BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Presented by Team MediAssist:

Yash Pankhania
Utkarsha Shirke
Shreya Jaiswal

INTRODUCTION

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin

Kenton Lee

Kristina Toutanova

Ming-Wei Chang

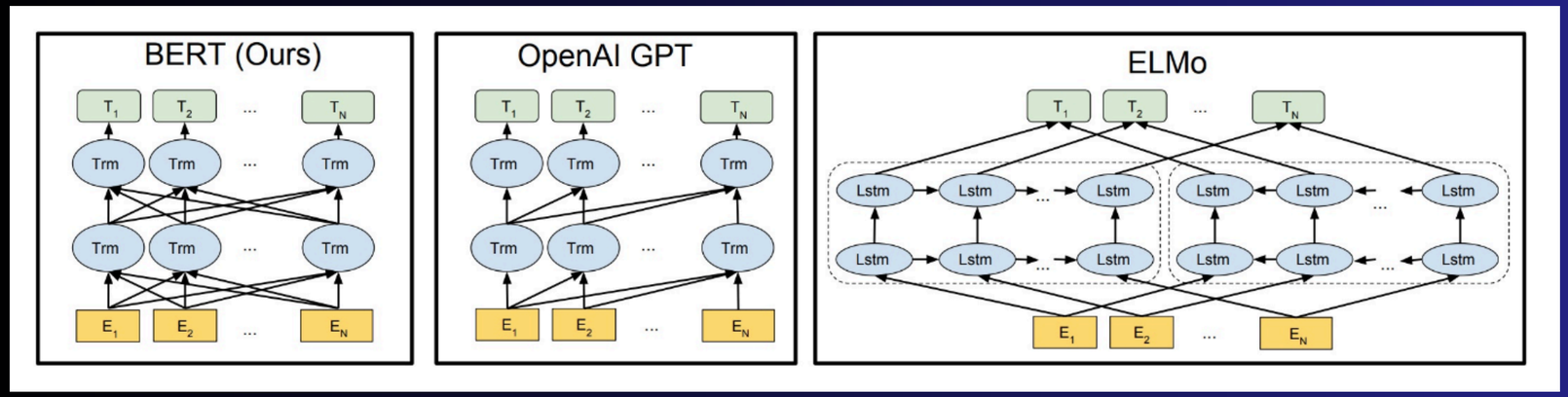
GOAL - designed to pre-train deep bidirectional representations conditioning on both left and right context in all layers

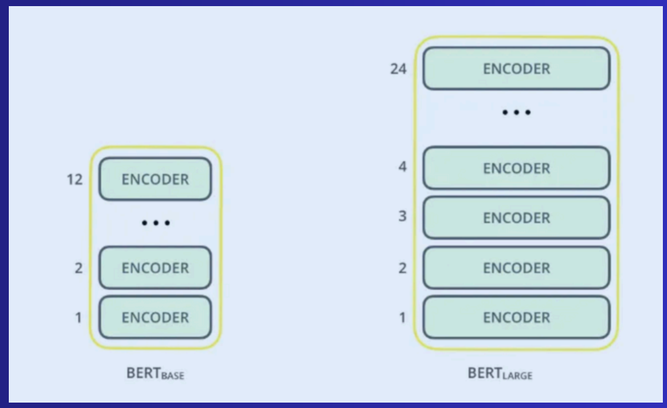
RESULT - easily fine-tuned with just one extra output layer without needing major changes to its structure

KEY OBJECTIVES OF THE PAPER

- Bert Architecture
  - Pre-Training
  - Fine-Tuning for DownstreamTasks
- BERT fine-tuning results on NLP tasks
- Ablation Studies
- Practical Implementation of BERT Model
- Implementation of BERT Models inside MediAssist

