基于众包时间同步评论的概念-动作映射的视频亮点检测和总结与滞后校准

清平和陈超美
德雷塞尔大学计算机和信息学学院
{qp27, cc345}@drexel.edu

摘要

随着视频共享的盛行，对高光检测等直观的视频消化的需求越来越大。最近，世界上出现了众包时间同步视频评论的平台，为高光检测提供了良好的机会。然而，这项任务并不简单：（1）时间同步评论往往落后于其相应的镜头；（2）时间同步评论通常是稀疏和嘈杂的；（3）确定哪些镜头是亮点是非常主观的。本文旨在通过提出一个框架来解决这些挑战：(1)使用con-
cept映射的词汇链进行滞后校准；(2)根据每个镜头的评论强度以及情感和概念集中的组合来建立视频高光模型；(3)使用改进的SumBasic与情感和概念映射对每个检测到的高光进行汇总。在大型真实世界数据集上的实验表明，我们的高光检测方法和求和方法都以相当大的幅度超过了其他基准的表现。

1 简介

每天，人们在YouTube上观看数十亿小时的视频，其中一半的浏览量来自移动设备[1]。随着视频分享的盛行

1 https://www.youtube.com/yt/press/statistics.html

目前，人们对快速视频浏览的需求越来越大。想象一下这样的场景：用户想快速掌握一段长视频，而不需要反复拖动进度条来跳过对用户来说不重要的镜头。有了自动生成的亮点，用户可以在几分钟内消化整个视频，然后再决定以后是否观看完整的视频。此外，自动视频高光检测和总结可以使视频索引、视频搜索和视频推荐受益。

然而，从视频中寻找亮点并不是一件简单的事情。首先，什么被认为是"亮点"可能是非常主观的。其次，通过分析图像、音频和视频中的低层次特征，不一定能捕捉到亮点。缺乏抽象的语义信息已经成为传统视频处理中亮点检测的一个瓶颈。

最近，众包的时间同步视频评论，或"弹幕评论"已经出现，实时生成的评论将在屏幕上方或旁边飞过，与视频逐帧同步。它已经在世界范围内得到了普及，如日本的niconico，中国的Bilibili和Acfun，美国的YouTube Live和Twitch Live。时间同步评论的流行为基于自然语言处理的视频亮点检测提供了新的机会。

尽管如此，使用时间同步的评论来检测和标记高光部分仍然是一个挑战。首先，与每个镜头相关的评论几乎不可避免地存在滞后。如图1所示，正在进行的关于一个镜头的讨论可能会延伸到下几个镜头。没有滞后校准的高光检测和标签可能会导致不准确的结果。第二。

时间同步的评论在语义上是稀疏的，包括每个镜头的评论数量和每个评论的标记数量。传统的词包统计模型在这类数据上可能效果不佳。

第三，在没有任何先验知识的情况下，在无监督的环境下，高光检测存在很多不确定性。亮点的特征必须被明确地定义、捕获和建模。

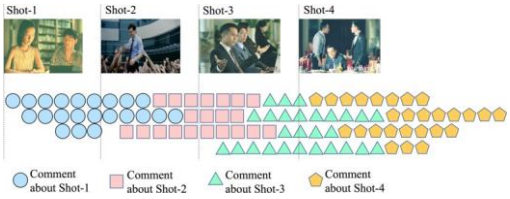

图1.时间同步组件的滞后效应逐个拍摄。

据我们所知，很少有工作集中在基于无监督的时间同步评论的亮点检测和标签上。最相关的工作提出了根据弹幕评论的语义向量的主题集中度来检测亮点，并根据预先定义的标签用预先训练好的分类器来标记每个亮点（Lv, Xu, Chen, Liu, & Zheng, 2016）。尽管如此，我们认为在高光检测中，情感集中比一般话题集中更重要。另一项工作提出了基于情绪分布的逐帧相似性来提取亮点（Xian, Li, Zhang, & Liao, 2015）。然而，这两项工作都没有提出同时解决滞后校准、情绪-
话题浓度平衡和无监督高光标签的问题。

