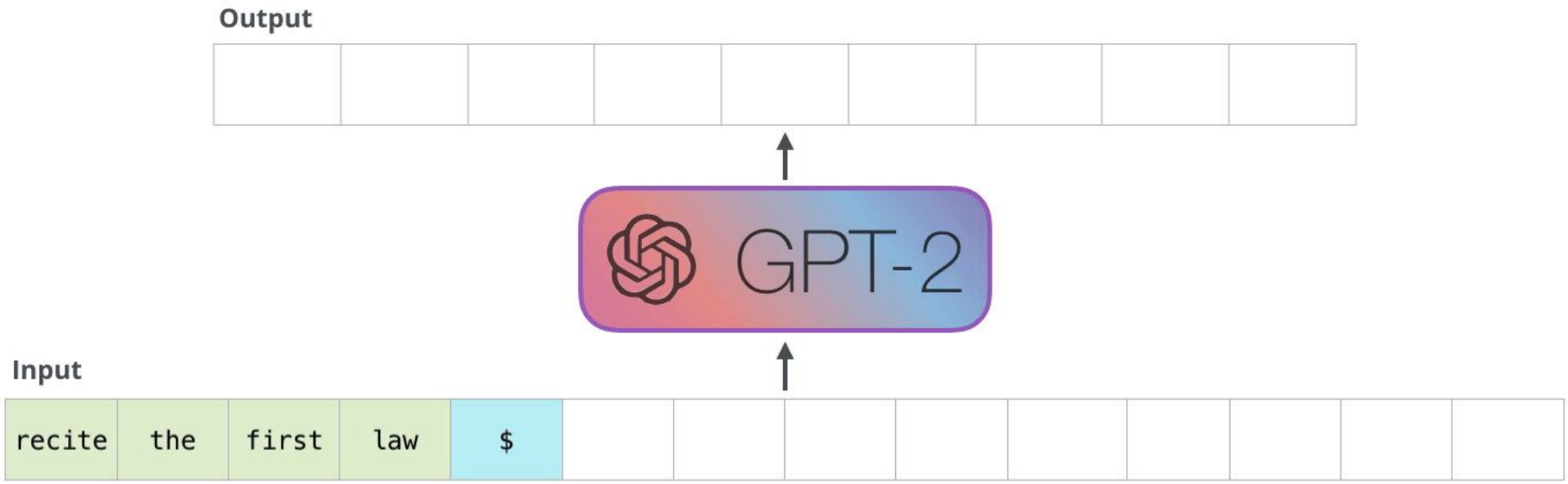
GPT

- Generative Pre-trained Transformer
- Use the Transformer Decoder architecture.
- Introduced in 2018 by OpenAI.

Model	Number of parameters	Training data size	Year
GPT	110M	4GB	2018
GPT-2	1.5B	40GB	2019
GPT-3	175B	≈2TB	2020

Openai [Accessed: 2023-03-01]. (2023). https://openai.com/

How it works?



XLNet

- Autoencoding (BERT):
  - [MASK] tokens do not appear during finetuning ⇒ pretrain-finetuning discrepancy.
  - Assume the predicted tokens are independent of each other given the unmasked tokens. Example: “New York is a city” ⇒ “[MASK] [MASK] is a city”
- Autoregressive (GPT):
  - Only trained to encode a unidirectional context (forward or backward).

