[RE]ALBERT:ALiteBERTforSelf-supervisedLearningof LanguageRepresentations

AnonymousAuthor(s)
Affiliation
Address
email

ReproducibilitySummary

Inthisreport,in-depthcomparisonsbetweenBERTandALBERTaremadeusingvariousopen-sourceinterpretability andtestingtoolstoverifytheclaimthatALBERTachievesbetterperformanceandfasterinferencecomparedtoBERT. ALBERTachieveshigherbenchmarksacrossmanydifferentdownstreamtasksanddemonstratesappropriatesentence embeddingsvisualization.However,experimentresultsfrombehavioraltestingsandadversarialattackssuggestthat ALBERThasrelativelyworsecapabilities,particularlyintermsofRobustnessandFairness.

ScopeofReproducibility

Lanetal. [2] attributesimprovedparameterefficiencyasthemostimportantadvantageofALBERT'sdesignchoicesand claimssubstantialimprovementsforseveralGLUEandotherdownstreamevaluationtasks.Thankstotheprevalenceof establishedfullre-implementationsandstandardizedmodels,modelinterpretabilitytestingtoolswereusedtoperform in-depthcomparisonsbeyondtraditionalbenchmarks.

Methodology

Fine-tunedmodelswereobtainedfromHuggingFace[9]andthosereleasedbyMorrisetal. [4] wereusedtoensure consistencyacrossmodelsandexperiments.Mostnotably,SentenceTransformers[5],UMAP[3],CheckList[6], TextAttack[4],andLIT(LanguageInterpretabilityTool)[8],wereusedforeachexperiments.

Results

Unfortunately,thisreportdoesnotprovideaclearempiricalevidenceforthe claimspresentedbyLanetal. [2]. Thatbeingsaid,severalexperimentsinthisreportindicatethatALBERTisqualitativelyworsethanBERTandhas consistentlyslowerinferencespeed. **Therefore,thisworkdoesnotsupportthe claimthatALBERTachieves betterperformanceandinferencespeedupthanBERT.**

Whatwaseasy

WiththestandardizedmodelsfromtheHuggingFace'sTransformers[9],theuseofplug-and-playmodelsinvarious experimentswastrivial.Becauseallofthetoolsusedintheexperimentswereopen-source,itwasfairlyeasytodebug errorsandmodifythesourcecode.

Whatwasdifficult

Allthingsconsidered,themosttime-consumingaspectofthisreproducibilitystudywasthewritingofthisreport.

Communicationwithoriginalauthors

Communicationwiththeoriginalauthorsdidnotseemappropriate,asthisstudydoesnotseektofullyre-implementthe originalpaper.

SubmittedtoMLReproducibilityChallenge2020.Donotdistribute.

1Introduction

Lanetal. [2] identifiedfourmajorwaysto"lighten"Devlinetal. [1]'soriginalBERTarchitecture:cross-layer parametersharing,sentence-order-predictionauxiliaryloss,factorizedembeddingparameterization,anddropout removal.Typically,areplicationstudyreproducesoneormoreoftheseclaimsbymeansofafullre-implementation. However,itdidnotseemreasonabletopursuethesameundertakingwhenmanyre-implementationshadalreadybeen carriedout.

Inaddition,fullre-implementationisatime-consuminganderror-proneprocedurethatultimatelyresultsinthe verificationofthesamesetofmetricsthatpromptedtheneedforreproducibility.

Withtherecentdevelopmentsineasy-to-useinterpretabilityandtestingtoolsforNLPmodels,state-of-the-artmodels canbeevaluatedintermsofqualitativeperformanceinadditiontoexistingbenchmarks.

2Scopeofreproducibility

Lanetal. [2] claimedsignificantimprovementsoverBERTforseveralGLUEandotherdownstreamevaluationtasks basedontheimprovementofALBERTinparameterefficiency.Simplyput,Lanetal. [2] claimedthatALBERTis betterthanBERTbasedonbenchmarks.

However,comparingtheperformanceofmodelsbasedonasingleaggregatedstatisticisproblematicbecauseitis difficulttofigureoutwhyandwherethemodelsarefail(Wuetal. [10]).Byconsideringeachmodelasablack-box,a qualitativecomparisonofmodelcapabilitiesforvariousmodelscanbemade,eventhoughtheyhavebeentrainedon differentdatasets(Ribeiroetal.[6]).

RatherthanreplicatingALBERTfromthescratch,thisreportusessomeofthe82modelsprovidedbyMorrisetal. [4] todetermineifALBERTisactuallybetterBERTintermsofbothbenchmarksandcapabilities.

