Continual BERT

Data-driven summarization of scientific literature

Jongwon Park, University of Illinois

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

What is Continual BERT?

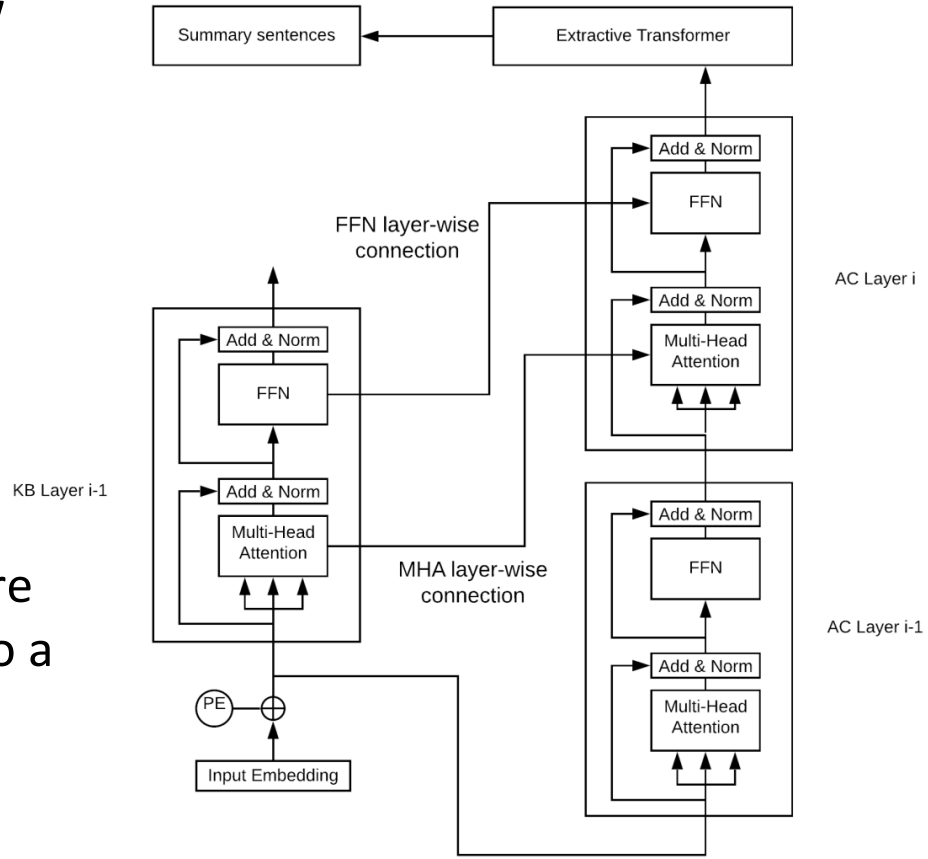
Science.org: **“Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?”**

- 23,000+ COVID-19 papers from Jan-May 2020 alone (150+/day)

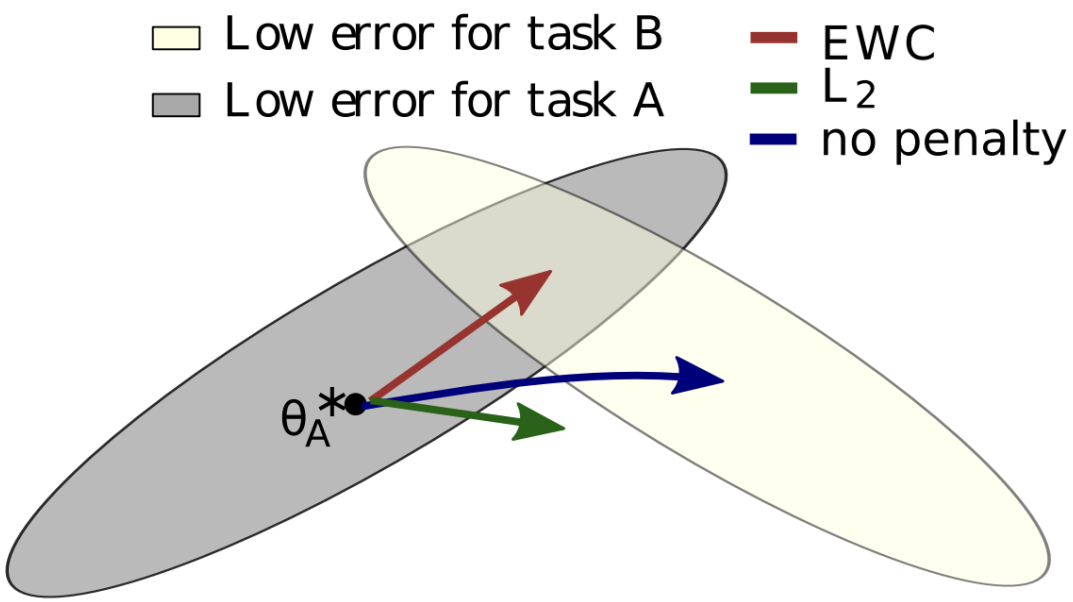
Question: How can we transform the massive influx of COVID-19 literature into concise feeds for researchers and readers to comprehend swiftly (in layman’s terms)?

