MEDVOC: Vocabulary Adaptation for Fine-tuning Pre-trained Language Models on Medical Text Summarization

Gunjan Balde*, Soumyadeep Roy*, Mainack Mondal, and Niloy Ganguly

Indian Institute of Technology Kharagpur
L3S Research Center, Germany
*Equal Contribution

IJCAI JEJU 2024

Summarization in Medical Domain

- Summarization in medical domain useful for variety of tasks:
  - Patient Health Query
  - Radiology Report
  - Patient Notes and many more…
- BART and PEGASUS are state of the art open-domain models

But PLMs generalizes poorly to medical domain

Medical Summarization Example

Source Document: KNEE OSTEOARTHRITIS … ruptured anterior cruciate ligament … meniscus… my question is can you treat me, or turn me can recommend doctors or hospitals to treat in the U.S.

Reference Summary: How can I find physician(s) or hospital(s) who specialize in knee osteoarthritis?

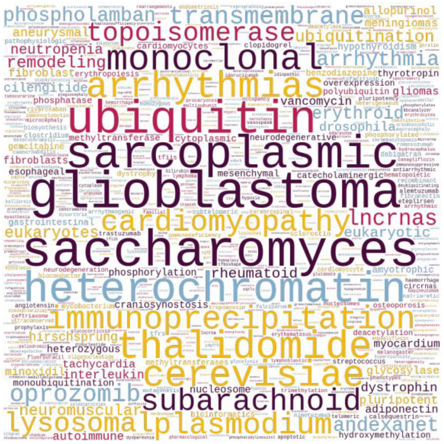
Generated Summary (PLM): What are the treatments for ruptured anterior cruciate ligament and meniscus?

Issue: PLM Poorly Tokenizes Medical Words

- Medical Out-of-vocabulary (OOV) words

  glioblastoma: g ##lio ##bla ##sto ##ma

- Affects both encoder and decoder:
  - **Encoder**: Poor tokenization blurs the meaning[1]
  - **Decoder**: Generate more tokens per word[2]



[1] Hofmann, Valentin et al. "An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers." ACL 2022.
[2] Rust, Phillip, et al. "How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models." ACL IJCNLP 2021.

Vocabulary Adaptation tackles OOV Words

- Vocabulary Adaptation strategies extends PLM vocabulary
  - Append domain-specific vocabulary **during fine-tuning**
- Existing works (AVocaDo[1], exBERT[2])
  - **Limited to only classification tasks** using **encoder-only** architecture

[1] Hong, Jimin, et al. "AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain." EMNLP 2021.
[2] Tai, Wen, et al. "exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources."EMNLP Findings. 2020.

Vocabulary Adaptation tackles OOV Words

- Vocabulary Adaptation strategies extends PLM vocabulary
  - Append domain-specific vocabulary **during fine-tuning**
- Existing works (AVocaDo, exBERT)
  - **Limited to only classification tasks** using **encoder-only** architecture

No work on vocabulary adaptation for domain-specific summarization

Vocabulary Adaptation tackles OOV Words

- Vocabulary Adaptation strategies extends PLM vocabulary
  - Append domain-specific vocabulary **during fine-tuning**
- Existing works (AVocaDo, exBERT)
  - **Limited to only classification tasks** using **encoder-only** architecture