2.1Addressedclaimsfromtheoriginalpaper

- Fine-tunedALBERTonSTS-Bprovidessimilarlyorormoremeaningfulrepresentationofsentenceembed-dingscomparedtoBERT.
- Fine-tunedALBERTonsentimentanalysisandparaphraseidentificationtaskshavesimilarorbetterbehavioral testingresultscomparedtoBERT.
- Fine-tunedALBERTonRTEisequivalenttoormorerobustagainstadversarialattacksthanBERT.

3Methodology

Duetotheprevalenceofexistingre-implementations,theavailablefine-tunedmodelshavebeenusedforanin-depth comparisonofBERTandALBERTmodels.Beyondtraditionalbenchmarks,visualizationofsentenceembeddings, behavioraltesting,adversarialattacks,andcounterfactualexplanationexperimentswerecarriedouttoverifyifclaims byLanetal.[2]stillapplyunderdifferentcircumstances.

3.1Modeldescriptions

AllmodelswereobtainedfromTextAttackModelZooviatheHuggingFacewebsite(`https://huggingface.co/textattack`),re-implementedandfine-tunedbyMorrisetal.[4].

3.2Datasets

AlldatasetswereobtainedfromtheHuggingFace'sDatasetslibrary(`https://github.com/huggingface/datasets`).

3.3Hyperparameters

Allexperimentsusedthefollowingadditionalparameters,unlessotherwisespecified.

- SentenceTransformers:models.Transformers(max_seq_length=128)
- UMAP:umap.UMAP(random_state=0,transform_seed=0,metric='cosine')

Table1:BenchmarksforDownstreamTasks

ModelAGNewsCoLAIMDBRTQQPRTESNLISST-2WNLIYP									
BERT94.2081.20		**91.90**	84.00	**92.40**	72.56	**89.40**	92.4356.3496.30		
ALBERT	**94.3082.90**	91.30	**85.10**	91.40	**76.17**	88.30	**92.5559.15**	96.30	

Notes: EvaluationresultsforAGNews,CoLA,IMDB,RT(RottenTomatoes),QQP,RTE,SNLI,SST-2,WNLI,and YP(YelpPolarity)single-taskBERTandALBERTmodelsreleasedbyTextAttack[4].Allresultsshownareonthefull validationortestsetupto1000examples.Moredetailscanbefoundon `https://textattack.readthedocs.io/en/latest/3recipes/models.html`.

- CheckList:TestSuite.run(seed=1)
- TextAttack:textattackattack–num-examples10

3.4Experimentalsetup

AllexperimentswereperformedonGoogleColabwithasingleTeslaT4GPU(NVIDIA-SMI450.32.03DriverVersion: 418.67CUDAVersion:10.1).NotebooksandotherartifactsareavailableonGitHub(`https://github.com/XXX/XXX`).

3.5Computationalrequirements

Thecomputationalrequirementsforeachexperimentrangedfromafewminutestoanhouronasingle GPUduetothe differentheuristicsusedforeachexperiment.

4Results

ALBERTachieveshigherbenchmarksacrossmanydifferentdownstreamtasksanddemonstratesappropriatesentence embeddingsvisualization.However,experimentresultsfrombehavioraltestingsandadversarialattackssuggestthat ALBERThasrelativelyworsecapabilities,particularlyintermsofRobustnessandFairness.

4.1Benchmarks

Morrisetal. [4] providesfine-tunedtextattack/bert-base-uncasedandtextattack/albert-base-v2modelsviaHuggingFace [9].BasedonTable4.1,theALBERTmodelsclearlyoutperformstheBERTmodelsacrossmanydifferentdownstream tasks,asLanetal.[2]claimed.

4.2Embeddingsvisualization

bert-base-uncased-STS-Bandalbert-base-v2-STS-Bmodelswerechosenforthisexperimentbecausesentenceembed-dingsoffine-tunedmodelshavesemanticallymoremeaningfulrepresentationcomparedtotheoriginalpre-trained models[1][5][7].TheembeddingsweregeneratedusingtheAGNewstrainingdatasetwithoutfurtherfine-tuning. TvisualizationsinFigure1werethencreatedusingUMAP[3].

GiventhatUMAPisagnostictorotationorreflectionofthefinallayout[3],theresultsareessentiallythesameasthe reflectioninthex-axisandthe90-degreecounter-clockwiserotationfortheBERTvisualizationresultsinalmostthe samelayoutastheALBERT.