Solution: *Continual BERT* (Bidirectional Transformers)

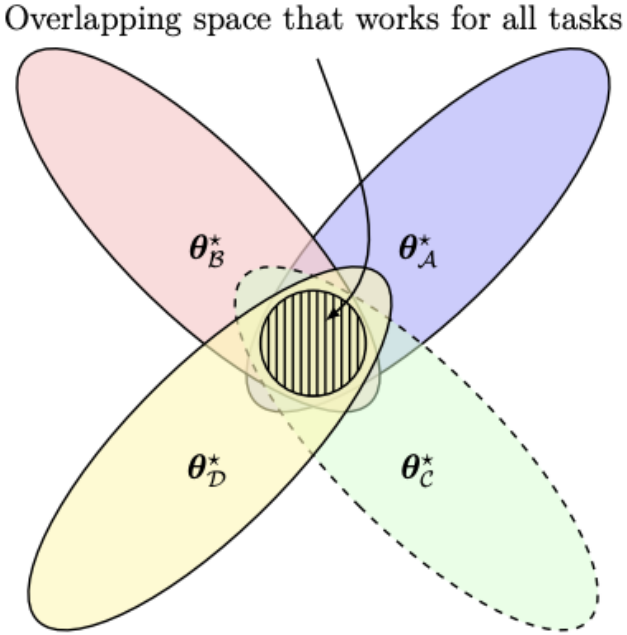
- **Continually learns** from chronological feed of scientific literature
- **Adaptively recognizes** critical sentences and combine them into a concise summary (in simpler words)
- Trained on **medical literature dataset**, including COVID-19, to deal with the rapid influx of scientific literature during the pandemic



Elastic Weight Consolidation (EWC)



$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$



(b) Example for four tasks

[1] Overcoming catastrophic forgetting in neural networks (Kirkpatrick et al)
[2] Elastic Weight Consolidation (EWC): Nuts and Bolts (Aich)

Data Drives the Continual Learning

What data does Continual BERT require? **Scientific literature/articles and high-quality summaries**

Examples of high-quality summaries include:

- **Well-written (summary-like) abstracts**
- **Concise summaries produced during peer reviews**
- **Manually written summaries by readers**

+ Well-grouped clusters of literature/articles

How are the data used?

1. Literature/articles grouped into clusters of certain categories (questions, targets, experiments, etc.)
2. Model trained on a cluster after cluster, learning parameters that store common information
3. Some articles used to test the model (requiring manual evaluation)
4. New influx of data are processed and clustered for training

Current Datasets

Name	Components	Acquisition
CTD-Pfizer	**88,000**+ scientific articles text (drug-disease and drug-phenotype interactions)	Manually mined text and abstracts (written by authors)
ScisummNet (Stanford)	**1,000**+ papers in ACL anthology network with comprehensive summaries	Human-annotated, manual summaries
CORD-19	**500,000**+ scholarly articles about coronaviruses since 1950s (most crude data)	Auto aggregated from PubMed Central and journals
Total	***~600,000** scientific literature & articles*	

Example Data

Structure of the full SARS-CoV-2 RNA genome in infected cells (Lan et. al, 2020)

Original

SUMMARY

SARS-CoV-2 is a betacoronavirus with a single-stranded, positive-sense, 30-kilobase RNA genome responsible for the ongoing COVID-19 pandemic. Currently, there are no antiviral drugs or vaccines with proven efficacy, and development of these treatments are hampered by our limited understanding of the molecular and structural biology of the virus. Like many other RNA viruses, RNA structures in coronaviruses regulate gene expression and are crucial for viral replication. Although genome and transcriptome data were recently reported, there is to date little experimental data on predicted RNA structures in SARS-CoV-2 and most putative regulatory sequences are uncharacterized. Here we report the secondary structure of the entire SARS-CoV-2 genome in infected cells at single nucleotide resolution using dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq). Our results reveal previously undescribed structures within critical regulatory elements such as the genomic transcription-regulating sequences (TRSs). Contrary to previous studies, our in-cell data show that the structure of the frameshift element, which is a major drug target, is drastically different from prevailing *in vitro* models. The genomic structure detailed here lays the groundwork for coronavirus RNA biology and will guide the design of SARS-CoV-2 RNA-based therapeutics.

