Uses cases in Natural Language Processing

Power of Pretraining

Niloy Ganguly
IIT Kharagpur
Complex Network Research Group

Machine Learning

TASK

Classification [Classify whether a sentence expresses positive or negative sentiment]

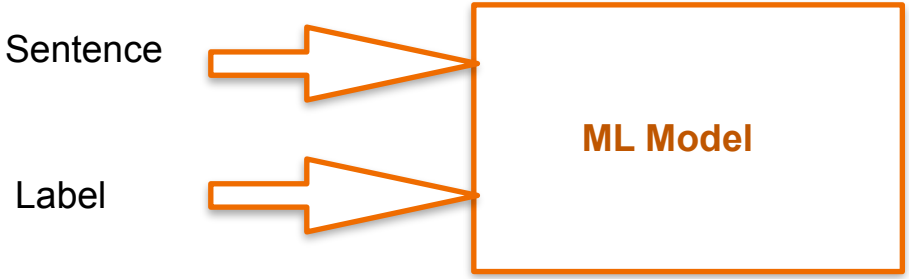
Annnotated Data

Given a sentence --> +ve, -ve, neutral

The food is not worth the price - -ve sentiment

It is a total Paisa Wassool - + ve sentiment

The lunch costs 400 bucks – neutral sentiment



Deep Learning/Machine Learning

TASK

Classification [Classify whether a sentence expresses positive or negative sentiment]

Prediction [Predict after how much time maintenance of a machine needs to be initiated]

Deep Learning is data-intensive

In order to perform certain tasks like classification, prediction, **DL** requires a lot of **annotated data** which may not be present in different situations

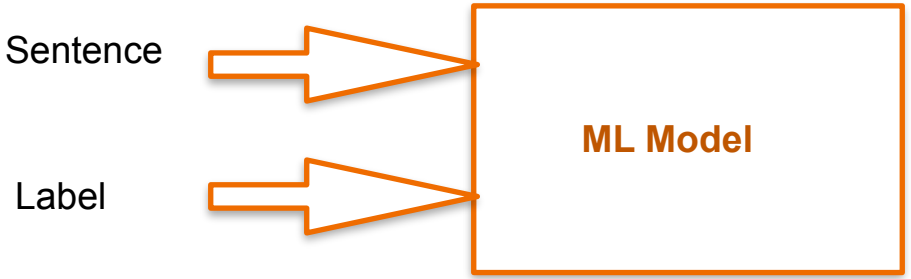
Sentence(1) — +ve sentiment, Sentence(2) — -ve sentiment Sentence(3) — +ve sentiment.

Classification

Deep Learning is data-intensive

In order to perform certain tasks like classification, prediction, **DL** requires a lot of **annotated data** which may not be present in different situations

Sentence(1) — +ve sentiment, Sentence(2) — -ve sentiment Sentence(3) — +ve sentiment.



Acquiring Labeled (annotated) data is costly

Abundance in domain-Specific NLP Applications

Very Long
E-Manual

How can I encrypt my SD Card?

Settings

Section

Encrypt or decrypt SD card

You can encrypt your optional memory card (not included) to protect its data. This only allows the SD card information to be accessed from your device with a password.

Answer

1. From Settings, tap Biometrics and security > Encrypt or decrypt SD card.
2. Tap Encrypt SD card and follow the prompts to encrypt all data on your memory card.

NOTE Performing a Factory data reset on your device prevents it from accessing an encrypted SD card. Before initiating a Factory data reset, make sure to decrypt the installed SD card first.

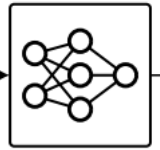
Decrypt SD card

You can decrypt an optional memory card (not included) if it was encrypted by this device. You may want to decrypt the memory card if you plan to use it with another device or before performing a Factory data

CUSTOMER SUPPORT DOMAIN

Question Answering

LEGAL DOMAIN



Output

Governing Law: "This Agreement shall be governed by the laws of the State of California without giving effect to conflict or choice of law principles." (Page 2)

⚠ **Covenant Not to Sue**: "In addition, Company shall not now or in the future contest the validity of Investor's ownership of its Intellectual Property." (Page 30)

► **Perpetual / Irrevocable License**: "Company grants to Investor a worldwide, royalty-free, exclusive, irrevocable license (with the right to grant sublicenses)." (Page 151)

Clause Span Extraction

FINANCIAL DOMAIN

Cash and Cash Equivalents

Debt Instrument Convertible Conversion Price

As of December 31, 2020, we had cash equivalents of $24.8 million and a closing stock price of $18.20 per share.

We also acquired a business loan from the U.S. Bancorp of $60.5 million.

Line of Credit Facility Maximum Borrowing Capacity

Impairment Loss

Finally, our firm reports no impairment loss for this year.

NER

Named Entity Recognition
NCBI disease, BC2GM, ...

Relation Extraction
EU-ADR, ChemProt, ...

Question Answering
BioASQ 5b, BioASQ 6b, ...

the adult renal failure cause ...
▶ O O B I O ...

