Centro de Información de COVID (CIC): Charlas científicas relámpago

Transcripción de una presentación de Jane Pan (Universidad de Princeton), 22 de septiembre de 2021

Título: Detección de contradicciones en ensayos aleatorizados controlados de COVID-19 a través de modelos de lenguaje BERT

Grabación de YouTube con diapositivas

Información del seminario web del CIC de Septiembre 2021

Editora de la transcripción: Macy Moujabber

Traductora: Isabella Graham Martínez

Transcripción

Lauren Close:

Diapositiva 1

Rápidamente pasamos a nuestro próximo orador, Jane Pan. Jane es en realidad una de las tres ganadoras de nuestro desafío inaugural de CIC [COVID Information Commons] para estudiantes de pregrado. Es una ganadora de primer lugar. El desafío se llevó a cabo a principios de esta primavera, y damos la bienvenida a Jane y estamos muy emocionados de compartir su investigación con la comunidad más amplia del CIC. Así que, Jane, adelante.

Jane Pan:

Diapositiva 2

Genial, déjame compartir mi pantalla rápidamente. Espero que eso funcione. ¿No funciona? De acuerdo, genial, genial. Así que, espero que todos hayan tenido un gran comienzo en otoño. Mi nombre es Jane. Me gradué de Columbia la primavera pasada y ahora estoy en Princeton en la escuela de posgrado. Estoy muy feliz de presentar el trabajo que hice durante mi último año de licenciatura con el profesor Chunhua Weng del Instituto de Ciencia de Datos de la Universidad de Columbia. Nuestro proyecto investiga la detección de contradicciones en estudios controlados aleatorizados COVID-19 utilizando modelos de lenguaje masivo como BERT.

Diapositiva 3

Primero un poco de contexto. Los resultados contradictorios en los estudios clínicos han sido un problema de larga data para académicos, investigadores y médicos por igual, especialmente en un campo como las publicaciones de alto volumen. Un estudio encontró que un tercio de los estudios clínicos originales son desafiados o incapaces de ser replicados, y otro encontró que un cuarto o ensayos aleatorizados controlados, en particular, son abiertamente contradecidos por conclusiones posteriores. Y este es un tema que se hizo especialmente tangible durante el

brote de COVID-19. Todos hemos oído hablar de la hidroxicloroquina cuyos estudios clínicos iniciales eran realmente optimistas, y más tarde los hallazgos fueron contradecidos de manera bastante decisiva. Por lo tanto, analizar e interpretar los resultados de un cuerpo de trabajo grande y continuamente cambiante es un desafío que es realmente importante en escenarios sensibles al tiempo como la pandemia global. Así que, para nosotros, facilitar el proceso de identificar estudios contradictorios o concordantes sería realmente crucial para los científicos que podrían querer, por ejemplo, realizar revisiones sistemáticas, identificar lo que podría causar resultados diferentes entre dos estudios, evaluar la veracidad de una afirmación de investigación, y caracterizar el estado de consenso o madurez en una cuestión de investigación particular. Y así, para nuestro proyecto de investigación, la pregunta que nos planteamos fue: ¿cómo podemos extraer sistemáticamente el conocimiento basado en la evidencia del texto en bruto para identificar rápida y automáticamente qué estudios están de acuerdo y en desacuerdo?

Diapositiva 4

Entonces, formulamos este problema como una inferencia estándar del lenguaje natural, o tarea de NLI, y la afirmación o el objetivo es clasificar un par de oraciones como contradictorias, que impliquen, o acepten y den a entender de manera neutral que las alegaciones de condena no están relacionadas o no se implican ni se contradicen entre sí. Entonces, el objetivo del modelo de lenguaje aquí formalmente recibe un par de oraciones x1 y x2 con un símbolo de clasificación de masa CLS y una matriz de parámetros. Elegimos una etiqueta que maximiza la probabilidad de que el estado final de CLS sea esa etiqueta para esa x específica. Y elegimos modelos de lenguaje masivo aquí, específicamente BERT, porque estos históricamente han tenido un desempeño muy fuerte de las tareas de NLI. Utilizamos modelos preexistentes pre-entrenados como modelo base para nuestros proyectos. El objetivo es utilizar el aprendizaje de transferencia adaptando estos modelos a nuestra tarea específica de NLI y consideramos tres modelos básicos. El primero es el modelo genérico BERT que está pre-entrenado en BookCorpus y Wikipedia y luego dos modelos específicos de dominio. BioBERT que está pre-entrenado en resúmenes y artículos PubMed y ClinicalBERT que está pre-entrenado en notas clínicas MIMIC 3.

Diapositiva 5

Así que, para nosotros la consideración más crucial era la rapidez con que el modelo se adaptaría a las nuevas preguntas de investigación, porque en la práctica querrías que el modelo encontrará contradicciones en la nueva investigación sobre las que no podría estar pre-entrenados. Así que, para ese fin sabíamos que necesitábamos un conjunto de datos con áreas de investigación y preguntas que nunca antes habían sido vistas por los modelos de base. Y así, creamos nuestro propio conjunto de datos. Anotamos manualmente un nuevo conjunto de datos utilizando LitCOVID, una base de datos pública de artículos PubMed COVID-19. Esto se debe a que COVID-19 es muy reciente y no habría estado presente en los datos utilizados para pre-entrenar los modelos base. Y en línea con otros métodos de anotación de informes analíticos biomédicos, identificamos 15 preguntas de investigación separadas y 103 estudios que las contestaron. Así, extrajimos manualmente una frase de cada resumen que aborda directamente la pregunta de investigación, y luego dos anotadores independientes etiquetaron manualmente los pares como contradicción, vinculación o neutral con respecto a la pregunta de investigación. Así que, cualquier etiqueta que tuviera conclusiones discrepantes fue descartada.

