

Extract Summary Finetune

[Full View](#)

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
Question & Improvement | Extract Your News

作者&小组成员

杨开创

赵心研

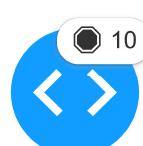
王媛

樊宇鑫

项目简介

[Introduction](#)

SWUFE-BERT 大型金融语言模型将在包含 1700 万条金融新闻和机构研报的超过 200G 语料库中进行预训练和微调，构建金融领域词典，提供一套面向金融场景的端到端自然语言处理框架，并开源以辅助金融学术研究和行业应用，填补该领域的空白。为了测试SWUFE-BERT预训练模型的效果，我们开启了篇章级抽取式自动文本摘要的下游任务实验。在此基础上，我们将 SWUFE-BERT 与 Google 原生中文 BERT、哈工大讯飞实验室开源的 RoBERTa-wwm-ext-large进行比较实验。该实验的具体内容是对比三种BERT预训练模型在自动生成金融短讯抽取式摘要上的表现。该实验的具体过程有：准备金融短讯有监督语料1万条（5000条为混有非金融相关内容的新闻短讯、5000条为高质量金融相关的新闻短讯）、采用三种BERT模型分别对语料进行预训练（形成Word embedding）、采用LSTM和TransformerEncoder的深度学习模型进行摘要抽取的训练和预测、对比训练效果和预测结果。



Extract Summary Finetune

[Full View](#)

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
Question & Improvement | Extract Your News

问题重述

面对呈指数级增长的网络文本资源，如何快速且准确地从中提取出重要的内容，是一个迫切而有意义的需求。自动文本摘要技术旨在利用计算机强大的计算能力，从较长文本中提炼出关键信息，生成简洁、通顺和凝练的摘要，以帮助用户快速全面地了解文本关键信息。由于不同学科领域词语表达、组合各有特色，现有预训练BERT模型在不同领域任务中表现也不尽相同。聚焦于财经领域，在当前数字金融蓬勃兴起的背景下，非结构化的文本数据在舆情预测、金融风控等领域发挥了独特的作用。

BERT 是由 Google 在 2018 年开发的大型自然语言处理模型，拥有强大的语言表征能力和特征提取能力。目前开源的 BERT 模型大多适用于通用语言领域，在金融文本的分支赛道上少有建树。现有的开源金融 BERT 模型存在使用的语料库较小、文本来源较乱、在金融场景中广泛应用存在局限性等问题。

CN_BERT

谷歌官方发布的BERT-base, Chinese，以每种语言的整个 Wikipedia 转储数据作为每种语言的训练数据。该模型使用的 WordPiece 的分词方式会把一个完整的词切分成若干个子词，在生成训练样本时，这些被分开的子词会随机被 mask。这种分词方式没有考虑到传统 NLP 中的中文分词。

WWM_BERT

wmm-bert 使用中文维基百科（包括简体和繁体）进行训练，并且使用了哈工大 LTP 作为分词工具，即对组成同一个词的汉字全部进行 Mask，是谷歌在 2019 年 5 月 31 日发布的 一项 BERT 的升级版本，主要更改了原预训练阶段的训练样本生成策略。在全词 Mask 中，如果一个完整的词的部分 WordPiece 字词被 mask，则同属该词的其他部分也会被 mask。

目前金融领域对文本数据的处理还缺乏成熟的技术工具，因此我们基于包含 1500 万条金融新闻和机构研报的超过 200G 语料库的大型金融语言模型 SWUFE-BERT，进行抽取式新闻摘要任务的 Fine-Tuning，期望对于金融文本处理的分支赛道上有更好的表现。

SWUFE_BERT

金融领域大型通用语言模型 SWUFE-BERT，在包含 1500 万条金融新闻和机构研报的超过 200G 语料库中进行预训练，采用 Transformer 模型构架与 Whole Word Mask 中文分词掩码技术。与现有 BERT 模型在通用语料库中训练不同，SWUFE-BERT 通过在大量高质量财经语料库中训练，且结合 Whole Word Mask 中文分词掩码技术，使其吸收了 WWM-BERT 的优势同时也弥补了其他现有 BERT 模型在财经领域的不足。

Extract Summary Finetune

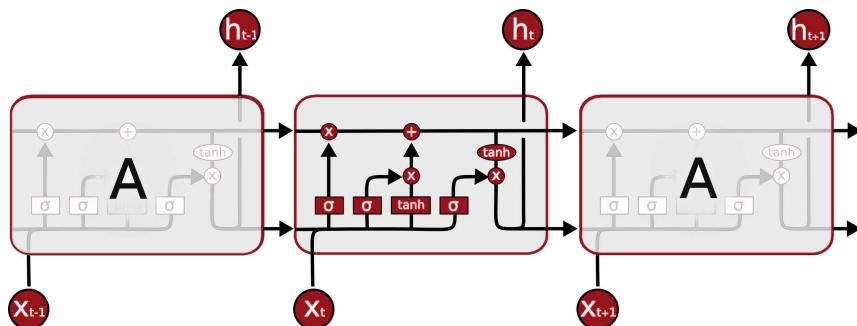
Full View

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
 Question & Improvement | Extract Your News