No work on vocabulary adaptation for domain-specific summarization

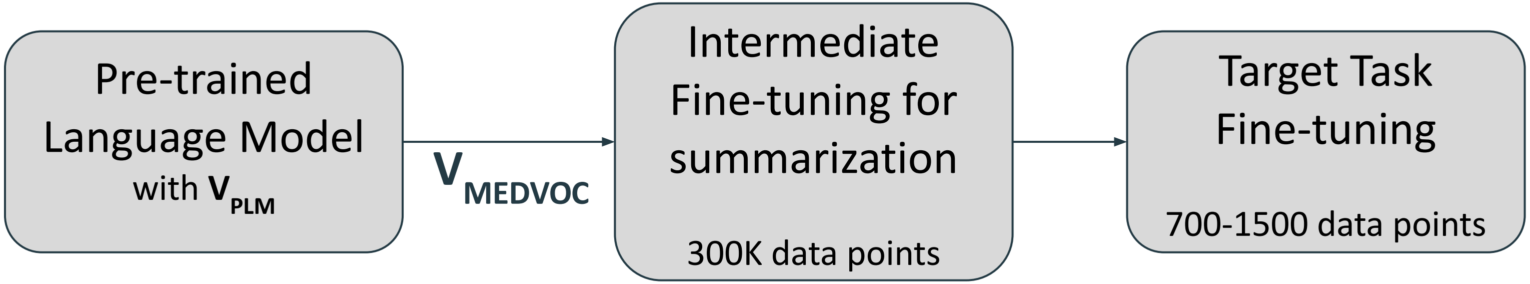
It requires tackling encoder-decoder / decoder-only architecture

Can we adapt PLM vocabulary during fine-tuning on domain-specific summarization tasks?

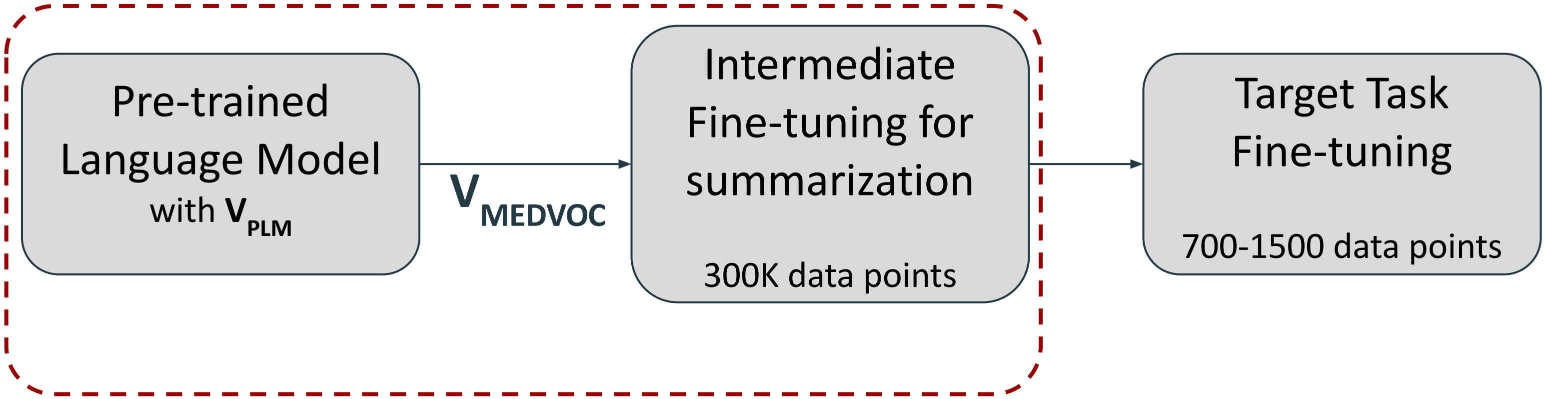
Can we adapt PLM vocabulary during fine-tuning on domain-specific summarization tasks?