为了解决这些问题，本研究提出了以下建议。(1)基于全局词嵌入的词到概念和词到情感的映射，从中构建词汇链，用于弹幕评论的滞后校准；(2)基于滞后校准的弹幕评论的情感和概念浓度和强度的高光检测；(3)用改进的基本和算法进行高光总结，将情感和概念视为弹幕评论的基本单位。

本文的主要贡献有以下几点。(1)我们提出了一个完全无超视的框架，用于基于时间同步评论的视频亮点检测和总结；(2)我们开发了一种基于概念映射词链的滞后校准技术；(3)我们构建了大型数据集，用于弹幕评论词------。嵌入、弹幕情感词库和地面真相，用于高亮检测和基于弹幕的标签评估。

2 相关工作

2.1 通过视频处理进行高光检测

首先，按照以前工作的定义（M. Xu, Jin, Luo, & Duan,
2008），我们把高光定义为视频中最令人难忘的具有高情感强度的镜头。需要注意的是，高光检测与视频总结不同，视频总结侧重于浓缩的故事情节表现，而不是提取情感内容（K.-S. Lin, Lee, Yang, Lee, & Chen, 2013）。

对于高光检测，一些研究者假设用唤醒-情绪平面上的曲线来表示视频中的情绪，低级特征如运动、声音效果、镜头长度和音频音调（Hanjalic & Xu, 2005），颜色（Ngo, Ma, & Zhang, 2005），中级特征如笑声和字幕（M. Xu, Luo, Jin, & Park, 2009）。然而，由于低级特征和高级语义之间的语义差距，基于视频处理的亮点检测的准确性是有限的（K.-S. Lin等人，2013）。

2.2 时间性文本总结

时间性文本总结方面的工作与本研究有关，但也有不同之处。一些工作将时态文本总结表述为一个受限的多目标优化问题（Sipos, Swaminathan, Shivaswamy, & Joachims, 2012; Yan, Kong, et al., 2011; Yan, Wan, et al,2011），作为一个图形优化问题（C. Lin等人，2012），作为一个监督学习-排名问题（Tran, Niederée, Kanhabua, Gadiraju, & Anand, 2015），以及作为在线集群问题（Shou, Wang, Chen, & Chen, 2013）。

本研究将高光检测建模为一个简单的带有约束条件的双目标优化问题。然而，选择用来评估一个镜头的"亮点"的特征与上述研究不同。因为据观察，一个高亮度的镜头与高情感强度和话题集中度相关，覆盖率和非冗余度不再是优化的目标，就像在时间性文本总结中一样。相反，我们在本研究中着重于对情感和话题集中度进行建模。

2.3 众包时间同步的评论挖掘

有几项工作侧重于通过手动标签和监督训练（Ikeda, Kobayashi, Sakaji, & Masuyama, 2015）、时间和个性化主题建模（Wu, Zhong, Tan, Horner, & Yang, 2014），或将视频作为一个整体进行标记（Sakaji, Kohana, Kobayashi, & Sakai, 2016）。有一项工作提出通过在文本和主题层面上联合进行数据重建来生成每个镜头的摘要（L. Xu & Zhang, 2017）。

一项工作提出了一个中心点扩散算法来检测亮点（Xian等人，2015）。镜头由LDA的潜在主题来表示。另一项工作提出使用预先训练好的评论语义向量，将评论聚类为话题，并根据话题的集中度找到亮点（Lv等人，2016）。此外，他们使用预先定义的标签来训练一个分类器，以进行高亮显示。本研究在几个方面与这两项研究不同。首先，在高光检测之前，我们进行滞后校准，以尽量减少评论滞后造成的不准确。第二，我们通过主题和情感浓度的组合来表示每个场景。第三，我们以无监督的方式进行高光检测和高光标记。