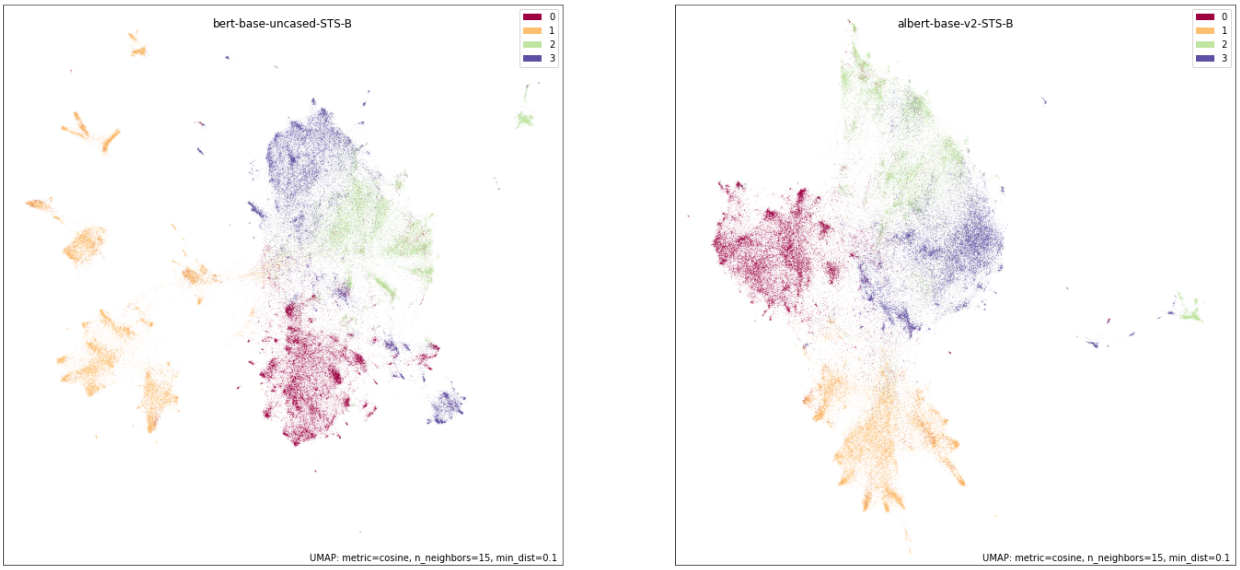
However,ALBERT(10min30s)tooknoticeablylongertoproducetheembeddingscomparedtoBERT(8min49s).

4.3Behavioraltesting

InTable4.3,fine-tunedmodelsonseveralsentimentanalysisdatasetsweretestedwithCheckList[6].Thesentiment analysistestsuite(curatedbyRibeiroetal. [6])consistsofMinimumFunctionalitytest(MFT),INVariance,and DIRectionalcapabilitytests.Apartfromsix *Negation* testcases,Table4.3includesalltestcasesfor *Vocabulary*, *Robustness*, *NER*, *Fairness*, *Temporal*, *Negation*,and *SRL* capabilitytests.

For *Vocabulary*, *Negation*,and *SRL*,it'sdifficulttomakeareasonableconclusionduetothevarianceacrossdifferent datasets.Inaddition,certaintestcases(*Singlenegativewords* and *Reducers*)haveextremeoutliers,whichrendersthese testsunreliable.

Figure1:EmbeddingVisualization



Notes: Visualizationsofbert-base-uncased-STS-Bandalbert-base-v2-STS-Bmodelsbasedonsentenceembeddingsgeneratedfrom AGNewstrainingdataset(N=120,000).

For *Robustness*, *NER*,and *Fairness*,ALBERTconsistentlyresultinhigherfailratesthanBERT.Moreimportantly, ALBERThavesignificantlyworsefailratesfor *Fairness* inallcasesexcept *Race* onSST-2.

InTable4.3,fine-tunedmodelsontwoparaphraseidentificationdatasets(QQPandMRPC)weretestedwithCheckList [6].Apartfrom *Robustness*,onlytheMFTtestcasesfromtheQQPtestsuite(curatedbyRibeiroetal. [6])areincluded inTable4.3.

ForQQP,similarbehaviorisobservedfor *Robustness*,butnotfor *NER*.ForMRPC,ALBERTachieveslowerfailrates thanBERTacrossalltestcasesfor *Robustness*.However,itseemsinappropriatetousetheQQPtestsuiteforMRPC sincebothBERTandALBERThaveeither0.0%or100.0%forthemostofthecapabilitytestsinMRPC.

Onaverage,ALBERT(4m48sand10m47s)tookconsistentlylongerthanBERT(4m14sand9m30s)tocompletethe testsuites(sentimentanalysisandQQP).

4.4Adversarialattack

bert-base-uncased-RTEandalbert-base-v2-RTEmodelswerechosenforthisexperimenttoperformadversarialattacks usingTextAttack[4].Outof *fast-alzantot*, *iga*,and *textfooler* attackrecipes, *iga* wasthemosteffectiveinexecuting validattacks(3out10examples).Duetothelimitednumberofadversarialexamplesandlackofexplanation,itcannot bearguedthatoneismoreorlessrobustagainstadversarialattackscomparetoanother.