LSTM的选择与搭建

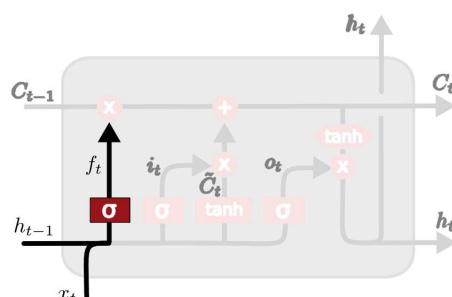
RNN是用于处理序列数据的深度神经网络，该神经网络会对前面的信息进行记忆并应用于当前输出的计算中。但RNN模型如果需要实现长期记忆则在每次计算时需与前n次计算相结合，使计算量呈指数式增长，因此RNN不适合进行长期记忆计算。

LSTM是一种特殊的RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题，相比普通的RNN，LSTM能够在更长的序列中有更好的表现。



Forget Gate

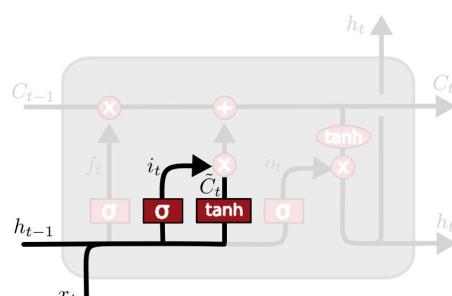
输入的 h_{t-1} , x_t 连接起来后通过 σ 函数将每个元素映射到0到1之间，得到 f_t ，与输入的 C_{t-1} 相乘，当 f_t 的某一位的值为 0 时， C_{t-1} 中对应位置的数据也为零，对应的信息即被遗忘。



$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Input Gate

与遗忘门类似， σ 函数决定我们要保留和更新哪些信息，得到输入门的输出， \tan 层创建新的候选值向量。

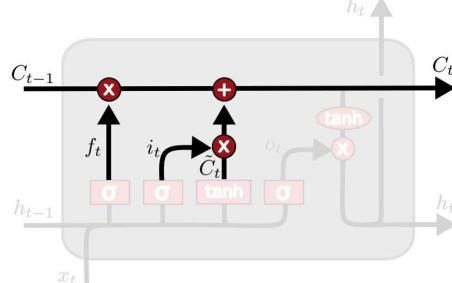


$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

New Cell State

细胞旧状态 C_{t-1} 与遗忘门中得到的 f_t 相乘，再与输入门中的两个计算结果的乘积相加，得到更新后的细胞状态。

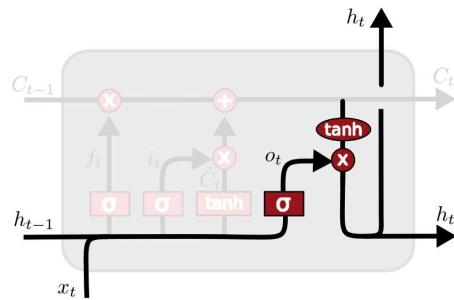


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



Output Gate

通过sigmoid层得输出门的输出，再与经tanh处理后的新的细胞状态 C_t 相乘，得到输出的新的隐藏细胞状态 h_t 。



$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

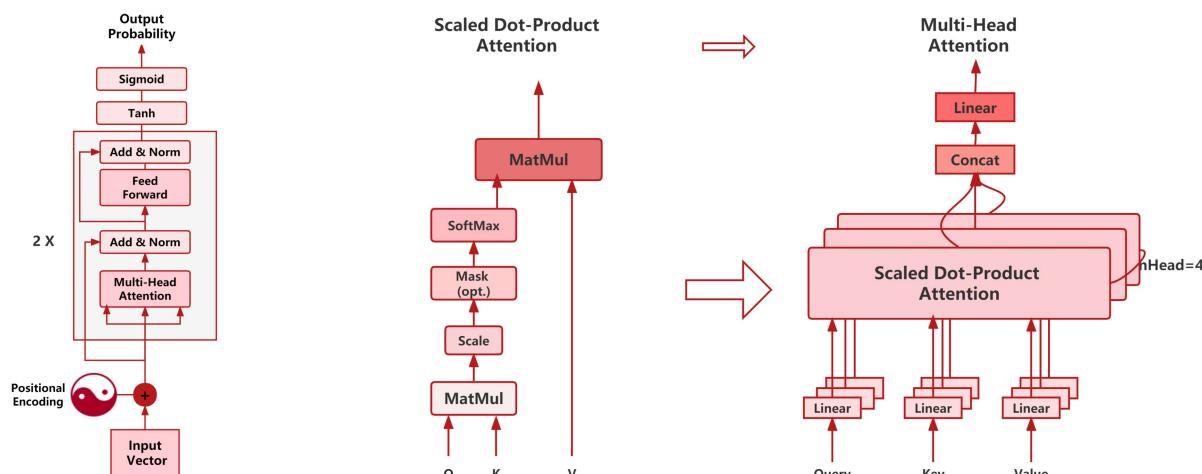
Extract Summary Finetune

Full View

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
 Question & Improvement | Extract Your News

Transformer Encoder

Transformer模型抛弃了传统的CNN和RNN，只使用注意力机制，以其优秀的表现成为机器学习领域热点，而翻译任务与摘要任务极为相似，我们有理由相信Transformer在文本摘要任务中也一定能够持续稳定输出。不同于论文中Transformer 使用6层encoder-decoder结构，在本次实训中经过多方面考量，我们采用至多2层encoder，由于本阶段不涉及decoder，所以我们不设置decoder。且本次实训直接提供了输入向量，所以不涉及embedding的部分。我们选取vrelu函数作为激活函数，对输出先使用Tanh函数处理，再使用Sigmoid函数得到介于0, 1之间的概率值，如图为采用的Transformer模型结构图。



