UMass · CS685 | Advanced Natural Language Processing (2020)

CS685 (2020)· 课程资料包 @ShowMeAI

视频
中英双语字幕

课件
一键打包下载

笔记
官方笔记翻译

代码
作业项目解析

视频 · B 站 [扫码或点击链接]
https://www.bilibili.com/video/BV1BL411t7RV

课件 & 代码 · 博客 [扫码或点击链接]
http://blog.showmeai.tech/umass-cs685

Awesome AI Courses Notes Cheatsheets 是 ShowMeAI 资料库的分支系列，覆盖最具知名度的 **TOP50+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets· 持续更新中			
知识图谱	图机器学习	深度强化学习	自动驾驶
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094

微信公众号

资料下载方式 2：扫码点击底部菜单栏称为 **AI 内容创作者？** 回复 [添砖加瓦]

Today: improvements on BERT

1. training improvements ⇒ RoBERTa
   more data
2. longer sequences
   TransformerXL (XLNet)
3. more efficient pretraining objectives
   ELECTRA
4. smaller models
   ALBERT

RoBERTa: very simple
Collection of modifications

1. train w/ bigger batches
   ↳ smaller # of batches w/ larger batch size
   ↳ gradient accumulation to bypass GPU mem. limitations
2. no next sentence prediction
   ↳ downstream perf. unaffected
   ↳ [CLS] token gets no pretraining
3. pretrain on more data

↳ 16 GB → 160 GB
↳ common crawl, URLs from Reddit

4. pretrain for longer (more batches/epochs)
↳ 500K steps

TransformerXL

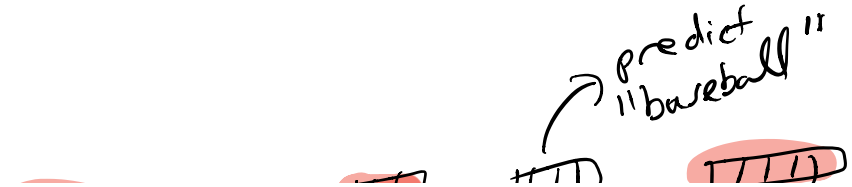
BERT has a fixed token limit of 512 for its inputs. How can we model longer sequences?

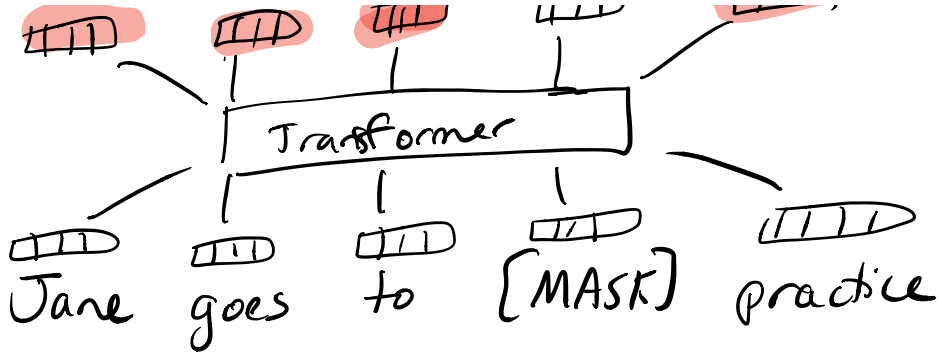
↳ idea: add a recurrent mechanism that connects adjacent segments

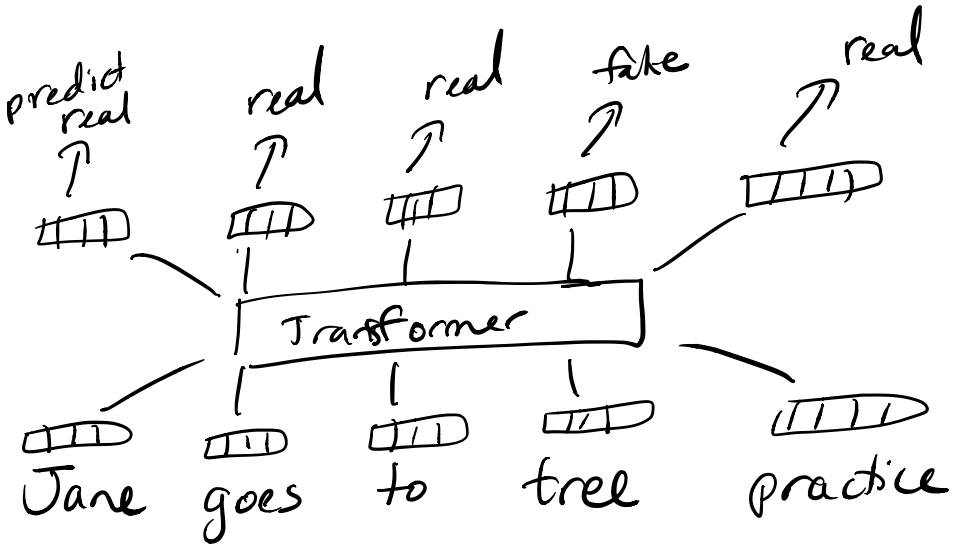
↳ no gradient flow to previous segments
hidden states from prev. segment are cached

→ practical limit to this extended context window
→ 900 words for TransformerXL

ELECTRA – cheaper obj. fn







↳ how do I decide which words to replace and with what?

↳ "generator" ⇒ coming up w/ fake words

↳ train a small BERT model

Jane goes to [MASK] practice

↓

football
basketball
baseball

↳ sampled words from generator form fake words for ELECTRA

- every single token is associated with a prediction of real/fake, not just 15% of words as in BERT

ALBERT - more params != better model

- cross-layer param sharing
  - Q, K, V projection matrices } shared across all layers
  - W matrices in FF layers }

BERT-large: 334M params

ALBERT-large: 18M params

↳ what if we make our shared set of params bigger?

ALBERT-XXL: 235M params, 4096d hidden state size

↳ outperforms BERT-large

UMass · CS685 | Advanced Natural Language Processing (2020)

CS685 (2020)· 课程资料包 @ShowMeAI

视频
中英双语字幕

课件
一键打包下载

笔记
官方笔记翻译

代码
作业项目解析

视频 · B 站 [扫码或点击链接]
https://www.bilibili.com/video/BV1BL411t7RV

课件 & 代码 · 博客 [扫码或点击链接]
http://blog.showmeai.tech/umass-cs685

Awesome AI Courses Notes Cheatsheets 是 ShowMeAI 资料库的分支系列，覆盖最具知名度的 **TOP50+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets· 持续更新中			
知识图谱	图机器学习	深度强化学习	自动驾驶
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094

微信公众号

资料下载方式 2：扫码点击底部菜单栏称为 **AI 内容创作者？** 回复 [添砖加瓦]