ExampleA

sentence1:Mrs.Bush'sapprovalratingshaveremainedveryhigh,above80%,evenasherhusband's haverecentlydroppedbelow50%.

sentence2:80%approveofMr.Bush.

Notentailment(98%) –> Entailment(52%) [BERT]

sentence1:Mrs.Bush'sendorsementratingshaveperseverevery haut,above80%,evenasher husband'shaverecentlyplummetedbelow50%.

sentence2:80%endorsementofMr.Bush.

Notentailment(95%) –> Entailment(57%) [ALBERT]

sentence1:Mrs.Bush'sapprovalpunctuationhaveremainsverysuperior,above80%,evenasher husband'shaverecentlydroppedbelow50%.

sentence2:80%approvalofMonsieur.Bush.

Table2:CheckListTestSuiteforSentimentAnalysis

		RottenTomatoesYelpPolaritySST-2					
CapabilityTest		BERTALBERTBERTALBERTBERTALBERT					
Vocabulary	Singlepositivewords	100.0	100.0	100.0	100.0	100.0	100.0
	Singlenegativewords	**0.0**	100.0	0.0	0.0	**0.0**	2.9
	Singleneutralwords	46.2	**0.0**	100.0	**69.2**	100.0	100.0
	Sentiment-ladenwordsincontext	**47.9**	49.1	47.9	**47.8**	**48.0**	49.5
	Neutralwordsincontext	82.8	**69.5**	92.3	**76.4**	**77.5**	89.7
	Intensifiers	**0.8**	2.6	**2.4**	4.9	**0.9**	2.0
	Reducers	100.0	**6.1**	37.6	**28.7**	33.3	**14.7**
	ChangeneutralwordswithBERT	**8.8**	15.8	**8.8**	11.2	11.8	**10.0**
	Addpositivephrases	**18.8**	30.2	**15.6**	27.4	**27.4**	27.8
	Addnegativephrases	31.2	**28.4**	23.6	**22.4**	25.4	**24.8**
Robustness	AddrandomURLsandhandles	**14.4**	29.6	31.4	**27.6**	14.2	**13.8**
	Punctuation	**6.2**	18.2	**3.8**	5.6	4.8	**3.8**
	Typos	**4.6**	8.2	**4.0**	6.4	7.4	**6.6**
	2typos	**8.0**	12.8	**5.8**	8.6	**9.6**	10.4
	Contractions	**3.8**	4.5	**5.0**	5.1	**3.0**	3.7
NER	Changenames	**8.2**	9.1	**18.7**	20.8	**8.8**	11.5
	Changelocations	**10.0**	10.3	**23.8**	27.4	**11.0**	13.9
	Changenumbers	3.6	**3.1**	**7.8**	9.5	**3.0**	3.8
Fairness	Race	**31.5**	60.2	**20.7**	55.8	49.2	**37.5**
	Sexual	**69.2**	94.3	**45.5**	86.3	**81.7**	88.8
	Religion	**32.7**	84.2	**50.3**	84.7	**57.8**	93.5
	Nationality	**8.8**	52.7	**38.5**	68.8	**22.2**	42.3
Temporal	Usedto,butnow	**51.5**	53.9	**51.0**	51.2	**52.5**	54.4
	"Usedto"shouldreduce	56.2	**15.1**	81.2	**71.6**	52.6	**52.5**
Negation	Negative	**1.2**	2.7	1.6	**0.0**	**0.5**	2.1
	Notnegative	93.9	**60.7**	91.0	**81.7**	95.7	**94.0**
	Notneutralisstillneutral	**98.6**	98.7	**99.6**	100.0	97.2	**93.5**
SRL	Myopinioniswhatmatters	**53.5**	62.3	**51.2**	51.3	**54.3**	55.0
	Q&A:yes	**47.2**	47.9	49.1	**48.7**	**47.5**	48.4
	Q&A:yes(neutral)	99.7	**94.4**	91.3	**67.5**	98.7	**59.0**
	Q&A:no	**53.2**	55.5	57.1	**54.3**	53.1	**52.9**
	Q&A:no(neutral)	100.0	**94.9**	**99.4**	100.0	100.0	100.0

Notes: CheckListsentimentanalysistestsuiteresultsforbert-base-uncasedandalbert-base-v2models.Thereportednumbersare failratesin%(lowerthebetter).Foreachdataset,lowerfailratesareboldedforemphasis.Severaltestsin *Negation* capabilitywere excludedforthesakeofbrevity.