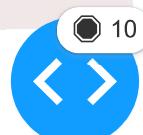
Muti-Head Attention

Transformer的注意力机制与人类的注意力机制相似，由于注意力有限，它将注意力资源更多的投入到目标区域，使用多头注意力机制（本次实训使用4头）从不同方面提取v特征。上右图是注意力机制的图解。

单个注意力机制用数学公式可以简单表达为（其中softmax除以根号下Q，K点乘的方差dk是为了防止梯度消失）：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Positional Encoding



没有采用RNN的Transformer没有捕捉序列信息的功能，它不能分清到底是“我很开心”还是“开心很我”，所以在输入词向量后还需要进行Positional Encoding，位置编码公式如下：

$$PE_{(pos, 2i)} = \sin(pos / 1000^{2i/d_{\text{model}}})$$

$$PE_{(pos, 2i+1)} = \cos(pos / 1000^{2i/d_{\text{model}}})$$

其中 pos 代表目前token在序列的位置， d_{model} 代表模型的维度也就是 input size 。



Extract Summary Finetune

[Full View](#)

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
 Question & Improvement | Extract Your News

评价指标选择

BECLoss函数：常用来处理二分类问题的损失函数。

AUC-ROC：在对不同模型进行比较时，若不同模型的ROC曲线发生交叉，则很难比较哪个效果更好，此时比较更为合理，即AUC的大小。 ROC曲线是基于混淆矩阵得出的，曲线越靠近左上角，模型的准确性越高。

参数选择与指标对比

序号	HIDDEN_SIZE	BATCH_SIZE	LR	DROP_RATE	LAYERS	学习率衰减	train_loss	test_loss
01	768	50	0.001	0	1	否	0.3461	0.3121
02	32	50	0.005	0.3	2	否	0.3985	0.3281
03	384	32	0.005	0	1	否	0.3515	0.3154
04	384	50	0.001	0	1	是	0.3653	0.3363
05	64	50	0.005	0.1	2	是	0.3745	0.3383

Hidden_size较大（大于等于input_size的一倍）及layers较大（大于等于2层）时模型训练速度太慢，数值较小时训练结果不理想。因此，我们结合运行时间和训练结果筛选出以下两组较优参数组。

参数组	HIDDEN_SIZE	BATCH_SIZE	LR	DROP_RATE	LAYERS	学习率衰减	test_auc	预训练模型
参数组1	384	50	0.001	0	1	是	0.8571	cn_bert
	384	50	0.001	0	1	是	0.8679	wwm_bert
	384	50	0.001	0	1	是	0.8755	swufe_bert
参数组2	64	50	0.005	0.1	2	是	0.8594	cn_bert
	64	50	0.005	0.1	2	是	0.863	wwm_bert
	64	50	0.005	0.1	2	是	0.8623	swufe_bert

Extract Summary Finetune

[Full View](#)

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
Question & Improvement | Extract Your News

Visualization

在这一节中, 基于通用的loss和auc score指标, 我们对抽取式摘要任务在下述不同情景中的表现进行比较:

1. 不同Bert预训练情景CN/WWM/SWUFE
2. 不同Network训练情景Transformer/LSTM
3. 不同Loss Function情景 等权重(未加权)BCEloss/ 加权BCEloss

Bar Charts

SWUFE bert

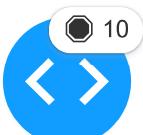
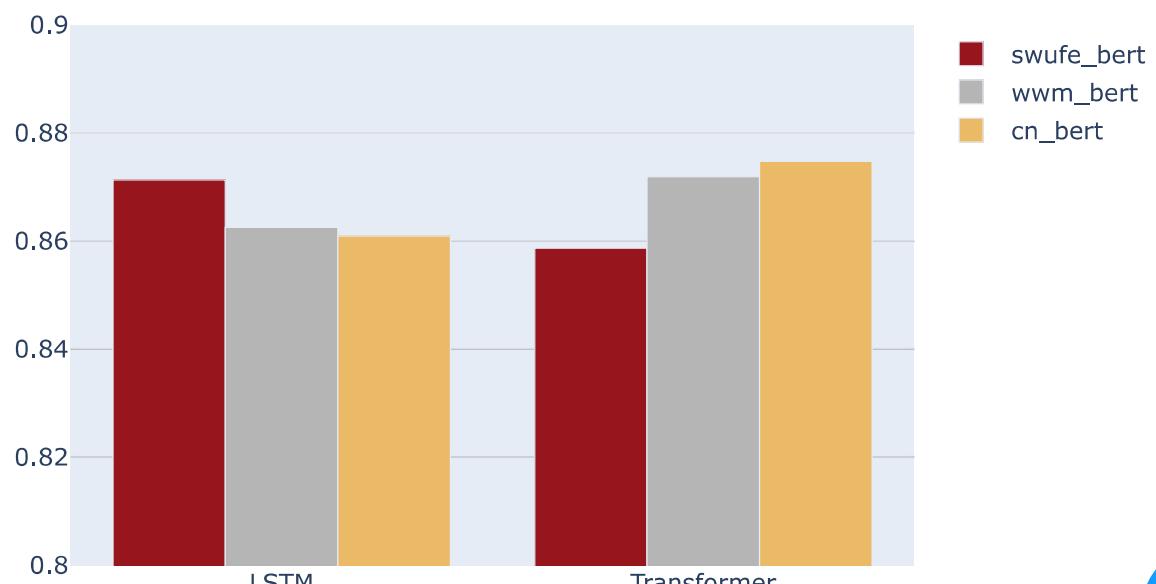
WWM bert