Trm Trm ... Trm

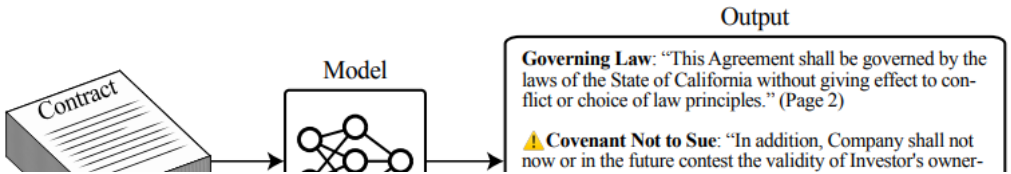
Variants in the @GENE$ region contribute to @DISEASE$ susceptibility.
▶ True

What does mTOR stands for?
▶ mammalian target of rapamycin

BIOMEDICAL DOMAIN

Several sentence and token level tasks

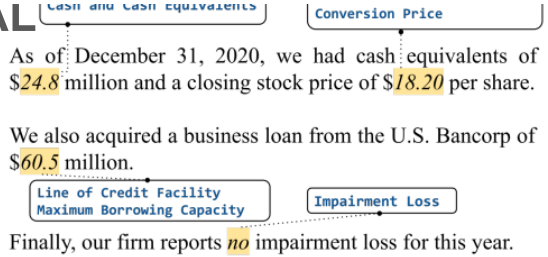
Abundance in domain-Specific NLP Applications



Domain-specific Datasets are often

- **small** in size
- costly to make, as **heavy domain-expertise** is needed
- **Unreliable** when annotated on a large-scale (e.g. crowdsourced datasets)

FINANCIAL DOMAIN



NER

Several sentence and token level tasks

Deep Learning/Machine Learning

Can we circumvent such a situation?

Can we leverage related unannotated dataset to reduce the need of annotated data.

If we can understand the semantics of unannotated data, it may help in quickly perform a specific task (classification, prediction, sentiment analysis)

Example:

Read a lot of story books to enable write good essays [Note: the first task is completely independent of the second task]

How to Leverage Unannotated Dataset

Perform a simple task of SELF-SUPERVISION

Example: The quick brown [redacted] jumps over the lazy dog.

How to Leverage Unannotated Dataset

Perform a simple task of SELF-SUPERVISION

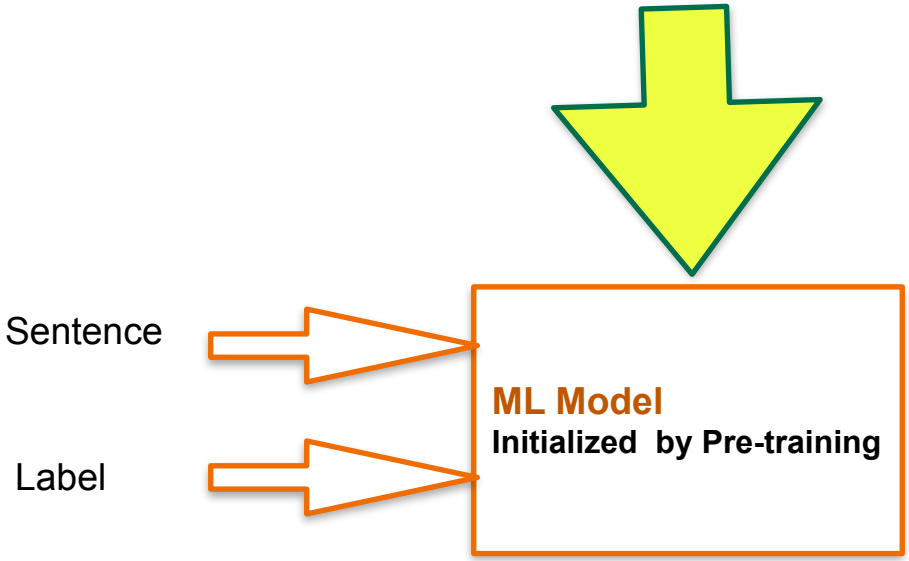
Example: The quick brown fox jumps over the lazy dog.

Perform this on millions and billions of dataset and some sorts of understanding of the language emerge.

[Pretraining - Masked Language Models]

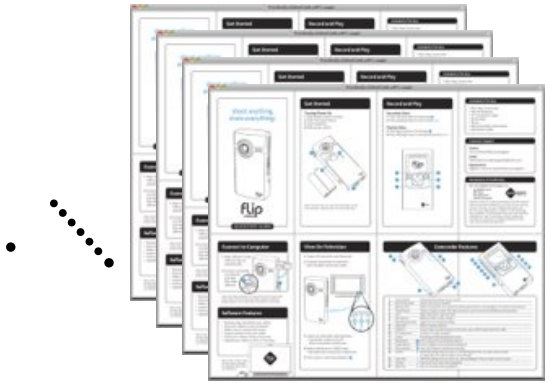
Use this understanding to perform specific tasks. [Domain Specific Tasks]

Classification



Question Answering over Electronic Devices

How can I encrypt my SD Card?



Settings

Section

Encrypt or decrypt SD card

You can encrypt your optional memory card (not included) to protect its data. This only allows the SD card information to be accessed from your device with a password.

Answer

1. From Settings, tap **Biometrics and security** > **Encrypt or decrypt SD card**.
2. Tap **Encrypt SD card** and follow the prompts to encrypt all data on your memory card.

NOTE Performing a Factory data reset on your device prevents it from accessing an encrypted SD card. Before initiating a Factory data reset, make sure to decrypt the installed SD card first.

Decrypt SD card

You can decrypt an optional memory card (not included) if it was encrypted by this device. You may want to decrypt the memory card if you plan to use it with another device or before performing a Factory data reset.