BERT ARCHITECTURE
Bidirectional Encoder Representations from Transformers



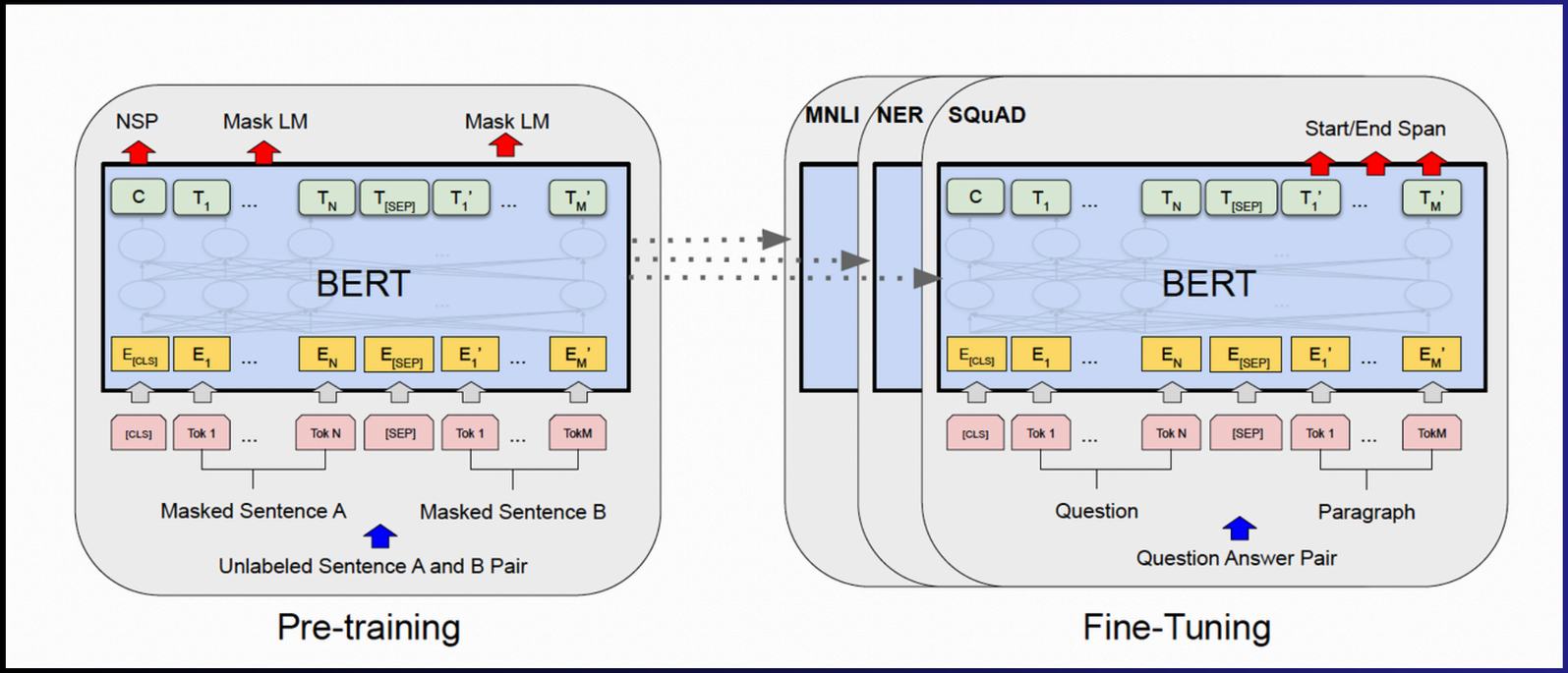


Model	Objective Function	Limitations	Why BERT is Better
ELMo (Peters et al., 2018a)	Shallowly bidirectional model	Limited contextual integration; combines separate forward and backward LSTMs only at the top layer.	BERT's deeply bidirectional model captures context from both directions at each layer, improving comprehension
OpenAI GPT	Unidirectional model	Left-to-right (unidirectional) structure limits full context understanding, critical for sentence and token tasks.	BERT's bidirectional transformer offers better context integration, enhancing performance on tasks like QA.