CN bert

Loss



auc_score



Variable:Network

BERT

SWUFE bert

x ▾

LOSS

Unweighted Loss

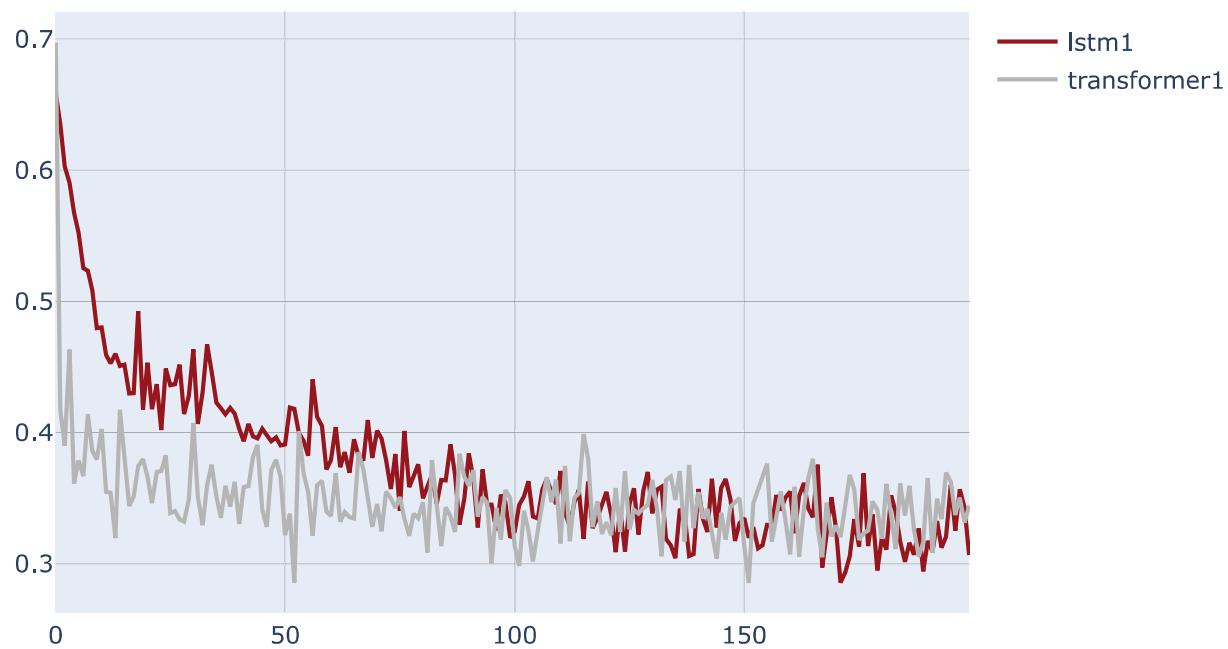
x ▾

NETWORK

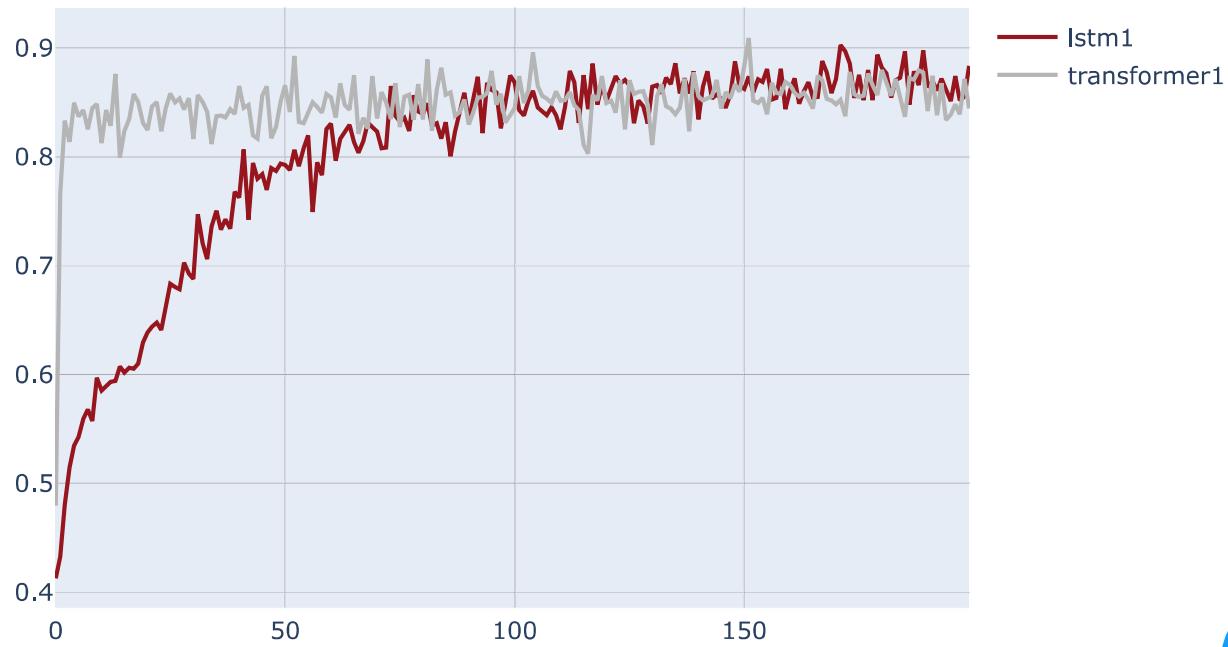
LSTM

Transformer

Loss