Question Answering over Electronic Devices Statement

Technical Challenges

Questions

Complex; multi- aspect

Answers

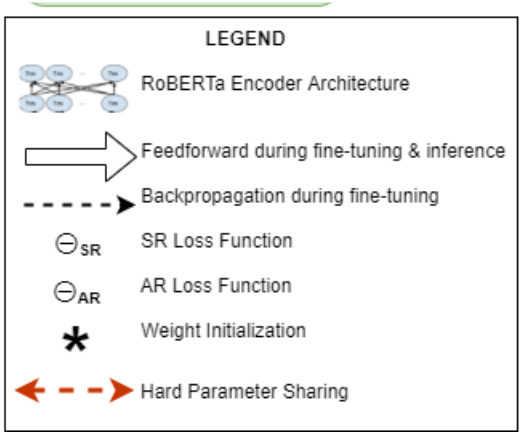
Multi-sentence
Non-contiguous multi-spans

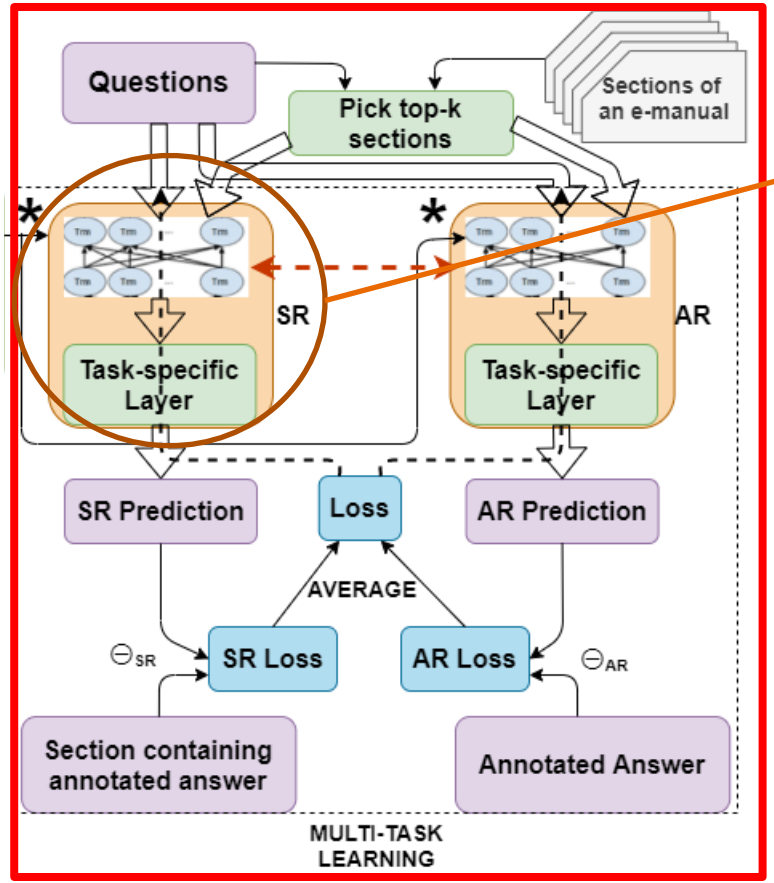
How can I save my selfies without flipping them?

Use the icons on the main camera screen and the settings menu to configure your camera's settings.

- From **Camera**, tap **Settings** for the following options:
  Intelligent features
  - **Scene optimizer**: Automatically adjust the color settings of your pictures to match the subject matter.
  - **Shot suggestions**: Get tips to help you choose the best shooting mode.
  - **Scan QR codes**: Automatically detect QR codes when using the camera.
- **Pictures**
  - **Hold shutter button to**: Choose whether to take a picture, take a burst shot, or create a GIF when holding the shutter button down.
  - **Save options**: Choose file formats and other saving options.
    HEIF pictures (Photo): Save pictures as high efficiency images to save space. Some sharing sites may not support this format.
    Save RAW copies: Save JPEG and RAW copies of pictures taken in Pro mode.
    Ultra wide shape correction: Automatically correct distortion in pictures taken with the ultra wide lens.
- **Videos**
  - **Rear video size**: Select a resolution. Selecting a higher resolution for higher quality requires more memory.
  - **Front video size**: Select a resolution. Selecting a higher resolution for higher quality requires more memory.
  - **Advanced recording options**: Enhance your videos with advanced recording formats.
    High efficiency video: Record videos in HEVC format to save space. Other devices or sharing sites may not support playback of this format.
    HDR10+ video: Optimize videos by recording in HDR10+. Playback devices must support HDR10+ video.
  - **Video stabilization**: Activate anti-shake to keep the focus steady when the camera is moving.
- **Useful features**
  - **HDR (rich tone)**: Enables the light sensitivity and color depth features of the device to produce a brighter and richer picture.
  - **Tracking auto-focus**: Keep a moving subject in focus.
  - **Pictures as previewed**: Save selfies as they appear in the preview without flipping them.