Magic behind BERT's success

PRE-TRAINING

BERT ARCHITECTURE - PRE-TRAINING
Bidirectional Encoder Representations from Transformers



PINK BLOCKS - Input Tokens and Tokenization

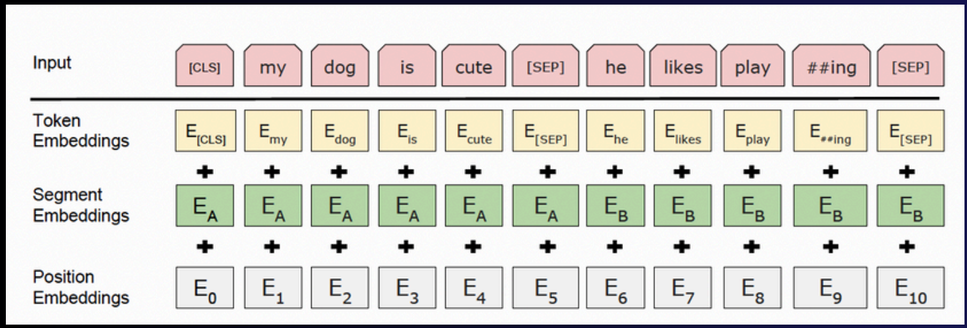
YELLOW BLOCKS: Embeddings

GREEN BLOCKS: Output Representations

Let's consider two clinical notes:

Sentence A:"Patient admitted to ICU."

Sentence B: "Patient received aspirin for chest pain."

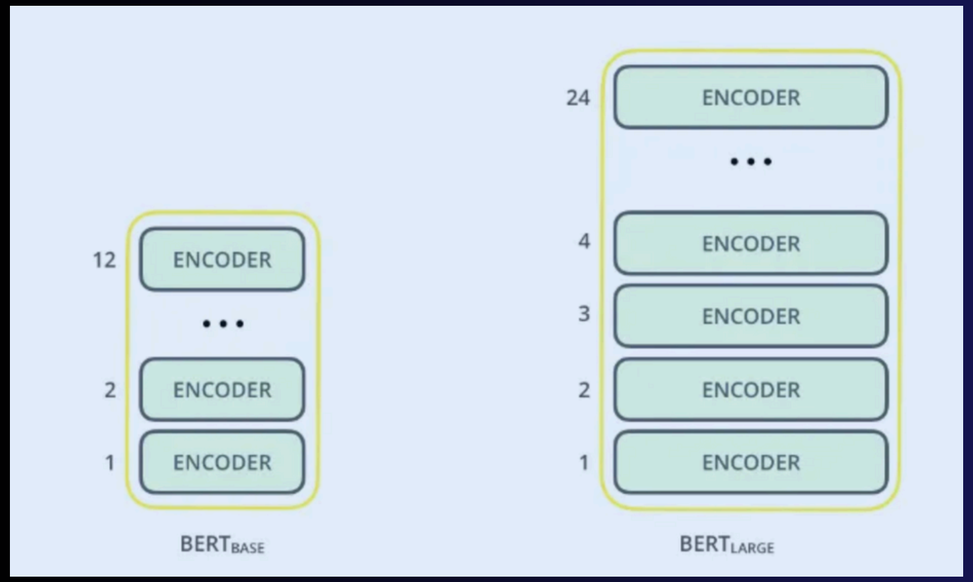


[CLS], Patient, admitted, to, ICU, [SEP], Patient, received, aspirin, for, chest, pain, [SEP]

"Patient" (token embedding) + Sentence A (segment embedding) + Position (position embedding)

The word "aspirin" in Sentence B will understand that it relates to "chest pain" based on the context provided in both sentences.

BERT ARCHITECTURE - PRE-TRAINING
Bidirectional Encoder Representations from Transformers



BERT is pre-trained on two NLP tasks:
Masked Language Modeling
Next Sentence Prediction

1. Masked Language Modeling

EXAMPLE:

"She enjoys watching horror movies”

"She enjoys watching [MASK] movies”

The model is then trained to figure out that the missing word should be "horror.”