AUC



Variable:LossFunction

BERT

SWUFE bert

x ▾

LOSS

Transformer

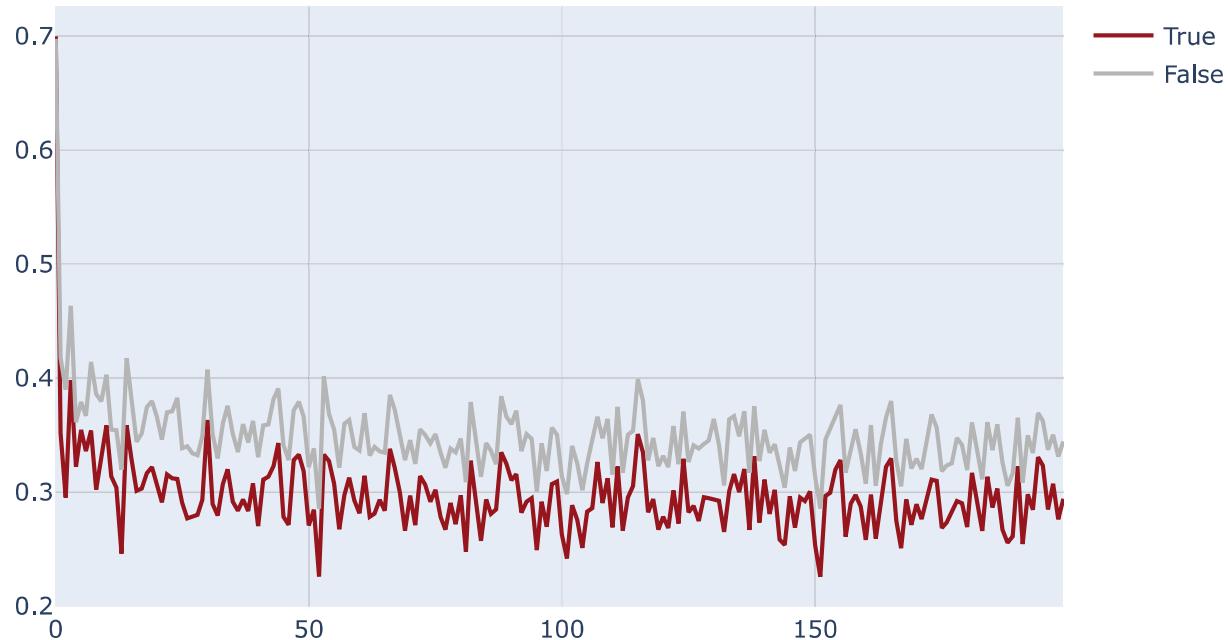
x ▾

LOSS

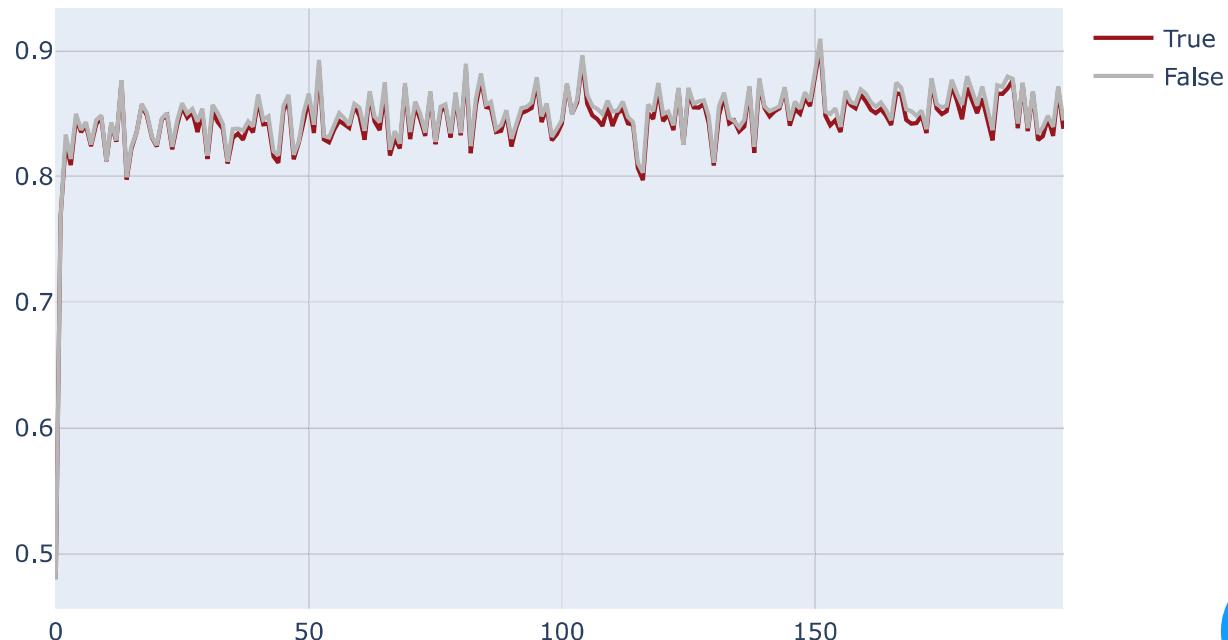
Unweighted Loss

Weighted Loss

Loss



AUC