Architecture - EMQAP



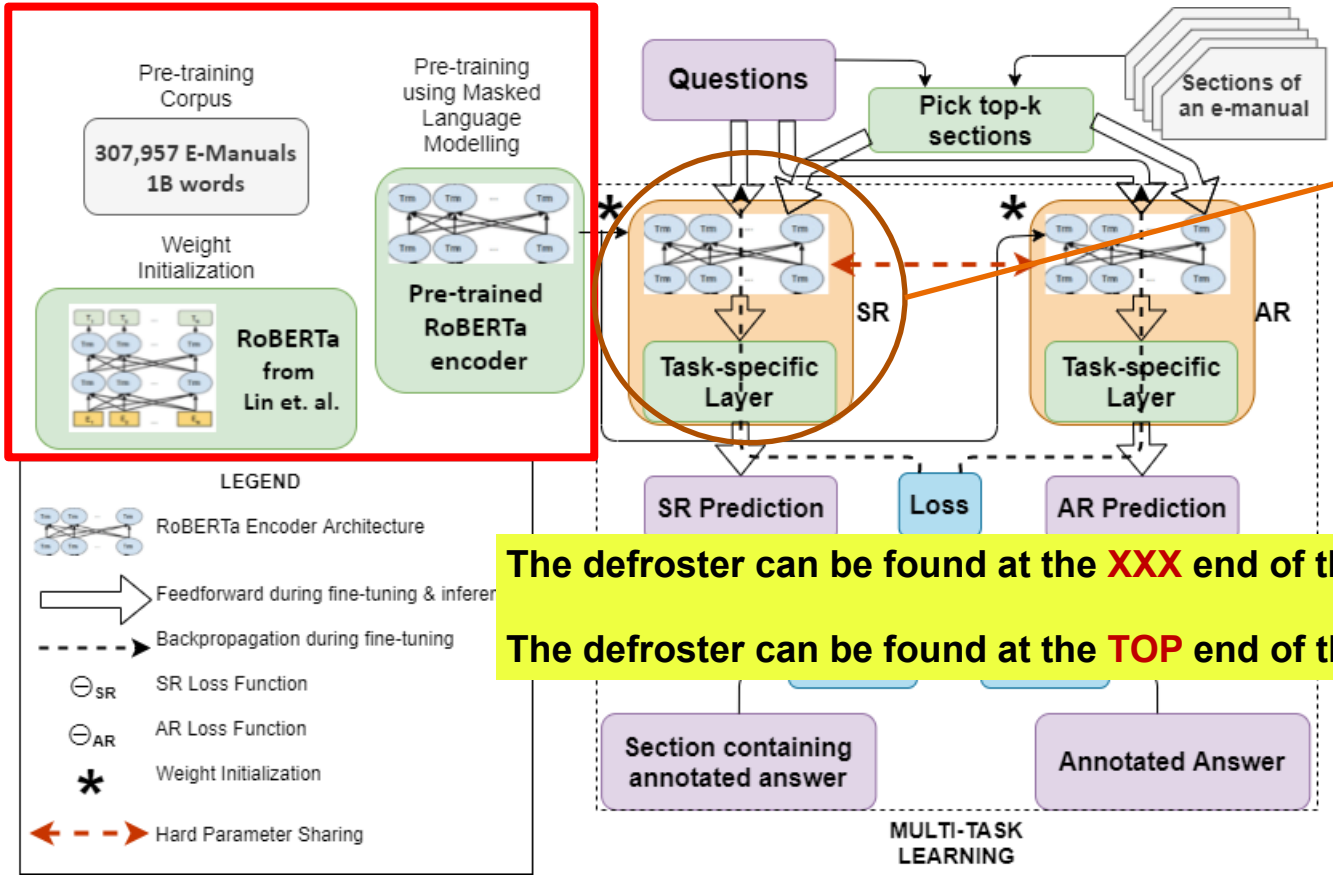


From Random Weights

Multi-Task Learning +Unsupervised IR Techniques

Architecture – with Pretraining - EMQAP

Domain-specific continual pre-training



From Weights obtained from Pretraining

The defroster can be found at the XXX end of the refrigerator.

The defroster can be found at the TOP end of the refrigerator.

Experimental Results

Model	EM	P	R	F
DPR	0	0.64	0.17	0.25
TAP	0.133	0.44	0.46	0.42
Multi-Span	0	0.94	0,14	0.22
EMQAP	**0.311**	**0.80**	**0.54**	**0.60**

Dense Passage Retrieval (DPR) (Karpukhin et al., 2020)
Technical Answer Prediction (TAP) (Castelli et al 2020)
MultiSpan (Segal et al., 2020)

EM stands for Exact Match. P(Precision), R(Recall) and F1 scores correspond to ROUGE-L.