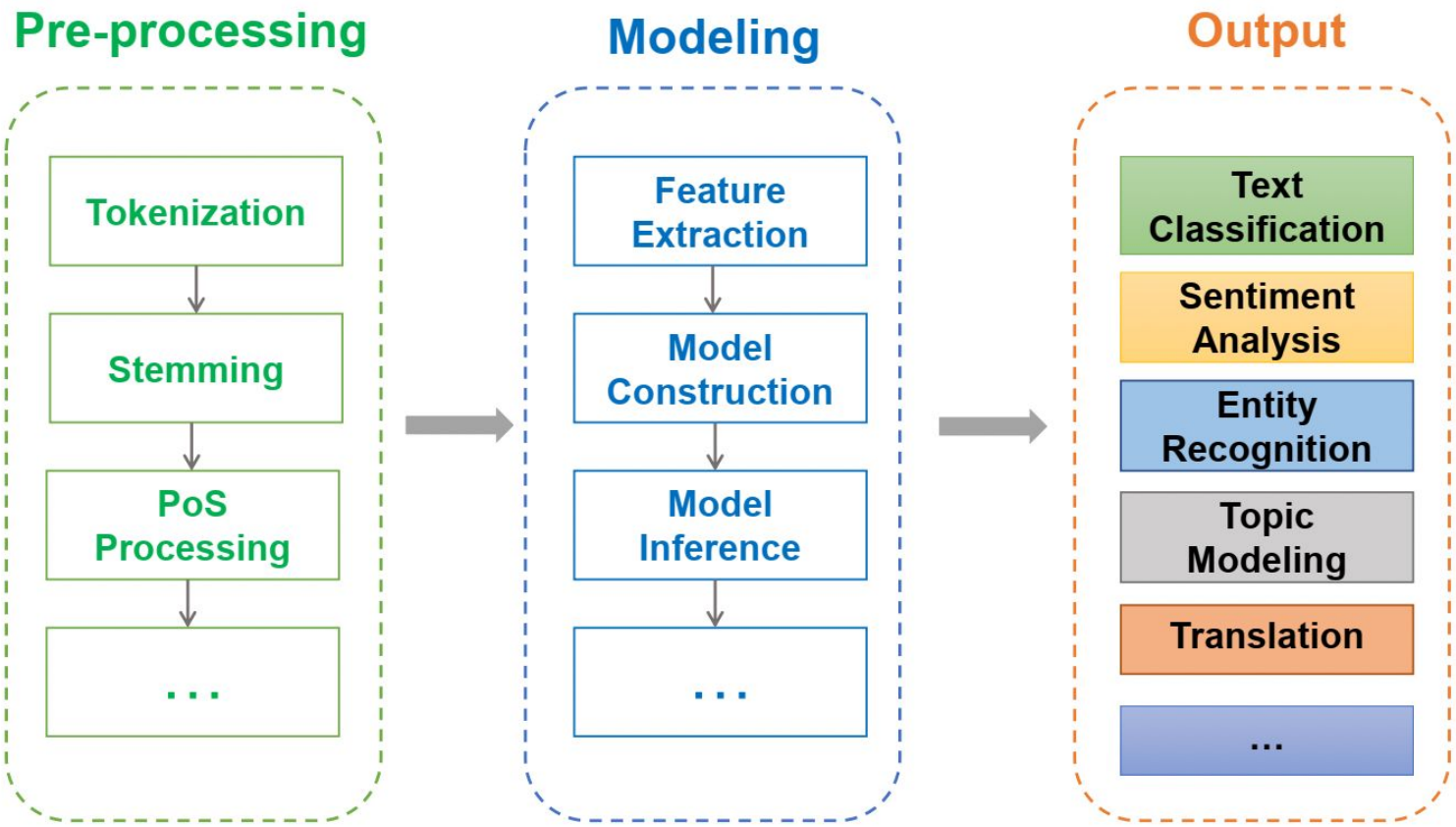
Yang, Z. et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding.” *NeurIPS* (2019)