MEDVOC: Vocabulary adaptation strategy for domain-specific summarization

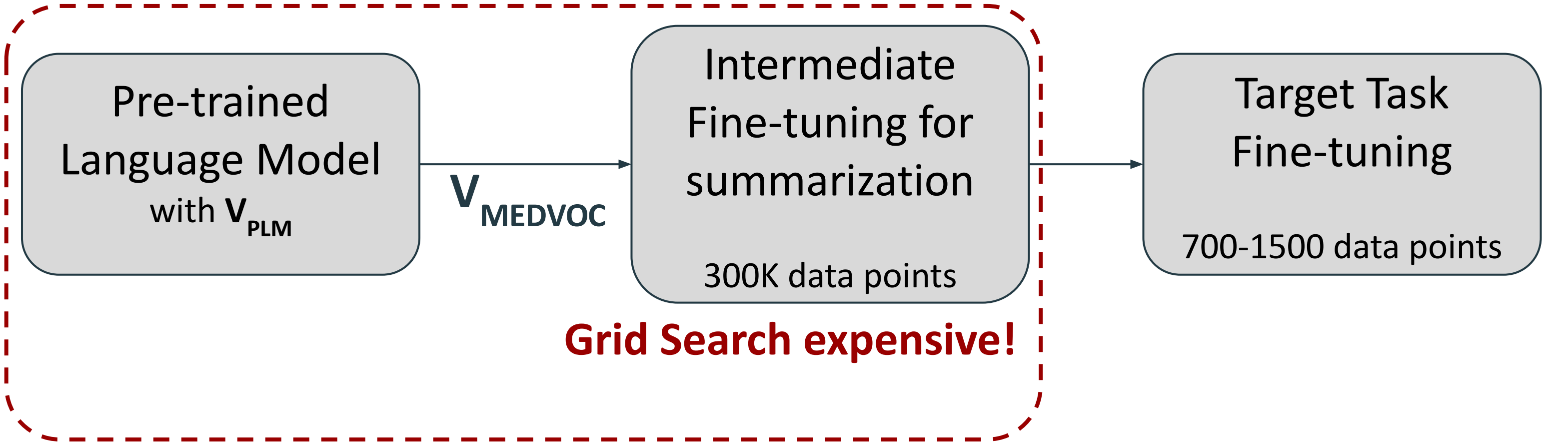
MEDVOC at a glance



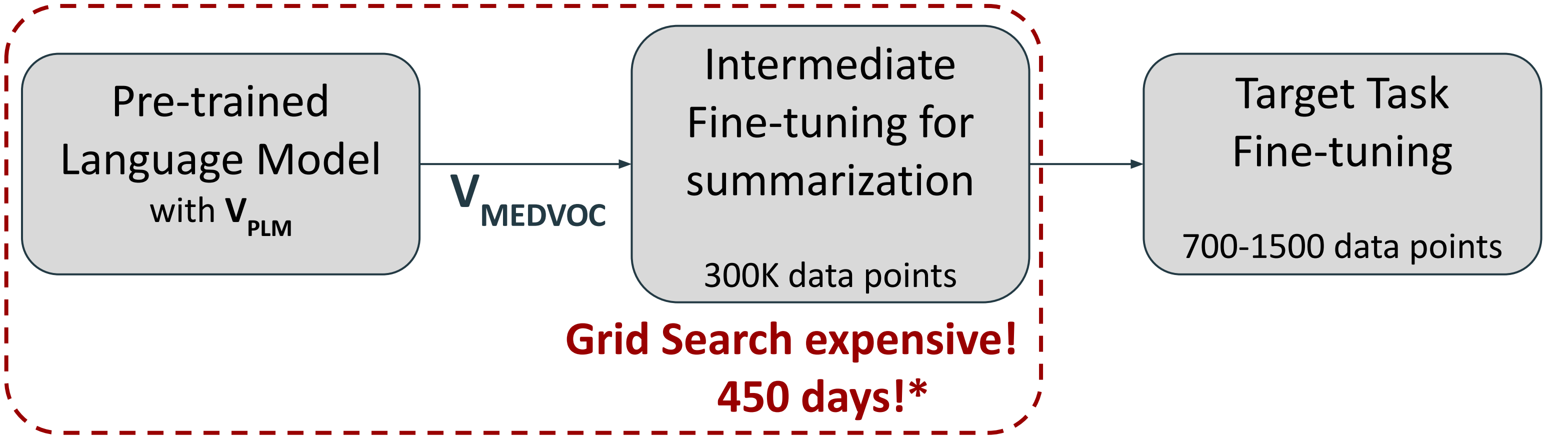
MEDVOC at a glance



MEDVOC at a glance

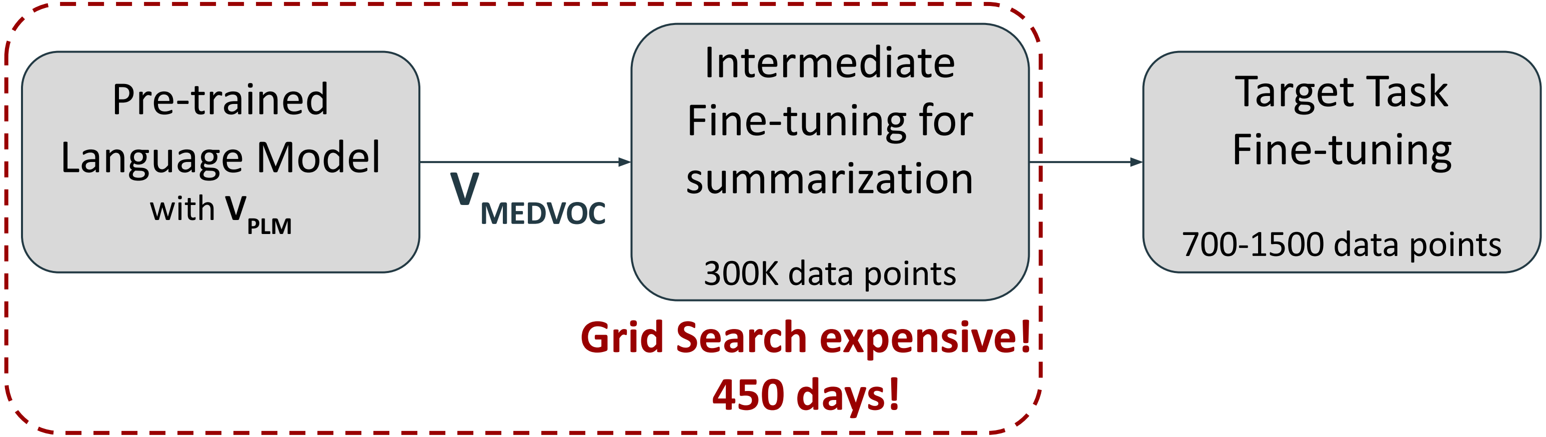


MEDVOC at a glance



*On 3 Tesla V100 GPUs

MEDVOC at a glance



- MEDVOC optimizes fragment score to optimize performance
- Brings down search time to **2 hours**

Evaluation Setup

- Three summarization models —BertSumAbs, BART, and PEGASUS
- Three kinds of baselines:
  - Vocabulary adaptation —AVocaDo[1] and PubMedBERT[2]
  - Domain-Specific Models —BioBERT and BioBART
  - two ablation baselines —IFT-CNN and IFT-PAC
- Metrics: **Rouge**, BertScore, and Concept-Score

[1] Hong, Jimin, et al. "AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain." EMNLP 2021.
[2] Gu, Yu, et al. "Domain-specific language model pretraining for biomedical natural language processing." ACM Transactions on Computing for Healthcare (HEALTH) 2021.

Datasets

- Query Focused Summarization
  - EBM, BioASQ
    - SD —Query + PubMed Abstract
    - RS —Answer
- Patient Healthcare Query Summarization
  - MeQSum, CHQSum
    - SD —Patient Query
    - RS —a concise Question

MEDVOC outperforms baselines across tasks

Dataset	R-L (Domain)	R-L (IFT-CNN)	R-L (IFT-PAC)	R-L (MEDVOC)
EBM	19.19	18.65	<u>19.35</u>	**20.03**
BioASQ	44.23	42.31	<u>43.60</u>	**45.98**
MeQSum	51.23	50.51	<u>52.47</u>	**53.63**
CHQSum	37.56	38.36	<u>38.45</u>	**38.75**