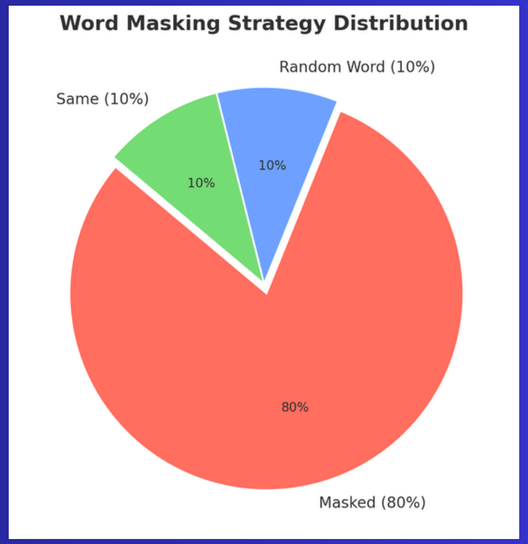
2. Next Sentence Prediction

EXAMPLE:
A: "I went to the store."

B: "I bought some apples."

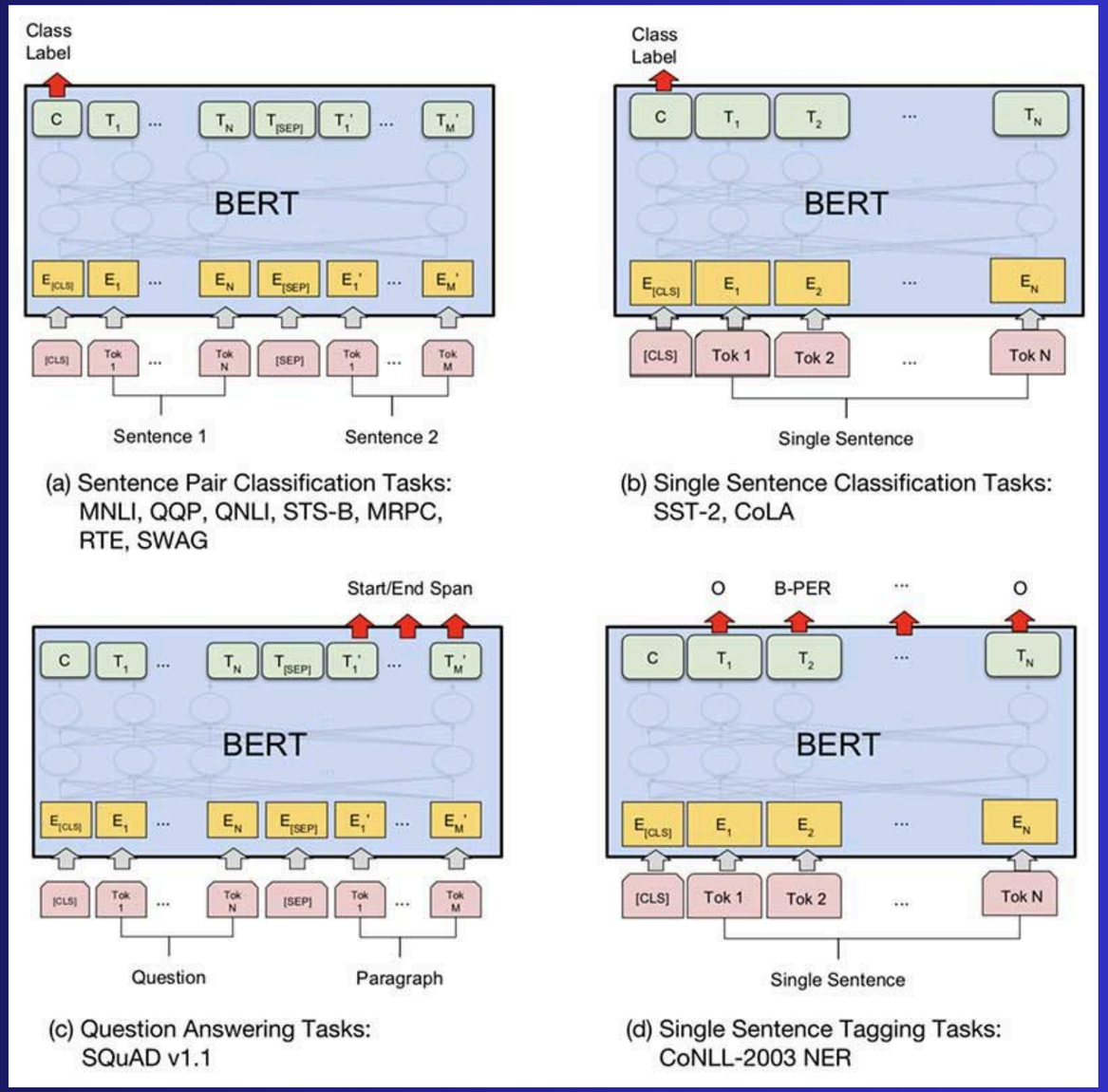
The model must predict if sentence B logically follows sentence A or if it's a random sentence