Variable:Bert

LOSS

Unweighted Loss

x ▾

NETWORK

Transformer

x ▾

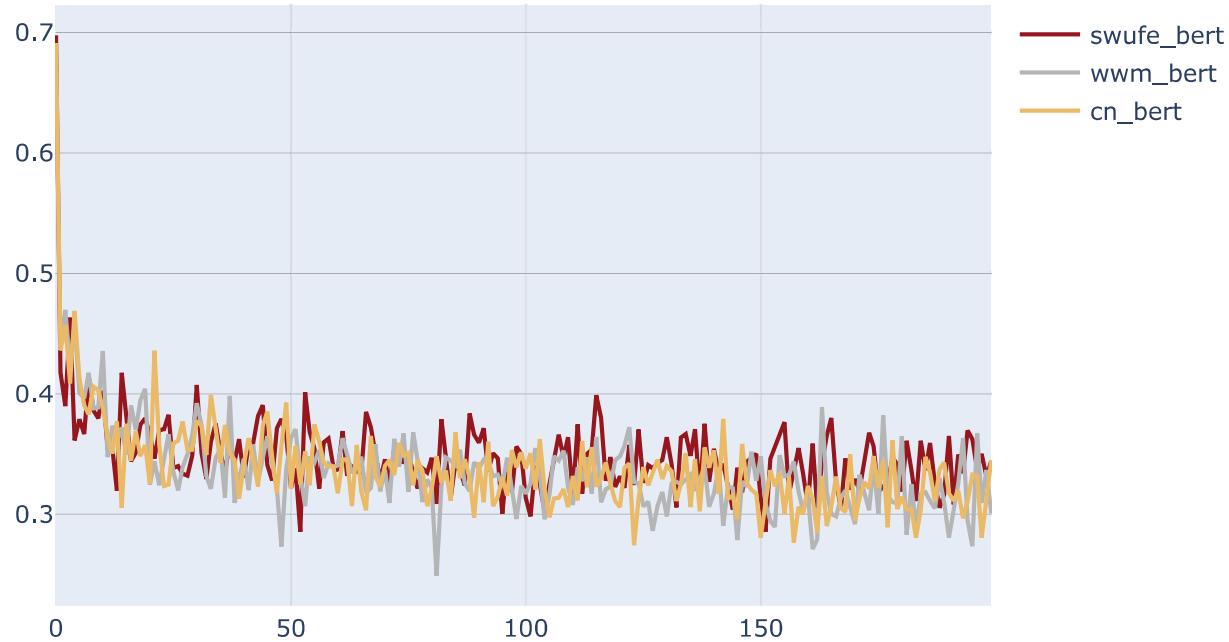
BERT

SWUFE bert

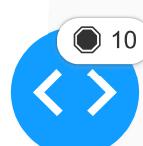
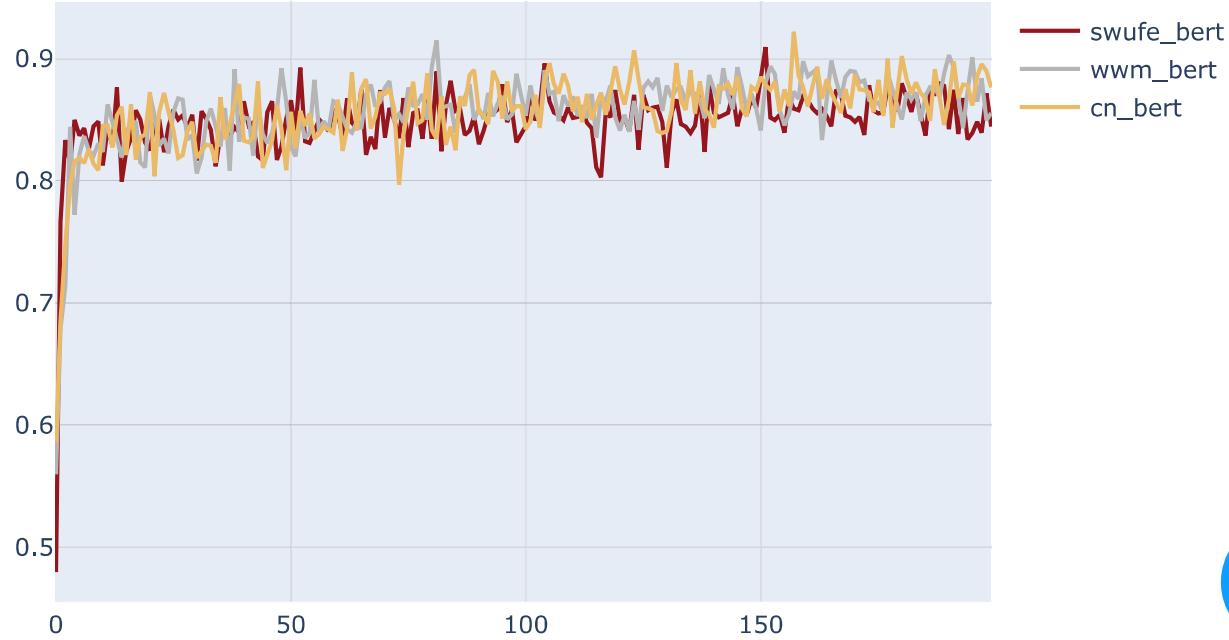
WWM bert