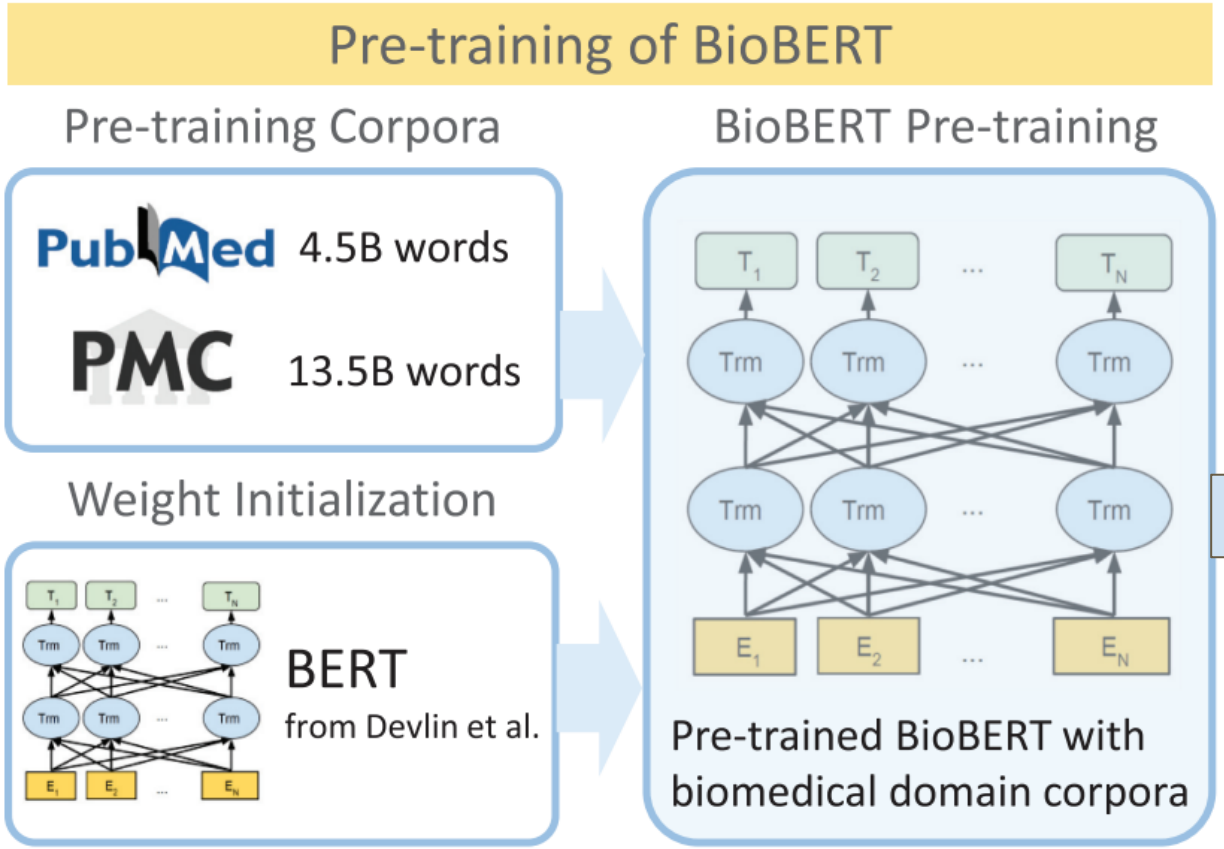
Summary

Domain Specific Pretraining Fine-Tuning Model significantly improves performance

NEXT

It takes a lot of time.

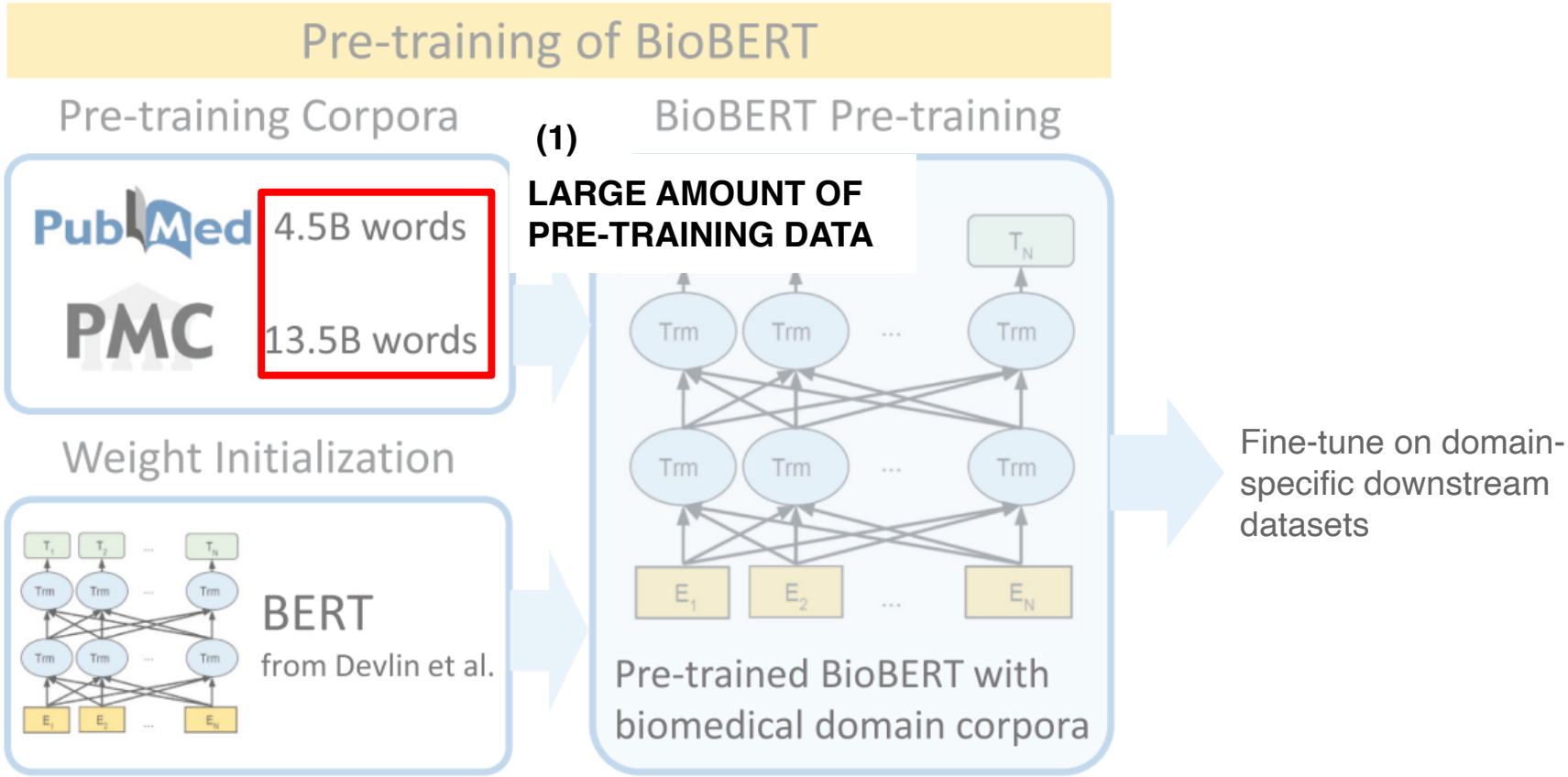
Domain-Specific Transformer Pre-training - ISSUES