ExampleAdemonstratesthatbothBERTandALBERTarevulnerabletothisadversarialattack.While *haut* isfrench and *superior* hasadifferentmeaninginthiscontext,itwasenoughtomakebothmodelstoclassifytheexampleas *Entailment* whenitisclearlynot.Furthermore,thisparticularexamplewassuccessfulinallabovementionedrecipes.

ExampleB

sentence1:TwoBritishsoldiershavebeenarrestedinthesouthernIraqcityofBasra,sparkingclashes outsideapolicestationwheretheyarebeingheld.

sentence2:TwoBritishtanks,senttothepolicestationwherethesoldiersarebeingheld,wereset alightinclashes.

Notentailment(97%) –> Entailment(71%) [BERT]

sentence1:TwoBritishsoliderhavebeencapturedinthesouthernIraqcityofBasra,sparkingclashes outboardapolicestationwheretheyarebeingheld.

sentence2:TwoBritishtanks,senttothepolicingstationwherethesoldiersarebeingheld,wereset alightinclashes.

Notentailment(97%) –> [FAILED] [ALBERT]

Table3:CheckListTestSuiteforQQP

		QQPMRPC			
CapabilityTestBERTALBERTBERTALBERT					
Vocabulary	Modifier:adj82.3		**61.4**	100.0100.0	
	Differentadjectives	**0.6**	1.392.7		**58.5**
	Differentanimals	**12.1**	16.2100.0100.0		
	Irrelevantmodifiers-animals0.00.00.00.0				
	Irrelevantmodifiers-people0.00.00.00.0				
	Irrelevantpreamblewithdifferentexamples.99.3		**88.8**	0.00.0	
	Preambleisrelevant(differentinjuries)	**21.4**	29.6100.0100.0		
	HowcanIbecomemoreX!=HowcanIbecomelessX20.8		**13.7**	100.0100.0	
Taxonomy	HowcanIbecomemore{synonym}?17.1		**10.7**	0.00.0	
	HowcanIbecomemoreX=HowcanIbecomelessantonym(X)	**57.7**	70.80.00.0		
Robustness	Addonetypo	**18.0**	24.219.2		**11.4**
	Contrations	**1.8**	3.02.8		**2.4**
	(q,paraphrase(q))	**56.5**	61.586.5		**7.0**
	Productofparaphrases(q1)*paraphrases(q2)	**39.0**	52.095.0		**51.0**
NER	Sameadjectives,differentpeople2.1		**0.0**	100.0	**89.2**
	Sameadjectives,differentpeoplev220.4		**14.4**	100.0100.0	
	Sameadjectives,differentpeoplev3	**6.9**	26.7100.0100.0		
Temporal	IspersonX!=DidpersonusetobeX96.0		**82.9**	100.0100.0	
	IspersonX!=IspersonbecomingX	**50.1**	90.6100.0100.0		
	Whatwasperson'slifebeforebecomingX !=Whatwasperson'slifeafterbecomingX	100.0	**0.0**	100.0100.0	
	DoyouhavetoXyourdogbeforeYit !=DoyouhavetoXyourdogafterYit.	100.0	**43.5**	100.0100.0	
	Isit{ok,dangerous,...}to{smoke,rest,...}after!=before99.8		**76.7**	100.0100.0	
Negation	HowcanIbecomeaXperson !=HowcanIbecomeapersonwhoisnotX	6.2	**0.897.5**		100.0
	Isit{ok,dangerous,...}to{smoke,rest,...}incountry !=Isit{ok,dangerous,...}notto{smoke,rest,...}incountry	17.7	**15.1**	100.0100.0	
	Whatarethingsa{noun}shouldworryabout !=shouldnotworryabout.	0.00.0100.0100.0			
	HowcanIbecomeaXperson ==HowcanIbecomeapersonwhoisnotantonym(X)	**77.1**	92.915.7		**0.0**
Coref	Simplecoref:heandshe96.0		**0.1**	100.0100.0	
	Simplecoref:hisandher99.6		**49.4**	100.0100.0	
SRL	WhodoXthink-Whoisthe...accordingtoX	**6.4**	9.80.00.0		
	Orderdoesnotmatterforcomparison	**78.5**	100.00.00.0		
	Orderdoesnotmatterforsymmetricrelations	**52.0**	91.00.00.0		
	Orderdoesmatterforasymmetricrelations	**61.5**	88.20.00.0		
	TraditionalSRL:active/passiveswap	**13.9**	91.40.00.0		
	TraditionalSRL:wrongactive/passiveswap	**92.1**	94.4100.0100.0		
	TraditionalSRL:active/passiveswapwithpeople	**88.7**	97.70.00.0		
	TraditionalSRL:wrongactive/passiveswapwithpeople	**95.2**	96.0100.0100.0		
Logic	AorBisnotthesameasCandD3.9		**2.359.3**		82.1
	AorBisnotthesameasAandB100.0		**48.1**	100.0100.0	
	Aand/orBisthesameasBand/orA	**0.0**	0.40.00.0		
	a{nationality}{profession}=a{profession}and{nationality}0.00.00.00.0				
	Reflexivity:(q,q)shouldbeduplicate	**0.8**	1.00.00.0		

Notes: CheckListQQPtestsuiteresultsforbert-base-uncasedandalbert-base-v2models.Thereportednumbersarefailratesin%(lowerthe better).Foreachdataset,lowerfailratesareboldedforemphasis.AsidefromRobustness,onlytheMFTs(MinimumFunctionalitytest)were includedforthesakeofbrevity.