2.4 词汇链

词链是指在一系列句子中具有内聚关系的词的序列。早期的工作是以罗杰特词典中的词的协同关系为基础构建词链，而不进行词义辨析（Morris & Hirst, 1991）。后来的工作是通过WordNet关系和词义辨析来扩展词链（Barzilay & Elhadad, 1999; Hirst & St-Onge, 1998）。词链也是基于词埋关系构建的，用于多词的消歧义（Ehren, 2017）。本研究在全局词嵌入的基础上构建词链以进行适当的滞后校准。

3 问题的提出

本文中的问题可以表述如下。输入是一组时间同步评论，$C=\{c_1, c_2, \ldots, c_n\}$，一组时间戳$T=\{t_1, t_2, \ldots, t_n\}$的视频$v$，压缩率$\tau_{123142315}$，用于生成高光的数量，压缩率为$\tau_{67889:;}$为每个高光摘要中的评论数量。我们的任务是：（1）生成一组高光镜头$S(v) = \{s_1, s_2, \ldots, s_m\}$，和（2）高光摘要A $v = \{I_1, I_2, \ldots, I_m\}$尽可能接近地面真相。每个亮点摘要包括这个镜头中所有评论的一个子集。$I_2 = \{c_1, c_2, \ldots, c_k\}$。高光镜头的数量$n$和摘要中的评论数量$n_2$是由$\tau_{123142315}$和$\tau_{67889:;}$分别为。

4 视频高光检测

在这一节中，我们介绍了我们的亮点检测框架。我们还描述了两个初步的任务，即构建全局时间同步的评论词嵌入和情绪列举。

4.1 预备工作

时间同步评论的文字嵌套

如前所述，分析时间同步评论的一个挑战是语义的稀疏性，因为评论的数量和评论的长度都非常有限。如果两个语义相关的词在一个视频中不经常出现，它们可能就没有关系。为了弥补这一缺陷，我们在大量的时间同步评论中构建了一个全局性的词嵌入。

嵌入词的字典可以重复发送。$D\{(w_1: v_1), (w_2: v_2), \ldots, (w_n: v_n)\}$，其中$w_2$是一个词，$v_2$是相应的词向量，$V$是语料库的词汇。

情感词库建设

正如前面所强调的，在时间同步的评论中提取情感是至关重要的，以便进行重点检测。然而，传统的情感词典不能用在这里，因为有太多的网际俚语是专门在这种类型的平台上诞生的。例如，"23333 "意味着 "哈哈哈"，而 "6666 "意味着
"真的很好"。因此，我们从上一步训练的词语嵌入词典中构建一个为时间同步评论量身定做的情感词典。首先，我们从语料库中出现频率最高的词中，将五个基本的情绪类别（*快乐*、*愤怒*、*悲伤*、*恐惧和惊讶*）的词标记为种子（Ekman, 1992）。这里省略了第六个情绪类别 "*厌恶*"，因为它在这个数据集中比较少见，而且可以很容易地纳入其他数据集。然后，我们通过搜索每个种子词的前N个邻居来扩展情感词库

在词嵌入空间中，如果邻居与所有种子的最小相似度为$sim_{82}@$，则添加邻居到种子中。邻居的搜索是基于单词嵌入空间的余弦相似性。

4.2 滞后校准

在这一节中，我们介绍了我们的滞后校正方法，包括概念图的绘制、嵌入词链的构建和滞后校正的步骤。

概念图

为了解决时间同步评论中的语义稀疏问题，并构建语义相关词的词链，应首先将意义相似的词映射到同一概念中。鉴于视频$\boldsymbol{v}$的一组评论C，我们首先提出了一个从词汇V_{T}o评论C到一组概念KT的映射$\mathcal{F}$，即。

$$\mathcal{F}: V_{\mathrm{t}} \rightarrow K\,\mathrm{t}\ \ (V\mid \geq \mid K\,\mathrm{t}\mid\ )\mid$$