Labels: "IsNext" for correct pairs and "NotNext" for random pairs.



FINE TUNING

- Purpose of Fine Tuning
- Fine tuning on “Pretrained” Bert Model
- Minimal Architecture Changes
- Task Specific Dataset



FINE TUNING EVALUATION ON NLP TASKS

GLUE

System	Average Accuracy Score
BiLSTM+ELMo+Attn	71.0
OpenAI GPT	75.1
BERT Base	79.6
BERT Large	82.1

SQuAD

System	F1
BiDAF+ELMo (Single)	85.8
R.M. Reader (Ensemble)	88.5
BERT Large (Sgl+TriviaQA)	91.8
BERT Large (Ens.+TriviaQA)	93.2

SWAG

System	F1
ESIM+GloVe	52.7
ESIM+ELMo	59.2
OpenAI GPT	78.0
BERT Large	86.3

ABLATION STUDIES

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Effect of Pre-training Tasks

- Bidirectionality and MLM are Essential
- Importance of NSP for Sentence Relationships
- Superiority Over Left-to-Right Models

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Feature Based Approach

- Fine-Tuning Yields Optimal Results
- Feature-Based Approach as an Alternative

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Effect of Model Size

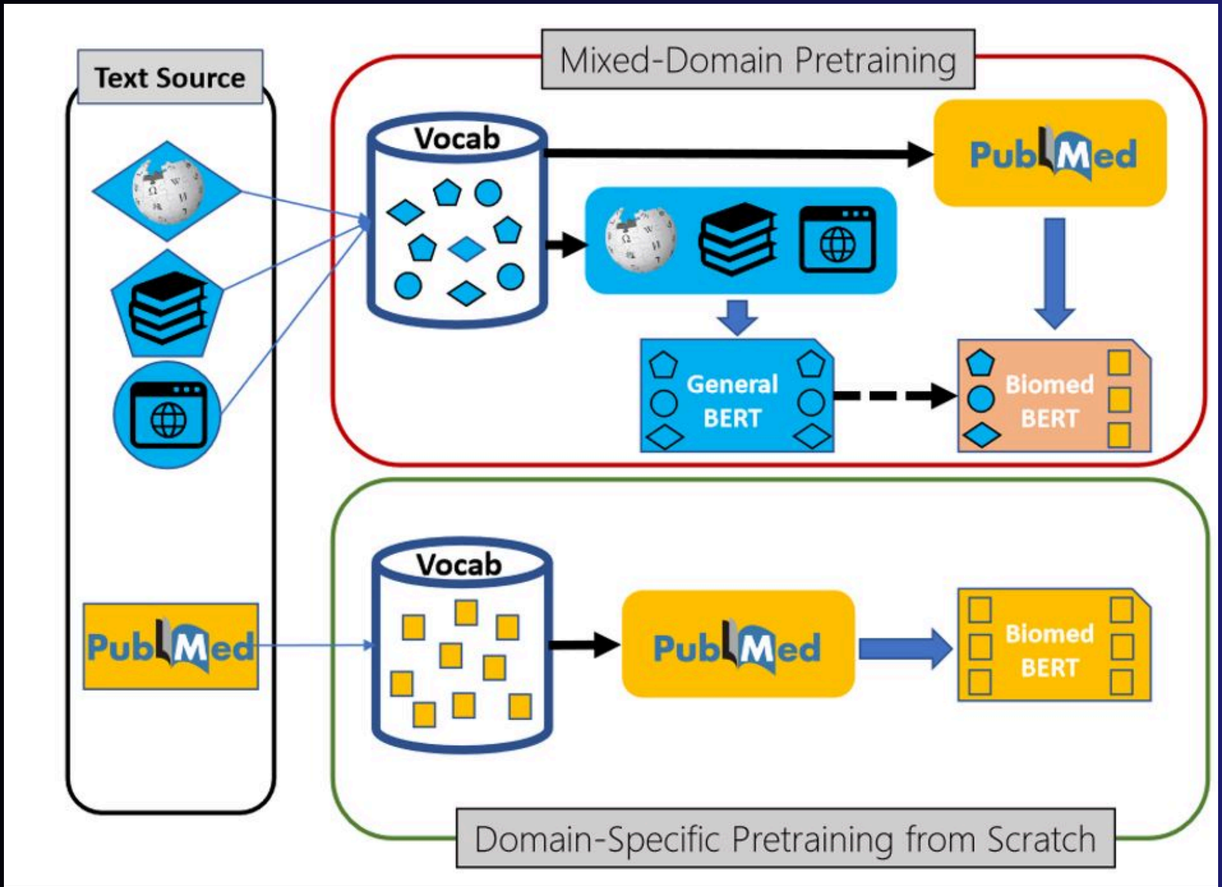
- Larger Models Yield Higher Accuracy
- Improvement even on Low Data Tasks