XLNet

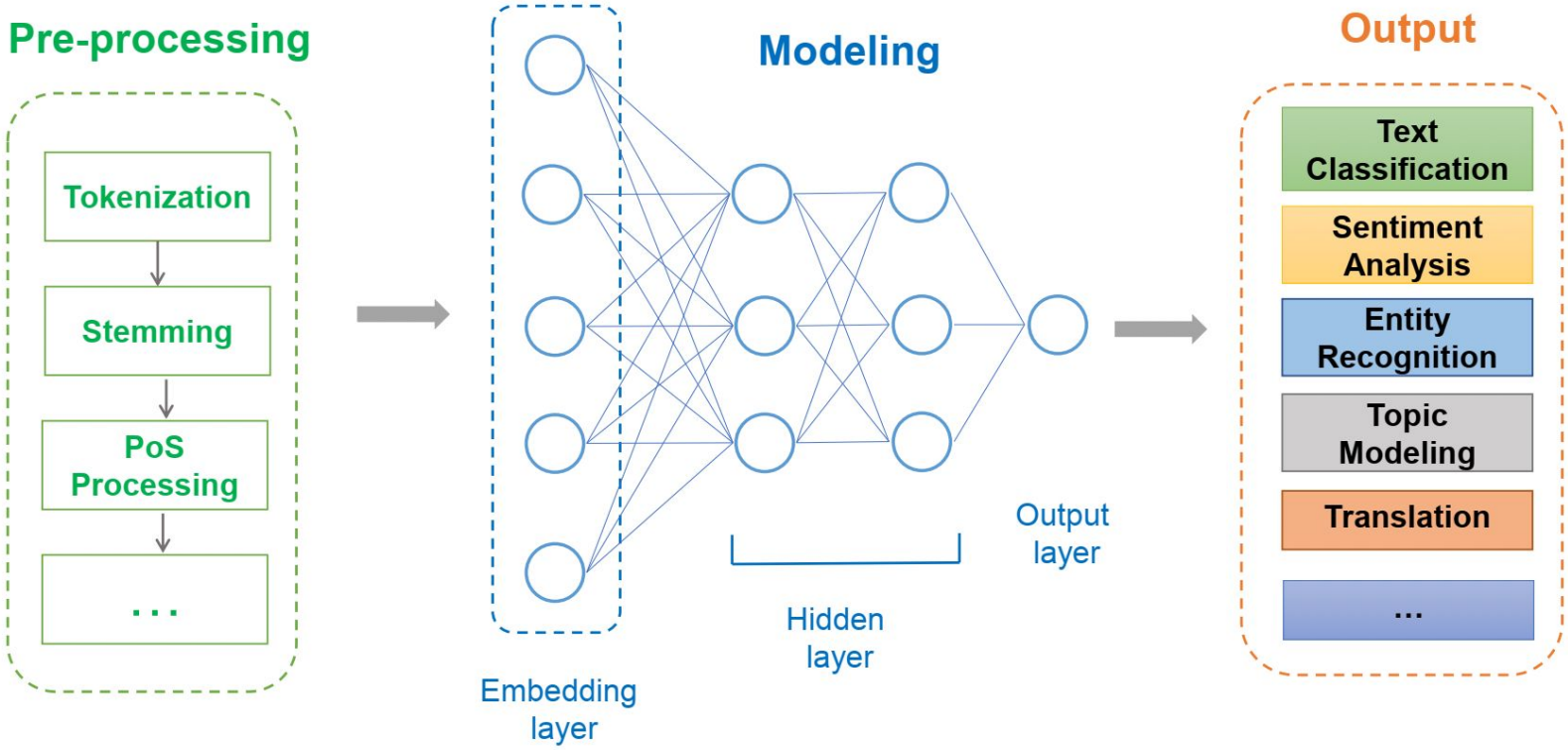
- XLNet combines pros from both while avoiding their cons.
- Techniques:
  - Permutation Language Modeling
  - Two-Stream Self-Attention for Target-Aware Representations
  - Incorporating Ideas from Transformer-XL
  - Modeling Multiple Segments