更具体地说，映射$\mathcal{F}$将每个工作都映射为$_w$X变成一个概念$k = \mathcal{F}(_w\mathrm{X})$。

$$\begin{cases} k, \exists k \in \ _K\mathrm{T} & and \ \frac{|\{\in 5\mathrm{MP\_@}(\backslash_{\mathrm{b}}) \wedge \mathcal{F}\ \mathcal{F}\ \mathrm{d}\ \mathrm{e}\}|}{|5\mathrm{MP\_@}(\mathrm{b})\ \mid} \geq \phi\mathrm{MNO}_{:49\mathrm{P}} \\ w, \ o\ t\ herwi\ s\ e & \end{cases} \tag{1}$$

而$top_n(_w\mathrm{X})$则根据余弦相似度返回单词$_w$X的前n邻人。对于评论中的每个词$_w$X，我们检查其邻居中已经被映射到概念k的百分比。如果百分比超过阈值$\phi\mathrm{MNO}_{:49}$，那么单词$_w$X及其邻居将被映射到k。否则，它们将被映射到新的概念$_w$X。

词汇链建设

下一步是在视频$\boldsymbol{v}$的当前时间同步评论中构建所有的词汇链，这样就可以根据词汇链来校准滞后评论。一个词链$l_{2\mathrm{m}}$包括一组三要素$l_{2\mathrm{m}} = \{(w, t, c)\}$，其中$w$是评论$k$中实际提到的概念$c_2$，$t$是评论$c$的时间戳。用于时间同步评论的词链词典$D_{4\mathrm{On2o94\ o192@}}$。

$\boldsymbol{v}$: $L_{4\mathrm{On2o94\ o192@}} = \{k\%{:}l\%, l\%{\cdot}, l\%(\cdots(\kappa{\cdot}$)

$(l'\%, \ _{\boldsymbol{l}''}, l'(\ \ldots), \ldots, k_{\mathrm{qr}}\ \ \{\ _{\mathrm{q\,r}}\%, l_{\mathrm{q\,r}}{\cdot}, l_{\mathrm{q\,r}}(\ldots)\}$，其中$k2 \in K_{\mathrm{T}}$是一个概念，$l_{2\mathrm{m}}$是概念$jth$的词链$_2$。词链构建的算法在算法1中描述。

具体来说，C中的每条评论既可以被附加到现有的词链中，也可以被添加到新的空词链中，其依据是它与现有词链的时间距离由Maxi- mum silence $l_{89\ \mathrm{n}}$控制。

请注意，词汇链中的词义与这里构建的概念图没有像大多数传统算法那样进行歧义处理。然而，我们认为词汇链仍然是有用的，因为我们的概念映射是由时间同步的评论按其自然顺序构建的，这种渐进的语义连贯性自然地加强了时间上接近的评论的类似词感。这种语义上的连续性以及全局性的词语嵌入确保我们的概念映射在大多数情况下是有效的。

算法1 词汇链构建

输入时间同步的评论C。词到概念的映射$\mathcal{F}$.最大限度的沉默$l_{89\mathrm{n}}$。
输出 词典中的词链$L_{4\mathrm{On2o94\ o192@}}$。
初始化$L_{\ 4\mathrm{On2o94o192@}} \leftarrow \{\}$。
对于C中的每个c，**做**
  $t_{\mathrm{o7::0@5}} \leftarrow t_{\mathrm{o}}$
  对于c中的每个*词*，**做**
    $k \leftarrow \mathcal{F}(word)$
    如果k在$L_{4\mathrm{On2o94\ o192@}}$**那么**
      $chains \leftarrow L_{4\mathrm{On2o94\ o192@}}(k)$
      $t_{\mathrm{P:0N2M76}} \leftarrow to192@6_{[4965]}$。
      如果$_{tcurent} - _{tprevious} \leq l_{89\mathrm{n}}$**则**
        $chains[last] \leftarrow chains[last] \cup c$
      否则
        $chains \leftarrow chains \cup \{c\}$。
      结束 如果
    否则
      $L_{4\mathrm{On2o94\ o192@}}(k) \leftarrow \{\{c\}\}$。
    end
  if end for
结束