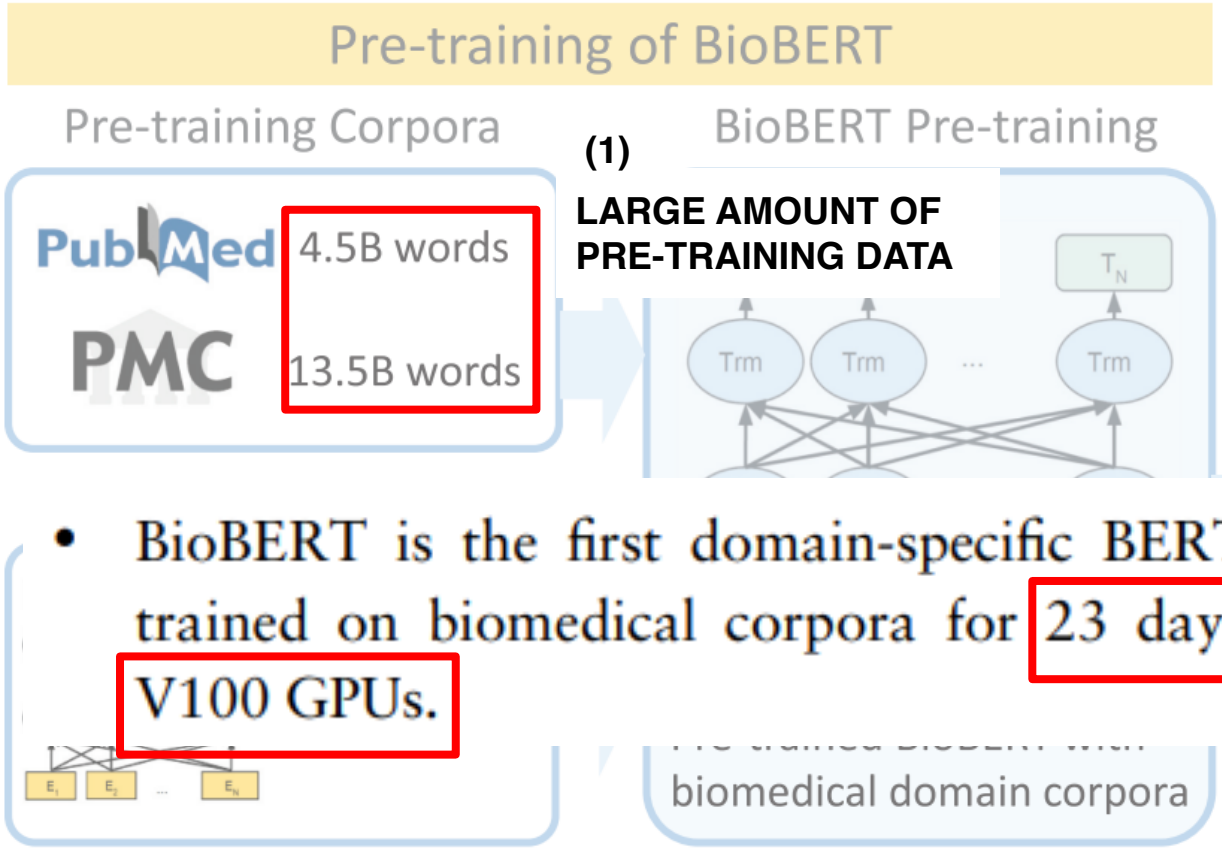
BIO-MEDICAL DOMAIN

Fine-tune on domain-specific downstream datasets

Domain-Specific Transformer Pre-training - ISSUES



Domain-Specific Transformer Pre-training - ISSUES



(1)
LARGE AMOUNT OF PRE-TRAINING DATA

(2)
LARGE AMOUNT OF PRE-TRAINING COMPUTE

- BioBERT is the first domain-specific BERT based model pre-trained on biomedical corpora for 23 days on eight NVIDIA V100 GPUs.

Agenda

A **compute and data efficient pre-training architecture** to solve sentence and token-level tasks

What are the assumptions of Masked Language Models?

Each sentence and the documents hosting them are independent entity.

In practice it is not so.
[there are several documents which are very similar to each other]

Examples: E-Manuals of a Phone Series, Movie review by different newspapers on a particular movie, scientific articles on pre-training

So can we leverage Document Level Similarity and their Categorization for pre-training

Agenda

Using **Document metadata and taxonomy** as potent supervision signals during pre-training

Domain and Data Source	Example Triplet	Example Hierarchy
Customer Support (E-Manuals Corpus)	stereo equalizer E-Manual, stereo equalizer E-Manual (of a different brand), *blu-ray player E-Manual*	**Stereo Equalizer** Electronics → Audio → Audio Players & Recorders → Stereo Systems
Scientific Domain (ArXiv)	Proximal Policy Optimization Algorithms Generating Natural Adversarial Examples *Autonomous Tracking of Intermittent RF Source Using a UAV Swarm*	**Generating Natural Adversarial Examples** Computer Science → Machine Learning
Legal Domain (EURLEX57k)	"· · · import licences · · · dairy products" "· · · market research measures · · · milk and milk products" *" · · · importations of fishery and aquaculture products"*	**"· · · importation of olive oil · · ·"** Agriculture → Products subject to market organisation → Oils and fats

Comparison

Supervision

Major Contribution - Drastic Reduction in Pre-training Compute

Domain	Model	Compute (in GPU-hours)
Customer Support	$\text{EManuals}_{\text{BERT}}$	576
	$\text{EManuals}_{\text{RoBERTa}}$	980
	DeCLUTR	370
	ConSERT	40
	SPECTER	600
	$\boldsymbol{FPDM}(CS)_{BERT}$	0.58
	$\boldsymbol{FPDM}(CS)_{RoBERTa}$	0.75
Scientific Domain	SciBERT	7680
	$\boldsymbol{FPDM}(Sci.)_{BERT}$	1.7
Legal Domain	$\text{RoBERTa}_{\text{BASE}}$ + Contracts Pre-training	710
	$\boldsymbol{FPDM}(Leg.)_{RoBERTa}$	1.49

1000x less!