Diapositiva 6

Para construir nuestro modelo, agregamos capas de clasificación no iniciadas a los modelos base y afinamos. Mantenemos los parámetros de la capa base congelados por ahora, ya que tenemos un conjunto de trenes relativamente pequeño y sólo afinamos las capas de clasificación. Para nuestro conjunto de trenes utilizamos ManConCorpus, un corpus de inferencia médica NLI anotado públicamente muy similar a lo que hicimos, pero más amplio, no sólo COVID-19.

Diapositiva 7

Y también reservamos una pequeña porción, alrededor de 20 del conjunto de datos LitCOVID para entrenamiento. Y fuimos muy cautelosos sobre la prevención de la contaminación entre la prueba y el tren porque el modelo tiene que generalizar a preguntas que no ha visto antes. Así que, con ese fin, eliminamos cualquier par que mezclara las frases de prueba y entrenamiento. Entonces, lo que eso significa es que si una pregunta de investigación aparece en el conjunto de trenes, no aparecerá en el conjunto de pruebas y viceversa. Mostramos los pares de frases, y para cada modelo base entrenamos dos modelos: uno que agregó esa pequeña porción de datos LitCOVID a su tren y otro que solo usó ManConCorpus. Y lo hacemos porque queremos ver cuánto mejora el modelo con solo una porción muy pequeña de datos específicos de COVID-19 añadidos a él.

Diapositiva 8

Así que aquí están los resultados de nuestras métricas de clasificación de todos los modelos, BioBERT y Clinical BERT con datos LitCOVID funciona mejor, lo que tiene sentido porque los modelos base están entrenados en un dominio similar a LitCOVID, y como era de esperar, si se añaden datos de tren LitCOVID, funciona mejor que un modelo que no lo hace. Pero me gustaría señalar la mejora–como la mejora muy drástica con una proporción muy pequeña de datos de entrenamiento COVID. Por lo tanto, las puntuaciones F están mejorando en gran medida. Casi todos ellos son como el doble con la excepción de la columna de precisión, que es bastante fuerte en todos los modelos.

Diapositiva 9

Aquí mostramos el recuerdo clase por clase y vemos algunos patrones bastante interesantes aquí. Hasta ahora, la contradicción es una clase que funciona mejor incluso con modelos que no tenían ningún dato de entrenamiento de datos LitCOVID agregado, y suponemos que esto puede ser porque negar términos como no o no son universales en todos los temas y dominios, así que tal vez el modelo pueda identificar las negaciones bastante rápido. Vimos que los datos de entrenamiento LitCOVID mejoran principalmente las predicciones neutrales. Pueden ver que es como duplicar el número de predicciones neutras correctas, lo cual es muy probable porque ahora que tiene datos de entrenamiento de LitCOVID sabe qué palabras COVID no necesariamente sugieren contradicción o vinculación. Y el derecho es relativamente débil en general, a excepción de BioBERT con LitCOVID, y creemos que esto podría ser porque 'BioBERT pre-entrenamiento corpora' realmente vino de PubMed, así que puede haber aprendido características que ayudaron a identificar mejor la afirmación o negación textual en un dominio biomédico.

Diapositiva 10

Así que, en resumen, tenemos pruebas sólidas que demuestran que los modelos BERT son un enfoque válido para la detección de conjugación en el dominio biomédico. Tenemos tres modelos pre-entrenados que necesitan sólo una pequeña cantidad de datos de entrenamiento para mejorar drásticamente el rendimiento. Y sólo un poco de análisis de errores muy brevemente. Algunos patrones comunes que encontramos fueron, como vieron, antes luchando con la identificación de términos mutuos y luego vimos cierta confusión con abreviaturas o terminología médica. Por ejemplo, HCQ e hidroxicloroquina, el modelo no sabe inmediatamente que son lo mismo, así que dirá que es neutral o no relacionado, y así sucesivamente.

Diapositiva 11

Así que, para el futuro, algunas de las preguntas interesantes que pensamos que podrían ser contestadas son, como, ¿cómo podemos seleccionar automáticamente la mejor oración sin necesidad de extraerla manualmente? Y ya hay algunas herramientas de nominación de textos para estudios clínicos disponibles abiertamente, como Trialstreamer o Weng Labs picoparser, así que me interesaría ver cómo podrían integrar esto con una herramienta de detección de conjugación. Y también, tenemos curiosidad por saber si podemos mejorar el rendimiento del modelo mediante, ya sabes, el suministro de una lista proporcionada por el usuario de acrónimos o sinónimos para ese dominio que el modelo es probable que se encuentre.

Es todo lo que tengo por hoy. Me gustaría concluir agradeciendo al profesor Weng y al Dr. Hao Liu por su tutoría y ayuda a lo largo de mi investigación y también un gran agradecimiento a Stan y Marguerite por su amable asistencia con el proyecto. Gracias por su tiempo y espero que todos tengan una gran semana.