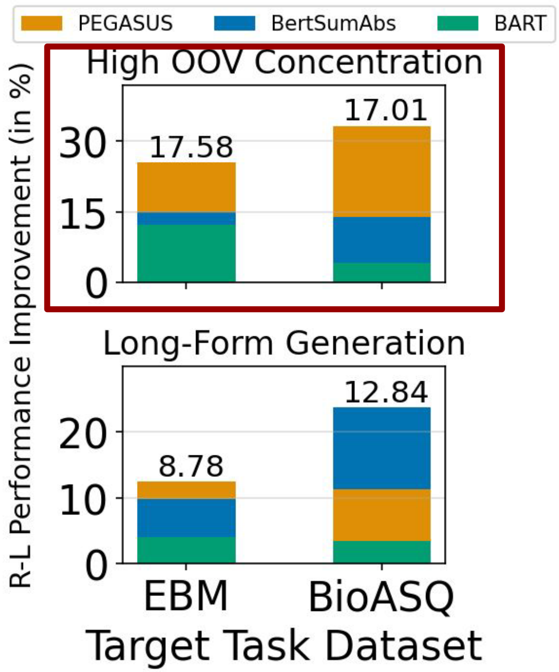
MEDVOC outperforms existing vocabulary adaptation baselines

- AVocaDo- strategy for classification
  - **4%** improvement
- PubMedBERT - Domain-specific vocabulary
  - **33%** improvement

	PubMed BERT	AVocaDo	MEDVOC
EBM	17.76	18.43	**19.59**
BioASQ	26.65	45.86	**47.54**
MeQSum	39.79	49.30	**51.49**
CHQSum	30.59	34.49	**35.11**

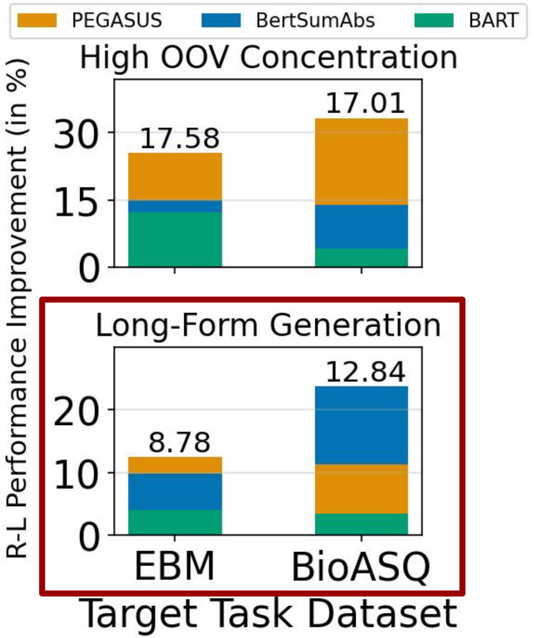
MEDVOC improves over baselines in challenging scenarios

- <u>**High OOV concentration**</u>
  - Subset where RS has high unigram OOV
  - **17.29%** improvement



MEDVOC improves over baselines in challenging scenarios

- <u>**High OOV concentration**</u>
  - Subset where RS has high unigram OOV
  - **17.29%** improvement
- <u>**Long-Form Generation**</u>
  - Subset where RS was longer in length
  - **10.81%** improvement



Conclusion and Key Takeaways

- **First vocabulary adaptation work** for domain-specific summarization
- MEDVOC adapts models of **varying parameter and vocabulary sizes**
- Minimal additional parameters upto 1.59% w.r.t. original model size
- Medical experts found MEDVOC summaries to be **more faithful**
- MEDVOC generalizable to other domains with high vocabulary mismatch

Thank You!

Soumyadeep Roy

Mainack Mondal

Niloy Ganguly

CNeRG

Codebase

Preprint

PMRF
Prime Minister's Research Fellows