PRACTICAL IMPLEMENTATION OF BERT MODEL

Objective: Fine-Tuning BERT Model on IMDB movie reviews dataset from Stanford University to utilize trained model to perform sentiment analysis

Notebook Link: https://www.kaggle.com/code/draconian10/sentiment-analysis-using-bert-model

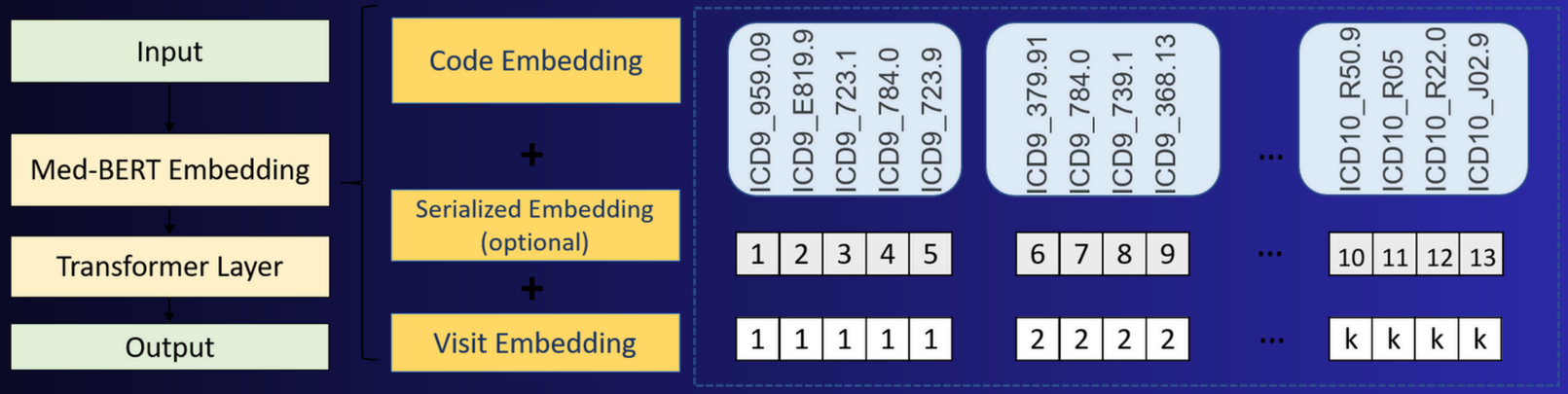
IMPLEMENTATION OF BERT MODELS INSIDE MEDIASSIST



Medical History Summarization - BiomedBERT

- Pretrained from Scratch on Biomedical Texts
- Deep Understanding of Clinical Terminology
- Accurate Medical History Summarization

