COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Jane Pan (Princeton University), September 22, 2021

Title: Contradiction Detection of COVID-19 Randomized Controlled Trials via BERT Language Models

YouTube Recording with Slides

September 2021 CIC Webinar Information

Transcript Editor: Shikhar Johri

Transcript

लॉरेन क्लोज़:

स्लाइड *1*

जेन वास्तव में हमारे उद्घाटन सीआईसी [COVID सूचना कॉमन्स] स्नातक छात्र पेपर चुनौती के तीन विजेताओं में से एक है। वह पहले स्थान की विजेता है। चुनौती इस वसंत से पहले आयोजित की गई थी, और हम जेन का स्वागत करते हैं और व्यापक सीआईसी समुदाय के साथ अपने शोध को साझा करने के लिए बहुत उत्साहित हैं। तो, जेन, इसे दूर ले जाओ।

जेन पैन:

स्लाइड *2*

बहुत बढ़िया, ठीक है, मुझे अपनी स्क्रीन को वास्तविक रूप से जल्दी साझा करने दें। उम्मीद है कि यह काम कर रहा है। काम नहीं कर रहा है? ठीक है, बहुत बढ़िया ठीक है। इसलिए, मुझे आशा है कि आप सभी की गिरावट की शानदार शुरुआत रही होगी। मेरा नाम जेन है। मैंने पिछले वसंत में कोलंबिया से स्नातक की उपाधि प्राप्त की और अब मैं ग्रेजुएट स्कूल के लिए प्रिंसटन में हूं। कोलंबिया विश्वविद्यालय के डेटा साइंस इंस्टीट्यूट के प्रोफेसर चुनहुआ वेंग के साथ अंडरग्रेजुएट के अपने वरिष्ठ वर्ष के दौरान मैंने जो काम किया था, उसे प्रस्तुत करने में मुझे बहुत खुशी हो रही है। हमारी परियोजना BERT जैसे बड़े पैमाने पर भाषा मॉडल का उपयोग करके COVID-19 यादृच्छिक नियंत्रित अध्ययनों के विरोधाभास का पता लगाने की जांच करती है।

स्लाइड *3*

तो, पहले थोड़ा सा संदर्भ। नैदानिक अध्ययनों में विरोधाभासी परिणाम शिक्षाविदों, शोधकर्ताओं और डॉक्टरों के लिए समान रूप से एक लंबे समय से चली आ रही समस्या रही है, खासकर उच्च मात्रा वाले प्रकाशनों जैसे क्षेत्र में। एक अध्ययन में पाया गया कि मूल नैदानिक अध्ययनों का एक तिहाई या तो

चुनौती दी जाती है या दोहराया जाने में असमर्थ होता है, और एक अन्य पाया जाता है कि एक चौथा या यादृच्छिक नियंत्रित परीक्षण, विशेष रूप से, बाद के निष्कर्षों से पूरी तरह से विरोधाभासी हैं। और यह एक ऐसा मुद्दा है जो COVID-19 के प्रकोप के दौरान विशेष रूप से मूर्त हो गया। हम सभी ने हाइड्रोक्सीक्लोरोक्वीन के बारे में सुना है जिनके प्रारंभिक नैदानिक अध्ययन वास्तव में आशावादी थे, और बाद में निष्कर्षों को बहुत निर्णायक रूप से विरोधाभास दिया गया था। और इसलिए काम के एक बड़े और लगातार बदलते शरीर से परिणामों का विश्लेषण और व्याख्या करना एक चुनौती है जो वैश्विक महामारी जैसे समय-संवेदनशील परिदृश्यों के दौरान वास्तव में महत्वपूर्ण है। इसलिए, हमारे लिए, विरोधाभासी या सहमत अध्ययनों की पहचान करने की प्रक्रिया को सुविधाजनक बनाना उन वैज्ञानिकों के लिए वास्तव में महत्वपूर्ण होगा, जो उदाहरण के लिए, व्यवस्थित समीक्षा करना चाहते हैं, यह पहचानें कि दो अध्ययनों के बीच अलग-अलग परिणाम क्या हो सकते हैं, एक शोध दावे की सत्यता का मूल्यांकन करें, और किसी विशेष शोध प्रश्न पर आम सहमति या परिपक्वता की स्थिति को चिह्नित करें। और इसलिए, हमारी शोध परियोजना के लिए, हमने खुद से जो सवाल पूछा वह यह था कि हम अकेले कच्चे पाठ से साक्ष्य-आधारित ज्ञान को कैसे व्यवस्थित रूप से निकाल सकते हैं ताकि जल्दी और स्वचालित रूप से पहचान सकें कि कौन से अध्ययन सहमत हैं और असहमत हैं?