CN bert

Loss



AUC



Extract Summary Finetune

[Full View](#)

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
Question & Improvement | Extract Your News

抽取摘要

本章利用训练好的LSTM和TransformerEncoder模型，选取未标记的时事短讯，对其进行自动式摘要抽取。具体而言，首先将未标记的短讯分别由三种BERT模型生成word embedding。其次将短讯的word embedding分别由训练好的LSTM和TransformerEncoder模型进行预测，获取短讯内每个字符是否被作为摘要内容抽取的0-1标签。最后，使用标记好的新闻短讯生成抽取是摘要。抽取摘要的结果展示如下：

新闻示例

近日，海信旗下家庭互联网AI云平台聚好看加入“5G云游戏产业联盟”，并率先通过了腾讯START云游戏的平台认证，成为OTT行业首家内测合作伙伴。图片来源：云游戏产业发展白皮书(2019)基于海信电视的流量优势和聚好看前身对大数据云平台的自主研发，海信聚好看已经成为全球领先的家庭互联网AI服务云平台，牢牢占据着5000万+家庭的“客厅C位”。与腾讯START云游戏的强强联合，让聚好看的AI服务再次升级。海信电视的高清“黑科技”与聚好看的AI交互和大数据平台底蕴，让玩家能尽情享受大屏玩游戏的乐趣，为5G云游戏的发展开“家庭场景”新赛道。图片来源：云游戏产业发展白皮书(2019)根据《云游戏产业发展白皮书(2019)》，2023年全球云游戏市场规模将达到25亿美元，预计未来国内大量主机游戏将会上线云游戏平台，带来增量规模。此前，聚好看游戏频道已经上线网易云游戏，此次率先通过腾讯START云游戏平台认证，或许在释放一种信号：作为主机游戏的一种，聚好看已经得到主流平台的认可，并将率先开启OTT行业的“客厅云游戏”时代。

摘要结果

`lstm(cn_bert)`:近日，海信旗下家庭互联网AI云平台聚好看加入“5G云游戏产业联盟”，并率先通过了腾讯START云游戏的平台认证，成为OTT行业首家内测合作伙伴。图片来源：云游戏产业发展白皮书(2019)基于海信电视的流量优势和聚好看前身对大数据云平台的自主研发

`lstm(www_bert)`:近日，海信旗下家庭互联网AI云平台聚好看加入“5G云游戏产业联盟”，并率先通过了腾讯START云游戏的平台认证，成为OTT行业首家内测合作伙伴

`lstm(swufe_bert)`:近日，海信旗下家庭互联网AI云平台聚好看加入“5G云游戏产业联盟”，并率先通过了腾讯START云游戏的平台认证，成为OTT行业首家内测合作伙伴。图片来源：云游戏19于海，聚好全球领先的家庭A,



transformer(cn_bert):近海信旗下家庭互联网AI云平台聚好看加入“5G云游戏产业联盟”，并率先通过了腾讯START云游戏的平台认证，成为OTT行业首家内测合作伙伴。

transformer(www_bert):近日，海信旗下家庭互联网AI云平台聚好看加入“5G云游戏产业联盟”，并率先通过了腾讯START云游戏的平台认证，成为OTT行业首家内测合作伙伴。

transformer(swufe_bert):海信旗下家庭互联网AI云平台聚好看加入“5G云游戏产业联盟”，并率先通过了腾讯START云游戏的平台认证，成为OTT行业首家内测合作伙伴。,

Extract Summary Finetune

[Full View](#)

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
Question & Improvement | Extract Your News

问题

问题一：摘要过度关注短讯首端 进行抽取摘要任务时，我们发现摘要的内容过分关注于短讯的前面和开头部分。该现象出现可能是因为数据特征如此（一些金融新闻的标题和关键句都在开头给出），迫使模型学习到此特征。也可能是模型学习效果不佳，导致损失了部分重要特征（比如短讯后面部分的特征）。

问题二：摘要断句存在语意不明 进行抽取摘要任务时，我们发现摘要的部分句子存在吞字、删字、句子不完整、阅读不同顺、语意不明的问题。该现象出现可能是因为标注数据存在语句不通、模型无法识别语言习惯等。

改进

加权损失函数 为了改善模型没有全面关注短讯内不同位置信息的问题，我们采用加权损失函数的方式进行改进。具体而言，给予连续出现的标注信息更低的权重，同时给予离其他标注信息较远的这类“孤立”标注信息更高的权重。迫使模型更多关注全局信息，而非只关注于标注信息较为集中的局部。

为了改善模型抽取摘要的断句效果，进一步提高抽取式摘要的可读性。我们设计了整句式摘要抽取规则来判定摘要的生成。具体而言，如果一个句子的大多数字符都被标记，我们则认为该句子应该完整的呈现在摘要结果中，