Applications in NLP

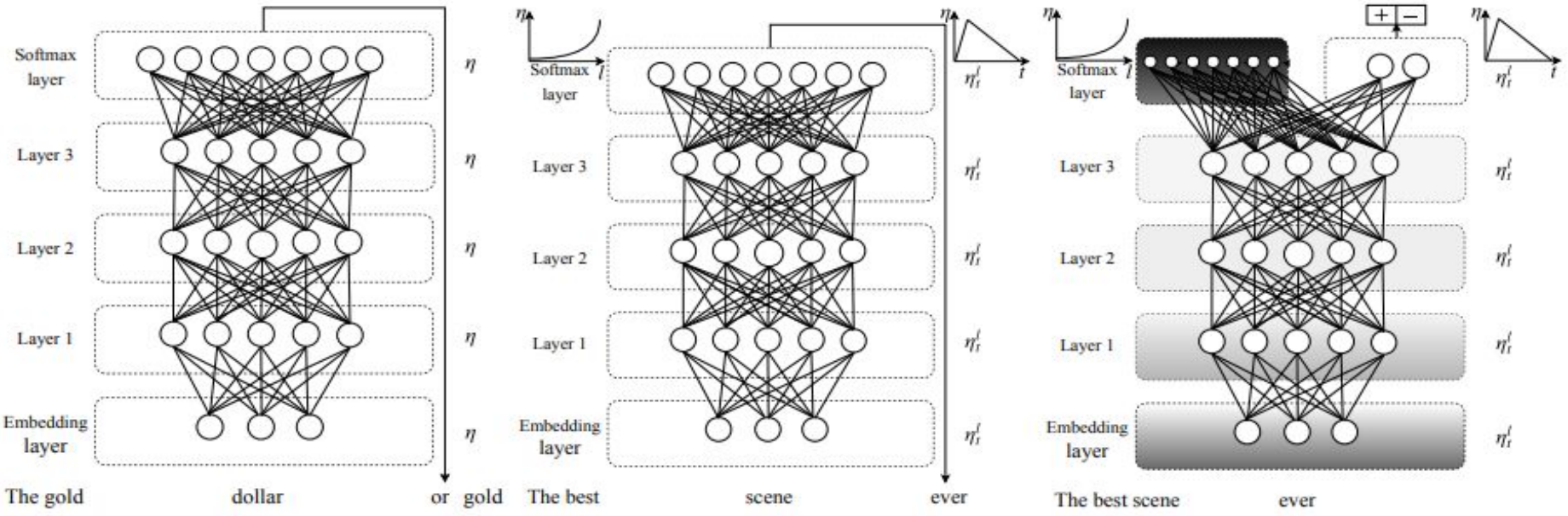
NLP typical pipeline



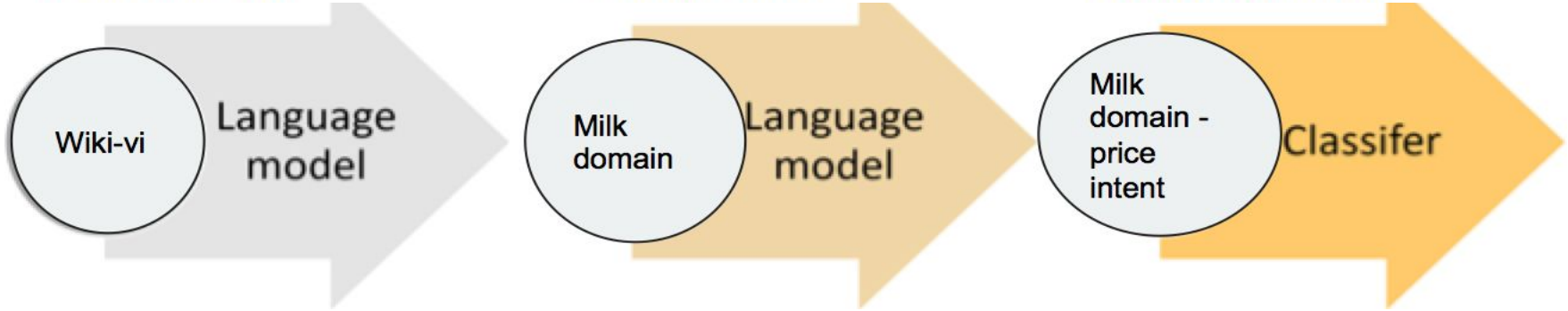
NLP DL-based pipeline



Pre-trained Neural Language Model

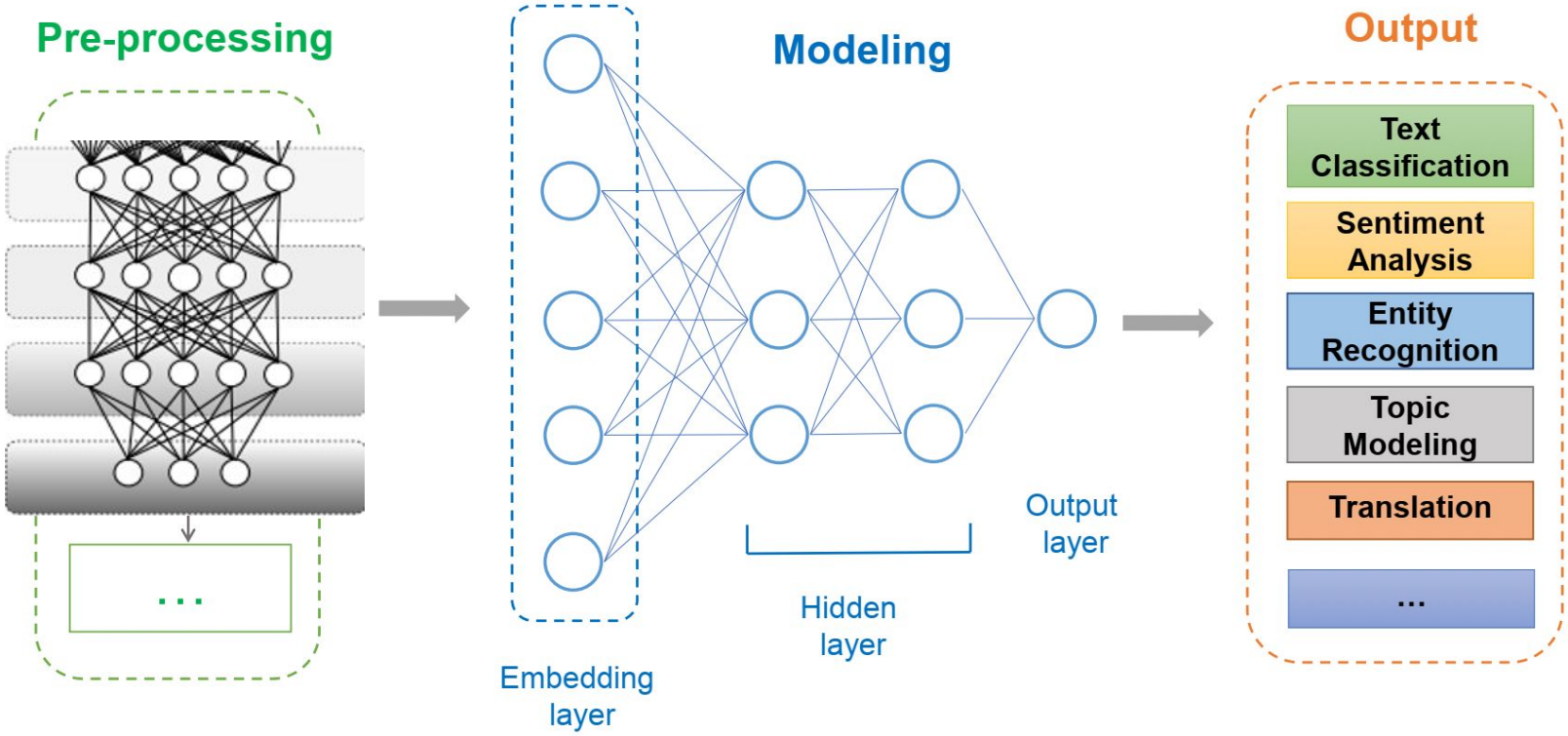


(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

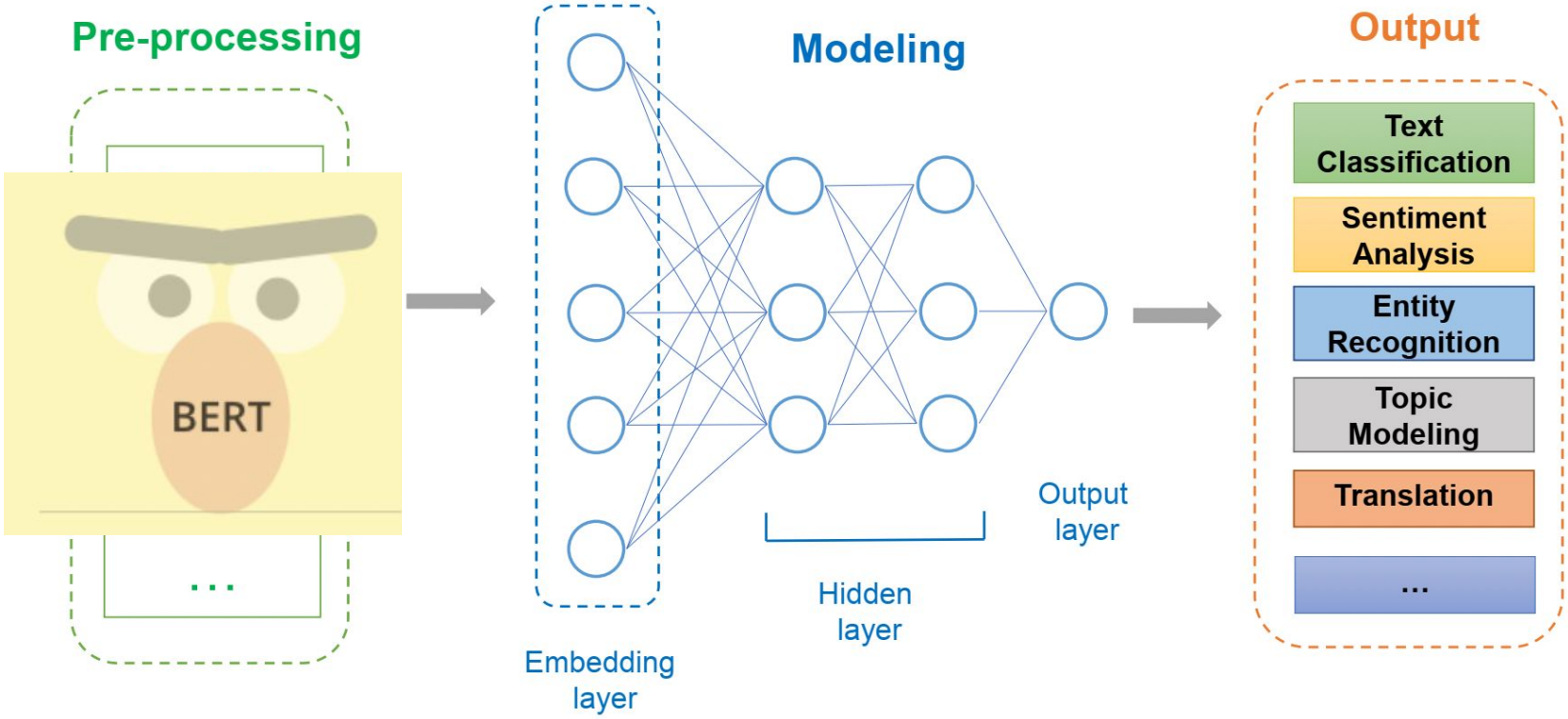