IMPLEMENTATION OF BERT MODELS INSIDE MEDIASSIST



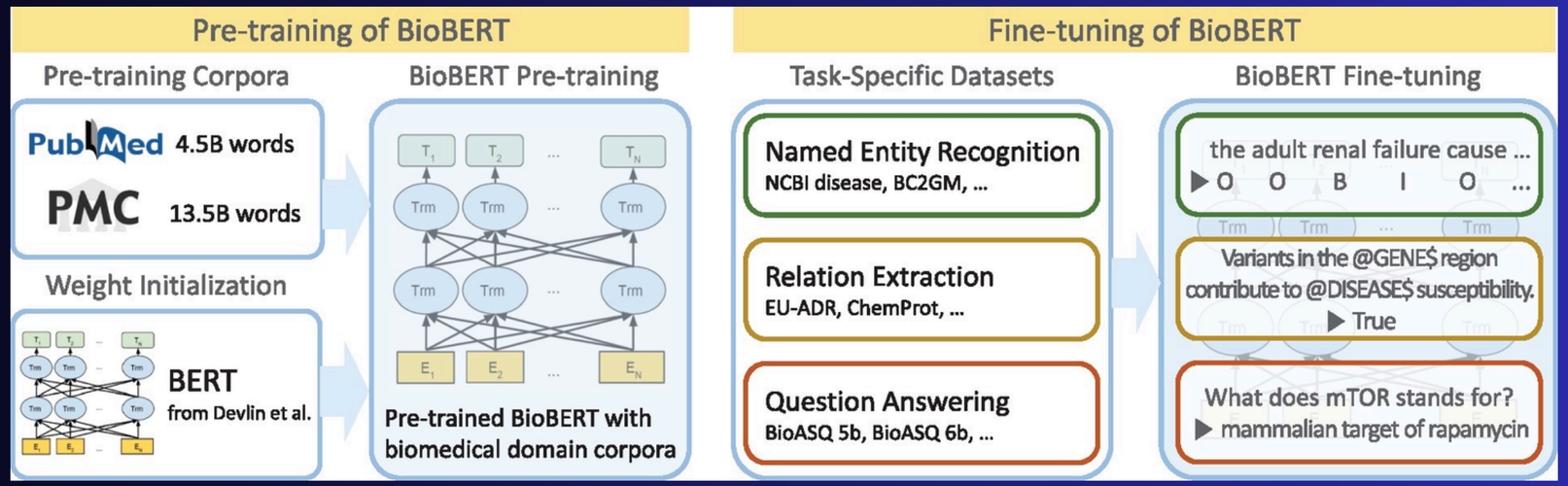
Medical Coding - MedBERT

- Trained on Large-Scale EHR Data
- Captures Sequential Patterns in Clinical Data
- Enhanced Coding Consistency

IMPLEMENTATION OF BERT MODELS INSIDE MEDIASSIST

- Biomedical-Specific Training on Large Datasets
- Automated Risk Prediction
- Enhanced Accuracy in Stratification

Patient Risk Stratification - BioBERT



REFERENCES USED

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://arxiv.org/pdf/1810.04805

- Scaler. (n.d.). Fine-tuning BERT: Concepts and applications in NLP. Scaler. Retrieved from https://www.scaler.com/topics/nlp/fine-tuning-bert

- Analytics Vidhya. (2020). BERT pre-training and fine-tuning. Medium. https://medium.com/analytics-vidhya/bert-pre-training-fine-tuning-eb574be614f6

- ResearchGate. (n.d.). BioBERT pre-training and fine-tuning overview [Figure]. Retrieved from https://www.researchgate.net/figure/BioBERT-pre-training-and-finetuning-overview-55_fig7_372074536

THANK YOU