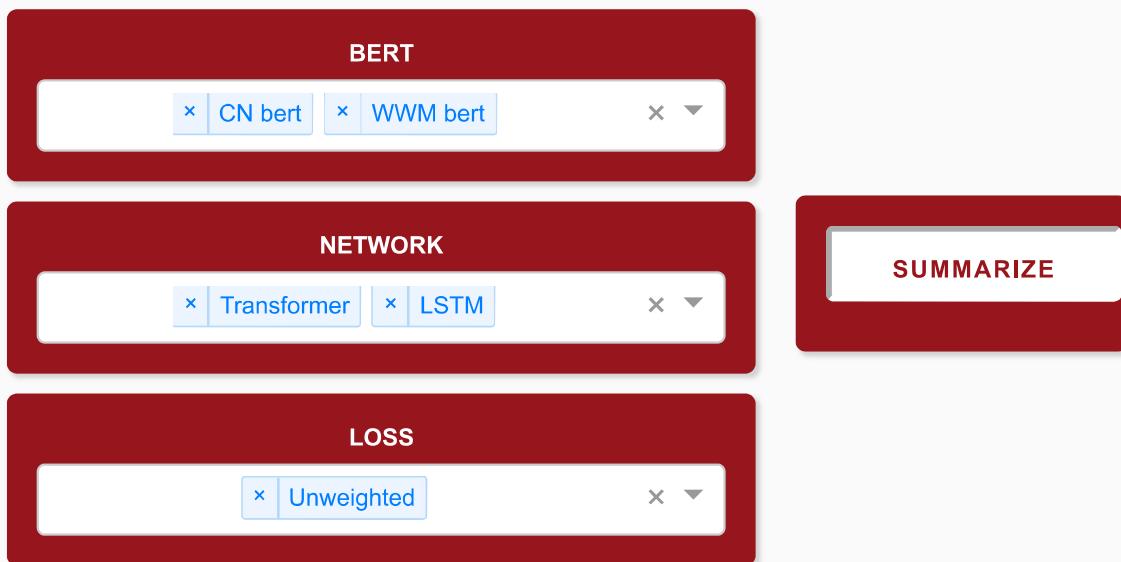
之，该句子小部分字符有抽取标记、或者所有字符都没有字符标记，则该句子无需呈现在摘要结果中。

Extract Summary Finetune

[Full View](#)

Introduction | Question Description | Lstm Model | Transformer Model | Measures | Figures | Extract Summary Dispaly
Question & Improvement | Extract Your News

自动抽取摘要APP



新闻短讯 **Input your original news!**

法巴：中手游(0302.hk)在游戏开发方面的持续投资有望在2023年转化为收入 予“买入”评级目标价4.7港元
格隆汇8月18日 | 法国巴黎银行发研报称，予中手游“买入”评级，目标价4.7港元
(意味着按2022年盈利计算，约为市盈率的15倍)。中手游是一家以IP为基础的游戏开发商和发行商，截至2021年底，拥有55个授权IP和68个专有IP。公司拥有已上线游戏超过80款和逾20款生命周期超过三年的游戏。2021年中手游游戏的月均活跃用户(MAU)达到1910万。

该行认为，公司在游戏开发方面的持续投资将在2023年开始转化为收入。法巴仍然看好该公司的估值，其市盈率为该行2022年预期收益的11倍和2023年预期收益的9倍。展望2022年下半年，法巴认为该公司将会有更好的表现。公司已经获得了《我的御剑日记》的新游戏版号。多款游戏将于2022年下半年推出，并将在下半年为公司带来收入。该行预测该公司在2022年下半年的收入将同比增长21%。

该行认为中手游将能够实现持续的收入增长，原因如下：1) 中手游转向长周期游戏

新闻自动摘要 **Summary**

CN BERT-TRANSFORMER-UNWEIGHTED-SUMMARY:

法巴：中手游(0302.hk)在游戏开发方面的持续投资有望在2023年转化为收入 予“买入”评级目标价4.7港元
格隆汇8月18日 | 法国巴黎银行发研报称，予中手游“买入”评级，目标价4.7港元 (意味着按2022年盈利计算，

WWM BERT-TRANSFORMER-UNWEIGHTED-SUMMARY:



法巴：中手游(0302.hk)在游戏开发方面的持续投资有望在2023年转化为收入 予“买入”评级目标价4.7港元格隆汇8月18日
| 法国巴黎银行发研报称，予中手游“买入”评级，目标价4.7港元（意味着按2022年盈利计算，

CN BERT-LSTM-UNWEIGHTED-SUMMARY:

法巴：中手游(0302.hk)在游戏开发方面的持续投资有望在2023年转化为收入 予“买入”评级目标价4.7港元 格隆汇8月18日
| 法国巴黎银行发研报称，予中手游“买入”评级，目标价4.7港元（意味着按2022年盈利计算，约为市盈率

WWM BERT-LSTM-UNWEIGHTED-SUMMARY:

法巴：中手游(0302.hk)在游戏开发方面的持续投资有望在2023年转化为收入 予“买入”评级目标价4.7港元格隆汇8月18日
| 法国巴黎银行发研报称，予中手游“买入”评级，目标价4.7港元（意味着按2022年盈利计算，约为市盈率的15倍）。行
将能够实现持入增