评论 滞后-校准

现在给定构建的词链词典$L_{4\mathrm{On2o94\ \ o192@}}$，我们可以根据其词汇链来校准$C$中的评论。根据我们的观察，关于一个镜头的第一条评论通常是在该镜头内出现的，而其他的可能不是这样。因此，我们将每条评论的时间戳校准为其所属词汇链的第一个元素的时间戳。在所有的词条中一个评论属于什么概念呢？挑选分数最高的一个$score_{\mathrm{o}}$。

$Score_{\mathrm{e,o}}$被计算为链中每个词的频率之和，并以其对数全局频率对数$(D\ \ w\ .\ count)$加权。因此。

每个评论将被分配到其最重要的语义链（概念）中进行校准。校准的算法在算法2中进行了描述。

请注意，如果有多个连续镜头

算法2 时间同步信号的滞后校准

输入时间同步的评论C。词到概念的映射　　$\mathcal{F}$.

词法　链　　词典

40n2o94o192@ L 。词条嵌入词典D。

输出滞后校准的时间同步评论 C'。

初始化 $C' \leftarrow$ 晶n

对于C'中的每个c, **做**

$chain$®065$_{,o} \leftarrow \{\}$。

$score$®065$_{,o} \leftarrow 0$

对于c中的每个*词*, **做**

$k \leftarrow \mathcal{F}(word)$

$chain_{e,o} \leftarrow L_{40n2o94\,o192@}(k)$ [c]

$score_{e,o} \leftarrow 0$

对于 (w,t,c) 在$chain$ **do**

$N(w) \leftarrow D(w)$。$count$

$_{e,o} \leftarrow scorescore_{e,o} + 1/\log(N(w))$

结束

如果$score_{e,o} > s\ c\ o\ r$ ®065 **，则**

$chain$®065$_{,o} \leftarrow chain_{e,o}$

end

if end

for

õ$_o \leftarrow$ $to192_{@}$´åñ¨$_{...[Æ2:65]}$

end for

返回 C

$\{s_{\%}, s, \ldots, s_{8}\}$，有类似内容的评论。

我们的滞后校正方法可能会将镜头s', $s_{(}$, ..., s_8中的许多评论校正为第一个镜头$s_{\%}$的时间戳，如果这些评论是通过词汇链从镜头$s_{\%}$连接起来的。这不一定是坏事，因为我们希望在固定的压缩率下，避免看到多余的连续高光镜头，为其他候选高光镜头留下机会。

射击重要性得分

在这一节中，我们首先通过时间长度相等的镜头

l6oO$_{@}$O来分割评论，然后建立镜头重要性模型。然后，可以根据镜头的重要性来检测亮点。

一个镜头的重要性被模拟为受到影响

有两个因素：评论的集中度和评论的强度。对于评论的集中度，如前所述，概念和情感的集中度都可能有助于亮点的识别。例如，一组概念集中的评论，如

"这个镜头的背景音乐/bgm/音轨是一流的/鼓舞人心的/最好的"，可能是一个指标。

与难忘的背景音乐有关的亮点。同时，诸如 "这个情节太有趣了/搞笑/lmao/lol/2333 "这样的评论可能暗示了一个单纯的情感集中的亮点。因此，我们在模型中结合这两种浓度。首先，我们根据时间同步评论C_6 给定的情感词汇E，定义镜头s的情感浓度如下。

$$\mathcal{C}_{O8M52M}@(C_6, s) = \frac{\%}{\text{ãÑ}_{\square}\ \text{Pä}\cdot\text{éèê (Pä)}} \tag{2}$$

$$p_0 = \frac{\{\text{l}\in T_{\text{ä}}\wedge\backslash\in\text{ï}(0)\}_{\text{。}}}{|T_{\text{ñ}}|} \quad | \tag{3}$$