1300x less!

4500x less!

480x less!

Summary

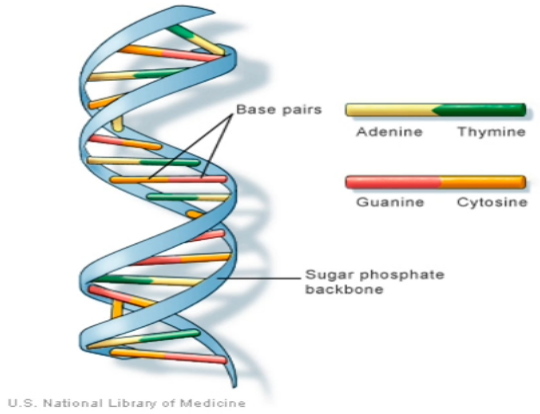
Frugal Pre-training leveraging Document Level Semantics shows dramatic improvement

Solves the requirement of huge compute infrastructure

Next

What do we do where semantics is not available – Non-Language Strings (Genes)

GeneMask: Fast Pretraining of Gene Sequences to Enable Few-shot Learning



Source: https://medlineplus.gov/genetics/understanding/basics/dna/

	Reference Genome	A Person's Genome
What is it?	+ Mitochondrial DNA	+ Mitochondrial DNA
How many chromosomes?	24 (22 + X + Y)	46 (23 PAIRS)
How many letters?	~3.2 bn	~6.4 bn
How to think about it?	• The Human Genome Project and its goal of a first draft of "the human genome" • Serves as a standard for comparison • A "consensus" genome sequence	• The genome of a person • The genome within a person's cells • The whole genome sequence of an individual

Source: https://www.veritasgenetics.com/our-thinking/whole-story/

By 2025, there will be 2-40 exabytes of human genome data [Stephens et.al., PLOS Biology, 2015]

Human genomes are very large in size!

Genomic Pre-training

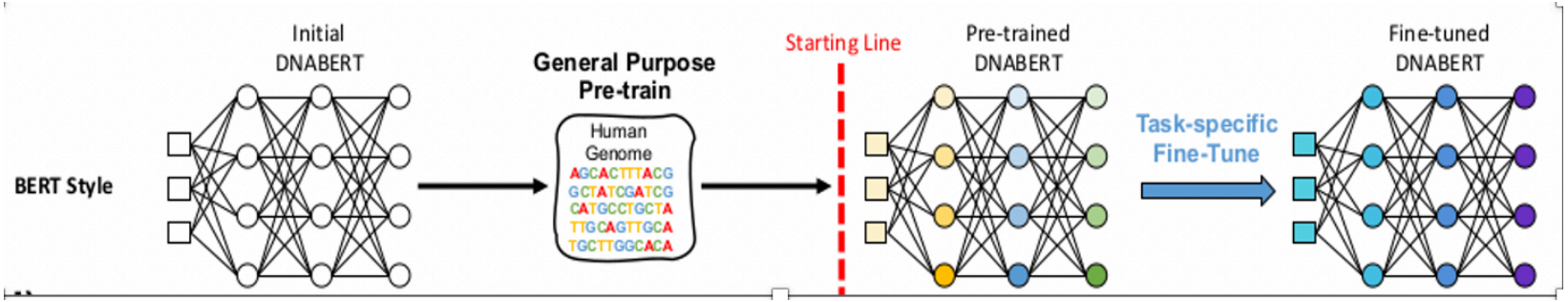
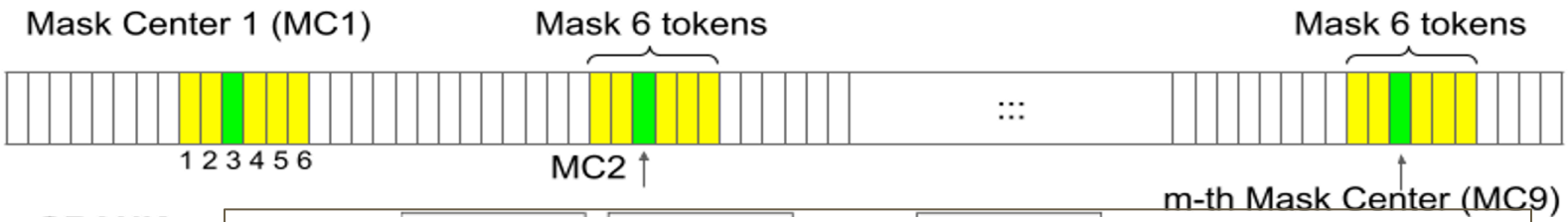
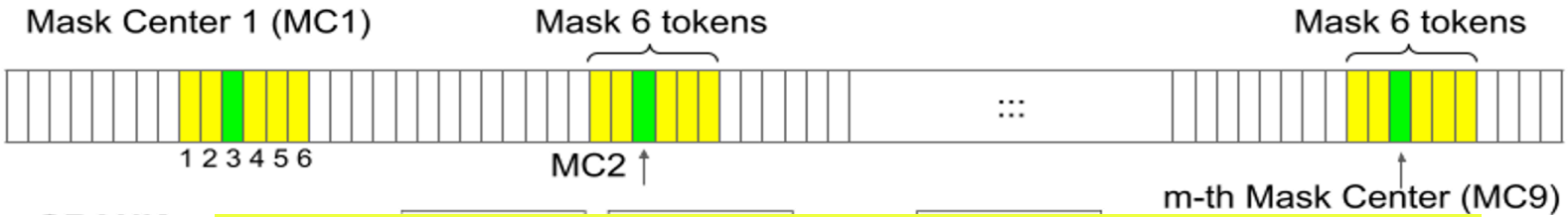


Image: Ji et al. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, Bioinformatics, pp. 1-9



Randomly Select m mask center – Following the MLM model

Pitfall of Random Masking



The quick brown fox jumps over the lazy dog.