स्लाइड *4*

इसलिए, हम इस समस्या को एक मानक प्राकृतिक भाषा अनुमान, या एनएलआई कार्य के रूप में तैयार करते हैं, और दावा या उद्देश्य वाक्यों की एक जोड़ी को विरोधाभासी, प्रवेश, या सहमत और तटस्थ अर्थ के रूप में वर्गीकृत करना है कि वाक्य के दावे असंबंधित हैं या वे न तो एक दूसरे में प्रवेश करते हैं और न ही विरोधाभास करते हैं। तो, यहां भाषा मॉडल उद्देश्य को औपचारिक रूप से कुछ द्रव्यमान वर्गीकरण टोकन सीएलएस और कुछ पैरामीटर मैट्रिक्स के साथ वाक्यों की एक जोड़ी एक्स 1 और एक्स 2 दी गई है। हम एक लेबल चुनते हैं जो इस संभावना को अधिकतम करता है कि सीएलएस की अंतिम स्थिति उस विशिष्ट एक्स के लिए वह लेबल है। और हम यहां बड़े पैमाने पर भाषा मॉडल चुनते हैं, विशेष रूप से बीईटी, क्योंकि इनमें ऐतिहासिक रूप से एनएलआई कार्यों का बहुत मजबूत प्रदर्शन रहा है। हम अपनी परियोजनाओं के लिए आधार मॉडल के रूप में पहले से मौजूद पूर्व-प्रशिक्षित मॉडल का उपयोग करते हैं। लक्ष्य इन मॉडलों को हमारे विशिष्ट एनएलआई कार्य के अनुकूल बनाकर स्थानांतरण सीखने का उपयोग करना है और हम तीन आधार मॉडल पर विचार करते हैं। पहला जेनेरिक BERT मॉडल है जो BookCorpus और Wikipedia पर पूर्व-प्रशिक्षित है और फिर दो डोमेन-विशिष्ट मॉडल हैं। BioBERT जो PubMed सार और लेखों पर पूर्व-प्रशिक्षित है और ClinicalBERT जो MIMIC 3 नैदानिक नोटों पर पूर्व-प्रशिक्षित है।

स्लाइड *5*

इसलिए, हमारे लिए सबसे महत्वपूर्ण विचार यह था कि मॉडल कितनी तेजी से नए शोध प्रश्नों के अनुकूल होगा, क्योंकि व्यवहार में आप चाहते हैं कि मॉडल नए शोध में विरोधाभासों को ढूंढे, जिस पर इसे पूर्व-प्रशिक्षित नहीं किया जा सकता है। इसलिए, उस अंत तक हमें पता था कि हमें अनुसंधान क्षेत्रों और प्रश्नों के साथ एक डेटा सेट की आवश्यकता है जो आधार मॉडल द्वारा पहले कभी नहीं देखा गया था। और इसलिए, हमने अपना डेटा सेट बनाया। हमने मैन्युअल रूप से LitCOVID का उपयोग करके एक उपन्यास डेटा सेट को एनोटेट किया, जो COVID-19 PubMed लेखों का सार्वजनिक रूप से उपलब्ध डेटाबेस है। ऐसा इसलिए है क्योंकि COVID-19 बहुत हाल ही में है और यह बेस मॉडल को प्री-ट्रेन करने के लिए उपयोग किए जाने वाले डेटा में मौजूद नहीं होता। और अन्य बायोमेडिकल विश्लेषणात्मक रिपोर्ट एनोटेशन विधियों के अनुरूप, हमने 15 अलग-अलग शोध प्रश्नों और 103 अध्ययनों की पहचान की जिन्होंने उनका उत्तर दिया। इसलिए, हम मैन्युअल रूप से प्रत्येक सार से एक वाक्य निकालते हैं जो सीधे शोध प्रश्न को संबोधित करता है, और फिर दो स्वतंत्र एनोटेटर मैन्युअल रूप से जोड़े को विरोधाभास,