在这里，我们把一个镜头内五种情绪的概率的熵的反向计算为情绪浓度。然后我们定义局部浓度$\mathcal{C}_{5MP2o}$。

$$\mathcal{C}_{5MP2o}(C_6, s) = \frac{\%}{\text{òng}\ \text{òng}\ \text{òng}\ \text{òng (P)}} \tag{4}$$

$$p_{\text{e}} = \frac{\sum_{\text{ó}\in\text{ò}(\text{ï}\cdot\text{î})} \mathcal{F}\ \text{ï}\text{ó}\notin\text{ú}\ \text{)@ò}\ \text{éèê (ò ô)}_{\text{ò}}/}{\sum_{\substack{\text{ó}\in\text{ò}(\text{ï}\cdot\text{î}) \\ \text{ô}\notin\text{ú}}} \sum_{\text{ô}\in\text{ï}\cdot\text{î}\wedge\mathcal{F}\ \text{ô}\ \text{ï}\text{ó}\wedge} \text{@ò}\ \text{éèê (öò)}} \tag{5}$$

其中，我们计算一个镜头中所有概念的熵的反向，作为话题集中度。每个概念的概率k是由其提到的词的频率之和加权后的全局频率决定的，并除以该镜头中所有词的这些值。

现在，评论的重要性$\mathcal{J}$oM88O$_{@}$5 $_{C\,6}$, s（　　）镜头的s可以定义为。

$$\mathcal{J}\text{oM88O@}(s) = \lambda \cdot \mathcal{C}_{O8M52M}(@_{C\,6}, s + (1-\lambda)\cdot \mathcal{C}_{5MP2o}(C_6, s) \tag{6}$$

其中

λ是一个超参数，控制情绪和概念集中之间的平衡。

最后，我们将射击的总体重要性定义为：。

$$\mathcal{J}(C_6, s) = \mathcal{J}\text{oM88O}_{@}(5_{C\,6}, s \cdot \log(|C_6|) \tag{7}$$

其中$|$，C。是镜头s中所有时间同步评论的长度，它是每个镜头评论强度的一个直接而有效的指标。

现在，高光检测的问题可以被建模为一个最大化问题。

$$Maximize \quad \sum_{6d\%}^{\ddot{o}} \mathcal{J}\,C_6, s \cdot \quad (8)$$

$$Subjctive\ to \quad \begin{cases} \sum_{\text{nd}\%}^{\ddot{o}} x \le \tau_{123142315} \cdot N \\ x_6 \in \text{Q}1\ \} \end{cases}$$

5 视频亮点归纳

给出一组检测到的高光镜头$S(v)$=

$\{s_{\%}, s', s_{(}, \ldots, s@\}$的视频$v$，每个都有该镜头的所有滞后校准的评论 C_6 ，我们是在

诱导生成摘要 A $\boldsymbol{v}=$ （）
$\{I_{\%}, \boldsymbol{I}^{\cdot}, I_{(}, \ldots, I@\}$所以$I_{6\subset C}6$，压缩比$\tau_{67889:;}$并且$I_6$尽可能地接近地面真相。

我们提出了一个简单但非常有效的总结模型，是对SumBasic（Nenkova & Vanderwende, 2005）的改进，具有情感和概念映射以及两级更新机制。

在修改后的SumBasic中，为了防止冗余，我们不是只对被测句子中的词的概率进行下采样，而是对词和其映射的概念的概率都进行下采样，以便对每个评论进行重新加权。这种两级更新机制可以(1)对有语义相似词的句子进行惩罚性选择；(2)如果这个词出现得更频繁，仍然选择已经在摘要中出现的句子。此外，我们使用一个参数*情感偏差*b±êèï¥èµ来决定。
在计算单词和概念的重量时
概率，因此，与非情绪化的词语和概念相比，情绪化的词语和概念的频率将增加b±êèï¥èµ。