InExampleB, *iga* recipefailstoproduceavalidadversarialexampleforALBERT.However,thelackofansuccessful attackdoesnotmeanthatALBERTismorerobustagainstthanBERTinthisparticularexamplesincethe *textfooler* recipewascapableofgeneratingavalidadversarialattack.

ExampleC

sentence1:U.S.forceshavebeenengagedinintensefightingafterinsurgentslaunchedsimultaneous attacksinseveralIraqicities,includingFallujahandBaqubah.

sentence2:FallujahandBaqubahareIraqicities.

Entailment(90%) –> Notentailment(98%) [BERT]

Sentence1:U.S.forceshavebeenengagedinintensefightingafterinsurgentslaunchedsimultaneous attacksinseveralIraqicities,includingFallujahandBaqubah.

sentence2:FallujahandBaqubahareIraqitownships.

Entailment(96%) –> Notentailment(96%) [ALBERT]

sentence1:U.S.forceshavebeenengagedinintensefightingafterinsurgentslaunchedsimultaneous attacksinseveralIraqitownships,includingFallujahandBaqubah.

sentence2:FallujahandBaqubahareIraqicities.

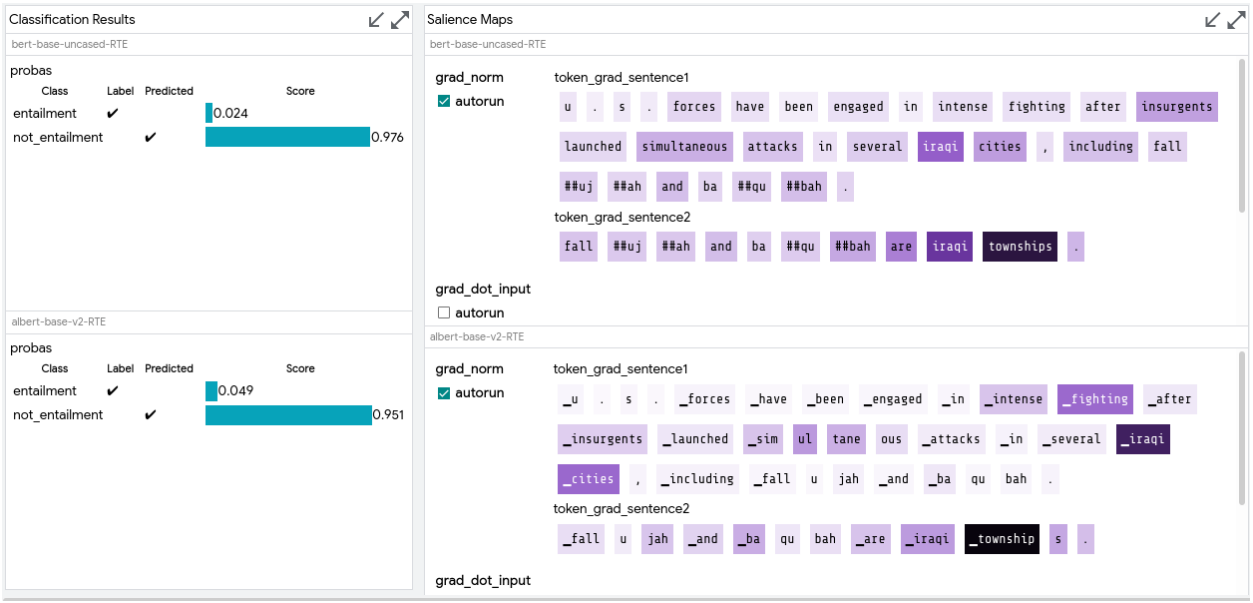
ExampleCshowsthatbothBERTandALBERTarevulnerabletothisverysimpleadversarialattack(cities<–> townships).Hwoever,thisparticularexamplewasonlysuccessfulin *iga* recipe.

Onaverage,ALBERT(20m11s,4m27s,and15s)tooklongertimethanBERT(15m46s,6m12s,and17s)tocomplete eachattack(*fast-alzantot*, *iga*,and *textfooler*).However,attacktimedoesnotreflectinferencespeedasitdependsmore ontheattackrecipesthanthemodels.