प्रवेश, या शोध प्रश्न के संबंध में तटस्थ के रूप में लेबल करते हैं। इसलिए, असहमत निष्कर्षों वाले किसी भी लेबल को फेंक दिया गया था।

स्लाइड *6*

हमारे मॉडल का निर्माण करने के लिए, हम आधार मॉडल और ठीक धुन के लिए uninitialized वर्गीकरण परतों को जोड़ने. हम अभी के लिए जमी हुई बेस लेयर पैरामीटर रखते हैं, क्योंकि हमारे पास अपेक्षाकृत छोटा ट्रेन सेट है और हम केवल वर्गीकरण परतों को ठीक करते हैं। हमारे ट्रेन सेट के लिए हम मैनकॉनकॉर्पस का उपयोग करते हैं, जो सार्वजनिक रूप से उपलब्ध मैन्युअल रूप से एनोटेट किए गए चिकित्सा अनुमान एनएलआई कॉर्पस के समान है, जो हमने किया था, लेकिन अधिक व्यापक, न केवल COVID-19।

स्लाइड *7*

और हम प्रशिक्षण के लिए LitCOVID डेटा सेट का लगभग 20 हिस्सा भी आरक्षित करते हैं। और हम परीक्षण और ट्रेन के बीच संदूषण को रोकने के बारे में वास्तव में सतर्क थे क्योंकि मॉडल को उन सवालों का सामान्यीकरण करना होगा जो उसने पहले नहीं देखे हैं। इसलिए, उस अंत तक, हमने किसी भी जोड़े को हटा दिया जो परीक्षण और ट्रेन वाक्यों को मिलाते थे। तो, इसका मतलब यह है कि यदि ट्रेन सेट में एक शोध प्रश्न दिखाई देता है, तो यह परीक्षण सेट में दिखाई नहीं देगा और इसके विपरीत। हम वाक्य जोड़े को टोकन करते हैं, और प्रत्येक आधार मॉडल के लिए हमने दो मॉडलों को प्रशिक्षित किया: एक जिसने लिटकोविड डेटा के उस छोटे हिस्से को अपनी ट्रेन में जोड़ा और एक जिसने केवल मैनकॉनकॉर्पस का उपयोग किया। और हम ऐसा इसलिए करते हैं क्योंकि हम यह देखना चाहते हैं कि इसमें जोड़े गए COVID-19 विशिष्ट डेटा के बहुत छोटे हिस्से के साथ मॉडल में कितना सुधार होता है।

स्लाइड *8*

तो यहां सभी मॉडलों के हमारे वर्गीकरण मीट्रिक के परिणाम दिए गए हैं, BioBERT और Clinical BERT with LitCOVID डेटा सबसे अच्छा प्रदर्शन करता है, जो समझ में आता है क्योंकि बेस मॉडल LitCOVID के समान डोमेन पर प्रशिक्षित होते हैं, और आश्चर्यजनक रूप से, यदि आप LitCOVID ट्रेन डेटा जोड़ते हैं, तो यह उस मॉडल से बेहतर प्रदर्शन करता है जो नहीं करता है। लेकिन मैं COVID प्रशिक्षण डेटा के बहुत कम अनुपात के साथ बहुत कठोर सुधार जैसे सुधार पर ध्यान देना चाहूंगा। इसलिए, एफ स्कोर में काफी हद तक सुधार हो रहा है। उनमें से लगभग सभी सटीक कॉलम के अपवाद के साथ दोहरीकरण के करीब हैं, जो सभी मॉडलों में बहुत मजबूत है।