6 实验

在本节中，我们在大型真实数据集上进行了亮点检测和总结的实验。我们将描述数据收集过程、评估指标、基准和实验结果。

6.1 数据

在本节中，我们描述了在我们的实验中收集和构建的数据集。所有的数据集和代码都将在Github上公开提供[2]。

众包的时间同步评论语料库

为了训练4.1.1中描述的词嵌入，我们从Bilibli[3]，一个中国的内容分享网站收集了大量的时间同步评论的语料。该语料库包含2,108,746条评论，15,179,132个标记，91,745个独特的标记，来自6,368段长视频。每条评论平均有7.20个tokens。

在训练之前，首先使用Chinse单词标记化包Jieba[4]，对每条评论进行Kenized。词汇中的重复字符，如"233333", "66666", "哈哈哈"被替换成两个相同的字符。

词汇嵌入是用word2vec（Goldberg & Levy, 2014）的跳格模型训练的。嵌入维数为300，窗口大小为7，下采样率为1-3，频率低于3次的词被丢弃。

情感词库建设

在单词嵌入训练完成后，我们从单词嵌入中500个最频繁的单词中手动选择属于五个基本类别的情感词。然后，我们使用算法1迭代地扩展情感种子。在每个

	快乐	悲伤	恐惧	愤怒	惊喜
种子	17	13	21	14	19
所有	157	235	258	284	226

表3.初始和扩展的情感词的数量。

扩增迭代时，我们还对扩增后的词库进行人工检查，删除不准确的词，以防止概念漂移效应，并在下一轮扩增中使用经过过滤的扩增种子。最小重叠度γMNO:49P被设定为0.05，最小相似度$sim_{82}@$被设定为0.6。γMNO:49P和$_{s}im82@$的选择是基于0,1范围内的网格搜索[1]。表3中列出了每种情感最初和最后扩展的字数。

视频亮点数据

为了评估我们的高光检测算法，我们构建了一个地面实证数据集。我们的真实数据集利用了用户在Bilibli上上传的关于某个特定视频的混合剪辑。混合剪辑是由用户自己的喜好来拼贴视频的高亮部分。然后，我们把投票最多的亮点作为视频的基础事实。

该数据集包含11个长度为1333分钟的视频，其中有75653条时间同步的评论。对于每个视频，我们从Bilibili收集了3~4个关于这个视频的混合剪辑。在所有混合剪辑中至少有2个出现的镜头被认为是地面真实的亮点。所有的地面实况高光都被映射到原始视频的时间轴上，高光的开始和结束时间被记录为地面实况。混合片段的选择基于以下启发式方法：（1）在Bilibli上用关键词搜索混合片段

[2]https://github.com/ChanningPing/VideoHighlightDetection
[3]https://www.bilibili.com/
[4]https://github.com/fxsjy/jieba

"视频标题+混合剪辑"；（2）混合剪辑

按播放时间降序排序；(3)
混合剪辑应该主要是关于视频的亮点，而不是逐个情节的总结或要点；（4）混合剪辑应该在10分钟以内；（5）混合剪辑应该包含几个亮点的混合
而不是只有一个镜头。

平均而言，每个视频有24.3个高光镜头。高光镜头的平均长度为27.79秒，而模式为8和10秒（频率=19）。

亮点总结数据

我们还构建了一个高亮总结（la-summarization）。
bing）数据集的11段视频。对于每个高光拍摄的评论，我们要求注释者
构建这些评论的摘要，在他们认为必要的情况下，尽可能多的评论。经验法则是。(1)
同一含义的评论不会被选取超过一次；(2)
在类似的评论中选取最有代表性的评论；(3)
如果某条评论本身很突出，而且与当前的讨论无关，则选取该评论。
的，将被丢弃。