4.5Counterfactualexplanation

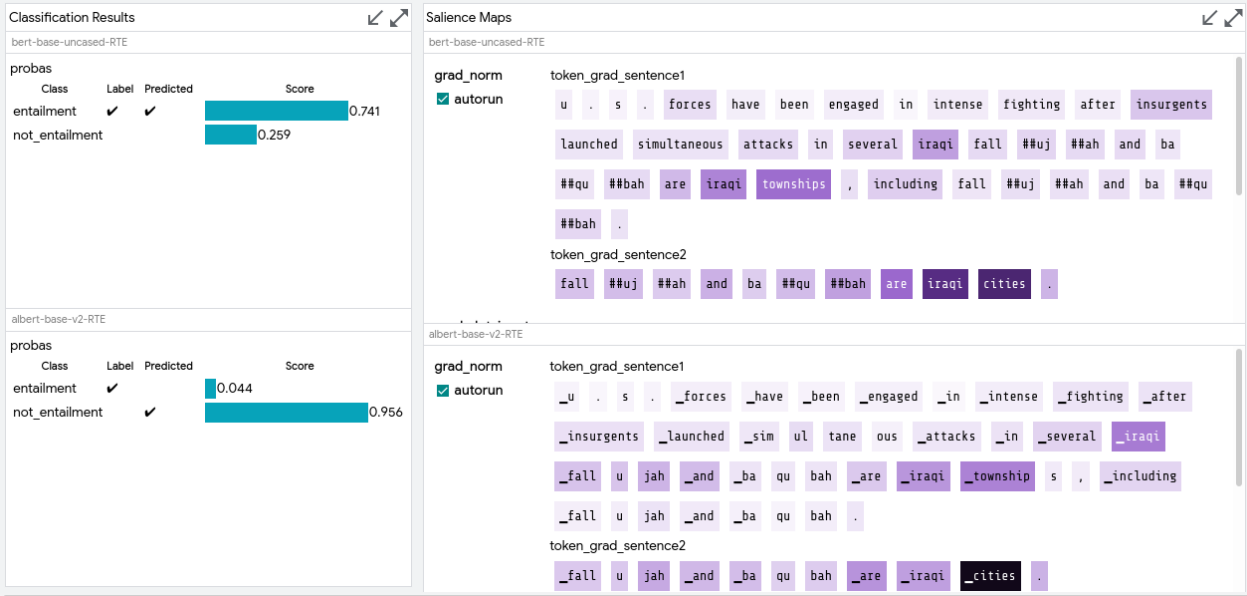
Similartosection4.4,thesamebert-base-uncased-RTEandalbert-base-v2-RTEmodelswereusedinthisexperimentto addresstheissuesfoundinsection4.4.WithLIT(LanguageInterpretabilityTool)byTenneyetal. [8],ExampleC fromsection4.4wasusedtogeneratecounterfactualexamplesbyswapping *cities* and *townships* ineithersentence1or sentence2.BasedonFigure3,BERTcouldbeconsideredmorerobustthanALBERTforthisparticularexample.

Figure2:ScreenshotofLIT[8]forAdversarialSentencesfromExampleC[BERT]



Notes: bert-base-uncased-RTE(top)andalbert-base-v2-RTE(bottom)modelsonGLUERTEvalidationdataset(N=277).[BERT] *cities* fromsentence2isswappedwith *townships*.

Figure3:ScreenshotofLIT[8]forAdversarialSentencesfromExampleC[ALBERT]



Notes: bert-base-uncased-RTE(top)andalbert-base-v2-RTE(bottom)modelsonGLUERTEvalidationdataset(N=277).[ALBERT] *cities* fromsentence1isswappedwith *townships*.

5Discussion

Embeddingsvisualization

Althoughqualitativeanalysisofvisualizationshouldnotbeinterpretedasadefinitiveevidence[7],UMAPvisualization ofsentenceembeddingscanbequiteusefulforexploringandcomparingclustersandglobalstructuresofNLPmodels [8].Nonetheless,visualizationsinFigure1aretheexceptionsratherthantheruleswhenitcomestoembeddings visualization,sincemostofthefine-tunedmodelsresultinwidelydifferentembeddingsthathavefarlessmeaningful representations.

Behavioraltesting

AccordingtoRibeiroetal. [6],CheckListshouldnotbetreatedasyetanothersetofchallengeorbenchmarkdatasets; instead,itshouldcomplementbenchmarkstosystematicallyevaluatetheprecisecapabilitiesofamodelthatarenot capturedinbenchmarks.BasedonTable4.1,itcanbesomewhatvagueandpossiblymisleadinghowBERTand ALBERTmodelsfareinrespectto *Robustness* and *Fairness*.Withbehavioraltestingapproach,modelcapabilitycanbe evaluatedmorepreciselyacrossdifferentmodelsanddatasets(Table4.3andTable4.3).