ULMFit (Howard and Rudder, 2018)

NLP LM-based pipeline

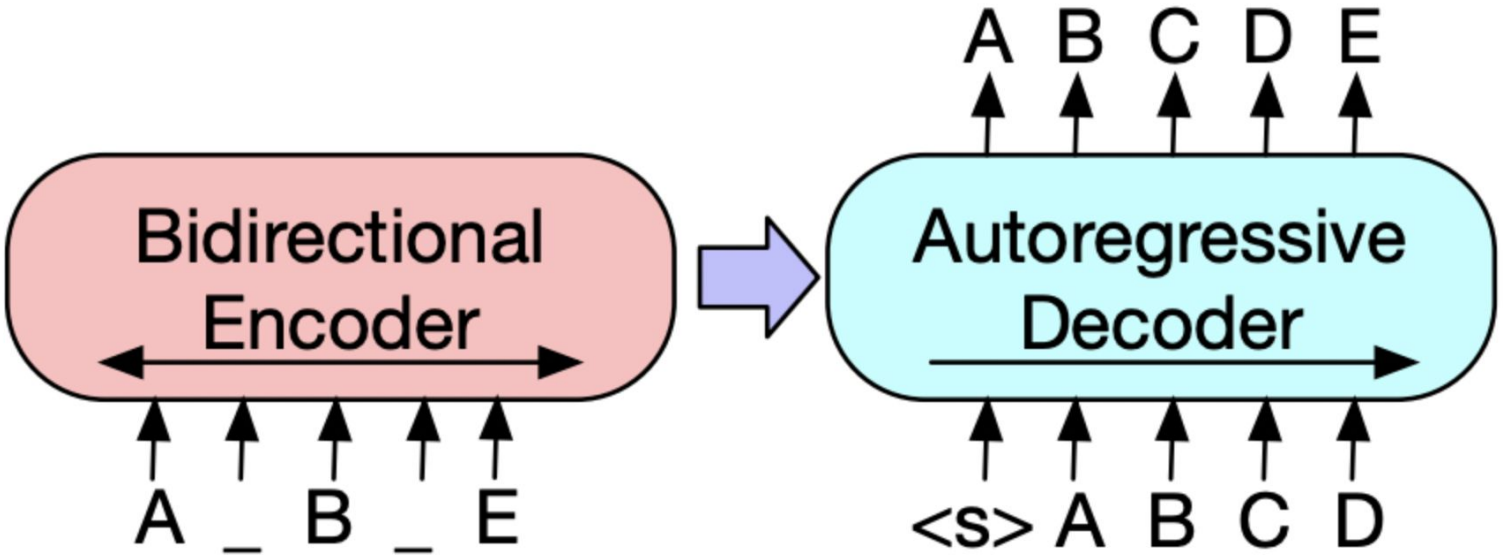


NLP LM-based pipeline



From BERT to BART

- BERT is not a fully Seq2Seq model (i.e. not a generative model)
- BART is introduced as an extended/complement



Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

From PhoBERT to BARTPho

VinAIResearch/
BARTpho

BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese (INTERSPEECH 2022)

1 Contributor | 0 Issues | 75 Stars | 6 Forks

BARTPho for Vietnamese translation applications

- Pretrained with Vietnamese
- Implicitly processing “aligning” task
- More powerful if the target language has similar language to Vietnamese (Chinese, Bahnaric, etc.)