Manual

RESEARCH HIGHLIGHTS

1. First in-cell map at single nucleotide resolution of the secondary structures of the SARS-CoV-2 genome
2. Uncover new structures in regulatory elements, including the genomic transcription-regulating sequences (TRSs)
3. Provide new structures of the frameshift element, which differ from previous in vitro models. There is evidence for heterogeneity of

SUMMARY

The authors performed the first in-cell map of the secondary structures of the SARS-CoV-2 genome. They used dimethyl sulphate mutational profiling with sequencing (DMS-MaPseq) to find unpaired nucleotides, which were used to constrain in-silico pairing predictions. The 5'UTR structures are similar to previous reports, but new structures were found for the genomic Transcription-Regulating Sequences (TRSs), which mostly lie within stem loops. The structure of the frameshift element (FSE), an important regulatory element of SARS-CoV-2 viral cycle, differed from previous in vitro observations, which the authors argue were an artefact of short length of the refolded viral RNA used for the in vitro studies. Clustering of DMS-MaPseq reads suggested the presence of two distinct FSE structures across different viral copies.

Example Result

Structure of the full SARS-CoV-2 RNA genome in infected cells (Lan et. al, 2020)

Original

SUMMARY

SARS-CoV-2 is a betacoronavirus with a single-stranded, positive-sense, 30-kilobase RNA genome responsible for the ongoing COVID-19 pandemic. Currently, there are no antiviral drugs or vaccines with proven efficacy, and development of these treatments are hampered by our limited understanding of the molecular and structural biology of the virus. Like many other RNA viruses, RNA structures in coronaviruses regulate gene expression and are crucial for viral replication. Although genome and transcriptome data were recently reported, there is to date little experimental data on predicted RNA structures in SARS-CoV-2 and most putative regulatory sequences are uncharacterized. Here we report the secondary structure of the entire SARS-CoV-2 genome in infected cells at single nucleotide resolution using dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq). Our results reveal previously undescribed structures within critical regulatory elements such as the genomic transcription-regulating sequences (TRSs). Contrary to previous studies, our in-cell data show that the structure of the frameshift element, which is a major drug target, is drastically different from prevailing *in vitro* models. The genomic structure detailed here lays the groundwork for coronavirus RNA biology and will guide the design of SARS-CoV-2 RNA-based therapeutics.

Evaluations

SARS-CoV-2 is an enveloped virus belonging to the genus **beta coronavirus**, which also includes SARS-CoV, the virus responsible for the 2003 **SARS** outbreak, and Middle East Respiratory Syndrome Coronavirus (**MERS** CoV), the virus responsible for the 2012 MERS outbreak. Despite the devastating effects these viruses have had on public health and the economy, currently **no effective antivirals treatment or vaccines exist**. There is therefore an urgent need to understand their **unique RNA biology** and develop new therapeutics against this class of viruses. Coronaviruses have **single-stranded and positive-sense genomes that are the largest of all known RNA viruses** (27 – 32 kb). Previous studies on Coronavirus structures have focused on **several conserved regions that are important for viral replication**. For several of these regions, such as the **5' UTR**, the **3' UTR**, and the **frameshift element** (FSE), structures have been predicted computationally with supportive experimental data from RNAse probing and nuclear magnetic resonance (NMR) spectroscopy.
… [truncated to 6 sentences]

Data-driven Training & Inferences

Data-centric

1) Inferences (gusses) are based on parameters learned from previous data (posterior probability)
2) Extracts crucial parameters from each learning cluster, which can summarize that cluster's articles into coherent summaries
3) Biases and variances in data clusters are directly reflected in that learning step's parameters

Data Assumptions

Bayesian treatment requires past data to contain good summaries

1) Summaries with simple words (that represent the critical information of the literature)
2) Literature contains detailed explanations on references to any other literature
3) Many articles fall under common clusters (for model to train cluster after cluster in online EWC manner)