स्लाइड *9*

यहाँ हम कक्षा के आधार पर एक वर्ग पर याद दिखाने के लिए और हम यहाँ कुछ बहुत दिलचस्प पैटर्न देखते हैं. अब तक और दूर विरोधाभास एक ऐसा वर्ग है जो उन मॉडलों के साथ भी सबसे अच्छा प्रदर्शन करता है जिनके पास कोई लिटकोविड डेटा प्रशिक्षण डेटा नहीं जोड़ा गया था, और हम अनुमान लगाते हैं कि ऐसा इसलिए हो सकता है क्योंकि विषयों और डोमेन में नहीं या नहीं जैसे शब्दों को नकारना सार्वभौमिक है, इसलिए शायद मॉडल बहुत जल्दी नकारात्मकता की पहचान कर सकता है। हमने देखा कि LitCOVID प्रशिक्षण डेटा मुख्य रूप से तटस्थ भविष्यवाणियों में सुधार करता है। आप देख सकते हैं कि यह सही तटस्थ भविष्यवाणियों की संख्या को दोगुना करने जैसा है, जो सबसे अधिक संभावना है क्योंकि यह- अब जब इसके पास लिटकोविड प्रशिक्षण डेटा है, तो यह जानता है कि कौन से COVID शब्द जरूरी नहीं हैं, जैसे, विरोधाभास या प्रवेश का सुझाव दें। और LitCOVID के साथ BioBERT को छोड़कर पात्रता समग्र रूप से अपेक्षाकृत कमजोर है, और हमें लगता है कि ऐसा इसलिए हो सकता है क्योंकि BioBERT

प्री-ट्रेनिंग कॉर्पोरा वास्तव में PubMed से आया था, इसलिए इसने ऐसी विशेषताएं सीखी होंगी जो बायोमेडिकल डोमेन में पाठ्य पुष्टि या नकारात्मकता की बेहतर पहचान करने में मदद करती हैं।

स्लाइड *10*

इसलिए, संक्षेप में, हमारे पास यह दिखाने के लिए मजबूत सबूत हैं कि बायोमेडिकल डोमेन में संयुग्मन का पता लगाने के लिए BERT मॉडल एक वैध दृष्टिकोण है। हमारे पास तीन पूर्व-प्रशिक्षित मॉडल हैं जिन्हें काफी बेहतर प्रदर्शन के लिए केवल थोड़ी मात्रा में प्रशिक्षण डेटा की आवश्यकता होती है। और बस कुछ त्रुटि विश्लेषण बहुत संक्षेप में। कुछ सामान्य पैटर्न जो हमने पाए, जैसा कि आपने देखा, पहले आपसी शब्दों की पहचान करने के साथ संघर्ष कर रहे थे और फिर हमने संक्षिप्त या चिकित्सा शब्दावली जैसे कुछ भ्रम देखे। इसलिए, उदाहरण के लिए एचसीक्यू और हाइड्रोक्सीक्लोरोक्वीन, मॉडल को तुरंत पता नहीं चलता है कि वे एक ही चीज़ हैं, इसलिए यह कहेगा कि यह तटस्थ या असंबंधित है, और इसी तरह आगे।

स्लाइड *11*

तो, भविष्य के लिए, कुछ दिलचस्प प्रश्न जो हमें लगता है कि उत्तर दिए जा सकते हैं, जैसे, हम इसे मैन्युअल रूप से निकालने की आवश्यकता के बिना स्वचालित रूप से सर्वश्रेष्ठ वाक्य का चयन कैसे कर सकते हैं? और नैदानिक अध्ययन के लिए पहले से ही कुछ पाठ नामांकन उपकरण हैं, जैसे, खुले तौर पर उपलब्ध हैं जैसे कि ट्रायलस्ट्रीमर या वेंग लैब्स पिकोपार्सर, इसलिए मुझे यह देखने में दिलचस्पी होगी कि वे इसे संयुग्मन पहचान उपकरण के साथ कैसे एकीकृत कर सकते हैं। और साथ ही, हम यह जानने के लिए उत्सुक हैं कि क्या हम मॉडल के प्रदर्शन में सुधार कर सकते हैं, आप जानते हैं, उस डोमेन के लिए उपयोगकर्ता द्वारा प्रदान की गई शब्दकोष या समानार्थक शब्द की सूची की आपूर्ति करना, जो मॉडल के सामने आने की संभावना है।

आज के लिए मेरे पास बस इतना ही है। मैं प्रोफेसर वेंग और डॉ हाओ लियू को उनके परामर्श के लिए धन्यवाद देकर समाप्त करना चाहता हूं और मेरे शोध में मदद करना चाहता हूं और परियोजना के साथ उनकी तरह की सहायता के लिए स्टेन और मार्गुराइट को भी एक बड़ा धन्यवाद देना चाहता हूं। आपके समय के लिए धन्यवाद और मुझे आशा है कि आप सभी के पास एक महान सप्ताह होगा।