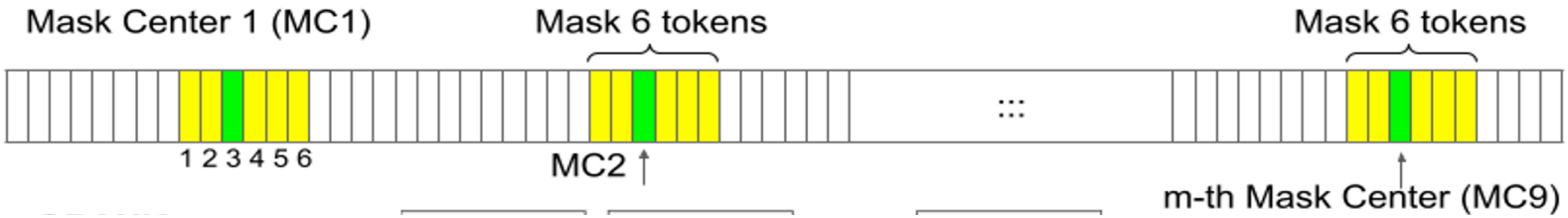
The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog.

Training steps are wasted for either too **"easy"** or too **"difficult"** predictions

Easy – other example New York

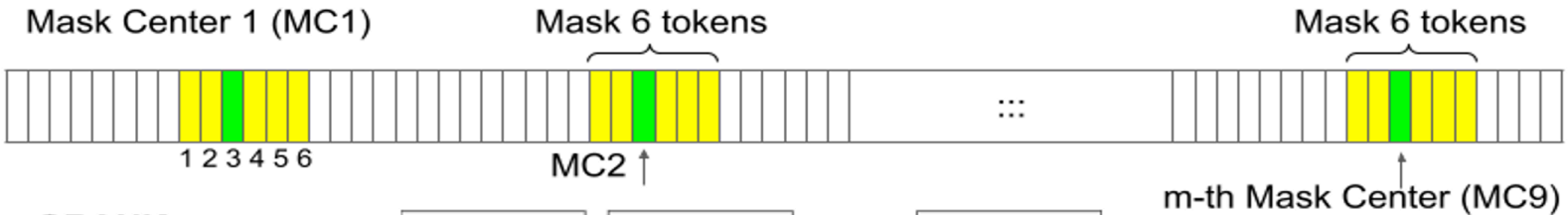
Pitfall of Random Masking



Alternate to → I live [redacted] New York

Mask entirely → I live [redacted] York

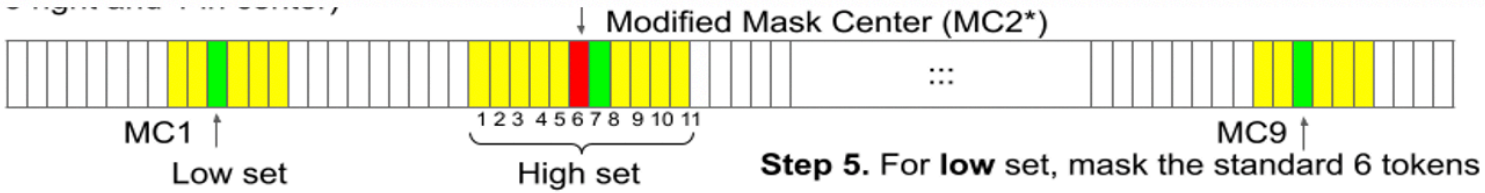
PMI-Based Masking



Alternate to → **I live in New York**

Mask entirely → I live in New York

PMI-based Technique to identify such frequent co-occurrence [Pointwise Mutual Information]



Step 5. For **low** set, mask the standard 6 tokens

Pretraining Steps: DNABert and LOGO

Both DNABert and LOGO is (pre)trained for **120K steps**

GeneMask is (pre)trained for **10K steps**

Downstream Tasks [Few Shot Setting]

- Promoter Region Prediction - binary classification
- Enhancer prediction - 500 bp
- Splice Donor and Acceptor Site Prediction - predict whether donor, acceptor or non-splice site (3-way classification) - 40 bp

k-shot	10	50	100	500	1000
DNABert	2.94%	0.93%	0.73%	0.40%	1.85%
LOGO	4.92%	5.87%	3.90%	7.74%	2.85%

Table 2. Percentage improvement in average accuracy over four datasets of GENEMASK 10K over ORI 10K model.

Summary

GeneMask ensures substantial speedup of 10x and performance improvement over random masking strategy of SoTA models (DNABert and LOGO) in few-shot settings

Incorporating domain knowledge while pretraining needs to be more explored

Complex Network Research Group (CNeRG)
IIT Kharagpur

- https://cnerg-iitkgp.github.io/ @cnerg
- facebook.com/iitkgpcnerg/

Thank You for Listening

Danke Schön

Any Questions?

Email: niloy@cse.iitkgp.ac.in
Complex Network Research Group (CNeRG) : @cnerg