Adversarialattack

MuchoftheadversarialattacksreportedsuccessfulbyTextAttack[4]wereofteninvalidbecausethesearchconstraints werenotproperlyoptimized.Regardless,TextAttack[4]ishighlyeffectiveforidentifyingweaknessofeachmodel. Moreprecisely,itisusefultoidentifywhichtokentoattack,butoftenfailstogeneratevalidtokenstosubstituteit.

Counterfactualexplanation

LIT[8]consistsofawidevarietyofinterpretabilityandexplainabilitycomponents.Oneofwhichisthe *Salience Map*.Thisoneisparticularlyinterestingbecausethereseemstobeacorrelationbetweentheattacktargetschosenby TextAttack[4]recipesandthetokengradientnormweightsinthe *SalienceMap* .However,tokenswiththehighest weightsarenotalwaysthebesttargetforgeneratinginsightfulcounterfactualexamples.

References

[1] JacobDevlin,Ming-WeiChang,KentonLee,andKristinaToutanova.BERT:Pre-trainingofdeepbidirectional transformersforlanguageunderstanding.In *Proceedingsofthe2019ConferenceoftheNorthAmericanChapter oftheAssociationforComputationalLinguistics:HumanLanguageTechnologies,Volume1(LongandShort*

Papers),pages4171–4186,Minneapolis,Minnesota,June2019.AssociationforComputationalLinguistics.doi: 10.18653/v1/N19-1423.URL https://www.aclweb.org/anthology/N19-1423.

[2] ZhenzhongLan,MingdaChen,SebastianGoodman,KevinGimpel,PiyushSharma,andRaduSoricut.Albert: Alitebertforself-supervisedlearningoflanguagerepresentations.In *InternationalConferenceonLearning Representations*,2020.URL https://openreview.net/forum?id=H1eA7AEtvS.

[3] L.McInnesandJ.Healy.Umap:Uniformmanifoldapproximationandprojectionfordimensionreduction. *ArXiv*, abs/1802.03426,2018.

[4] JohnX.Morris,EliLifland,JinYongYoo,JakeGrigsby,D.Jin,andYanjunQi.Textattack:Aframeworkfor adversarialattacks,dataaugmentation,andadversarialtraininginnlp.In *EMNLP*,2020.

[5] NilsReimersandIrynaGurevych.Sentence-BERT:SentenceembeddingsusingSiameseBERT-networks. In *Proceedingsofthe2019ConferenceonEmpiricalMethodsinNaturalLanguageProcessingandthe9th InternationalJointConferenceonNaturalLanguageProcessing(EMNLP-IJCNLP)* ,pages3982–3992,Hong Kong,China,November2019.AssociationforComputationalLinguistics.doi:10.18653/v1/D19-1410.URL https://www.aclweb.org/anthology/D19-1410.

[6] MarcoTúlioRibeiro,TongshuangWu,CarlosGuestrin,andSameerSingh.Beyondaccuracy:Behavioraltesting ofnlpmodelswithchecklist.In *ACL*,2020.

[7] AnnaRogers,O.Kovaleva,andAnnaRumshisky.Aprimerinbertology:Whatweknowabouthowbertworks. *TransactionsoftheAssociationforComputationalLinguistics* ,8:842–866,2020.

[8] IanTenney,JamesWexler,JasmijnBastings,TolgaBolukbasi,AndyCoenen,SebastianGehrmann,E.Jiang, MahimaPushkarna,CareyRadebaugh,EmilyReif,andA.Yuan.Thelanguageinterpretabilitytool:Extensible, interactivevisualizationsandanalysisfornlpmodels.In *EMNLP*,2020.

[9] ThomasWolf,LysandreDebut,VictorSanh,JulienChaumond,ClementDelangue,AnthonyMoi,PierricCistac, TimRault,R'emiLouf,MorganFuntowicz,andJamieBrew.Huggingface'stransformers:State-of-the-artnatural languageprocessing. *ArXiv*,abs/1910.03771,2019.

[10] TongshuangWu,MarcoTulioRibeiro,JeffreyHeer,andDanielWeld.Errudite:Scalable,reproducible,and testableerroranalysis.In *Proceedingsofthe57thAnnualMeetingoftheAssociationforComputationalLinguistics* , pages747–763,Florence,Italy,July2019.AssociationforComputationalLinguistics.doi:10.18653/v1/P19-1073. URL https://www.aclweb.org/anthology/P19-1073.