为11段267个亮点的视频，每个亮点平均有3.83条评论作为其摘要。

6.2 评价指标

在这一节中，我们介绍了高光检测和总结的评价指标。

视频高光检测评估

对于视频高光检测的评估，我们需要定义什么是高光候选和参考之间的
"命中"。一个严格的定义是，候选高光和参考高光之间的起点和终点要完美匹配。然而，这对任何模型来说都是太苛刻了。一个更宽容的定义是候选高光和参考高光之间是否有过度的搭接。然而，这仍然会低估模型的性能，因为用户对高光部分的开始和结束的选择有时是很随意的。取而代之的是，我们提出了一个具有放松性的 "命中 "定义
候选者h和参考者H之间的ε如下。

$$h_i t\pi(h,\widehat{H}) = \begin{cases} 1, & \exists 1\in ö:(6\text{Å}\tilde{a}\pi,0\text{Å}^{9}\pi)\in\emptyset \text{M}51\text{O}: \\ 0, & \text{260} \end{cases} \quad (9)$$

其中$s1$，$_e1$是高光的开始时间和结束时间h，ε是参考的放松长度。
经验集H。此外，精度、召回率和F-1度量可以定义为：。

$$Precision(H,H) = \frac{\Sigma^{|\varnothing|}125\ 1,ö}{|ö|} \quad (10)$$

$$Recall(H,H) = \frac{\Sigma^{|\hat{}|}\varnothing 125(1,ö)}{|ö|} \quad (11)$$

$$F1(H,\widehat{H}) = \frac{\text{'}\cdot\neg:0o262M@(ö,\hat{ö})û\ Oo944(ö,\hat{ö})}{\neg ö,ö^{9}\hat{u}0o9(4\ ö,\hat{ö})} \quad (12)$$

在本研究中，我们将松弛长度设定为为5秒。此外，候选亮点的长度被设定为15秒。

视频亮点总结评估

我们使用ROUGE-1和ROUGE-2（C.-Y.Lin, 2004年）作为召回的候选摘要进 行 评估。

$$ROUGE\text{-}n(C,R) = \frac{\Sigma_{i\in\text{À}}\ \Sigma_{n\text{-}gram\in i}\text{TM7@5}\ldots}{\Sigma_{i\in\text{À}}\ \Sigma_{n\text{-}gram\in i}\text{TM7@5}(n\text{-}gram)} \quad (13)$$

我们使用BLEU-1和BLEU-2（Papineni, Roukos, Ward, & Zhu, 2002）作为精度。我们选择BLEU有两个原因。首先，对于较短的评论来说，天真的精度指标会有偏差，而BLEU可以通过BP乘积因子来补偿这一点。

$$BLEU\text{-}n(C,R) = \text{BP}\frac{\Sigma_{n\text{-}gram\in i}\text{TM7@5}(n\text{-}gram)}{\Sigma_{n\text{-}gram\in i}\text{TM7@5}(n\text{-}gram)} \quad (14)$$

$$BP = \begin{cases} 1, & if\ C > R \\ e^{(1-R/C)} & if\ C \le R \end{cases}$$

其中，C是候选摘要，R是参考摘要。其次，虽然参考文献摘要不包含冗余内容，但候选摘要可能会错误地选择与参考文献中相同关键词非常相似的多个评论。在这种情况下，精度会被严重高估。BLEU只对匹配的词进行逐一统计，也就是说，一个词的匹配次数将是候选人和参考文献中的最小频率。

最后，F-1措施可以被定义为。

$$F1\text{-}n(C,R) = \frac{2\cdot BLEU\text{-}n(C,R)\cdot ROUGE\text{-}n(C,R)}{BLEU\text{-}n(C,R) + ROUGE\text{-}n(C,R)} \quad (15)$$

6.3 基准方法

视频高光检测的基准

对于亮点检测，我们提供了我们的模型与三个基准的不同组合的比较。

- **随机选择。** 我们选择亮点

从一个视频的所有镜头中随机抽取镜头。

- **统一选择。**我们以相等的时间间隔选择高光镜头。

- **尖峰选择**。我们选择那些在镜头中拥有最多评论的高光镜头。
- **尖峰+E+T**。这是我们的方法，考虑到了情感和主题浓度，没有滞后校准步骤。
- **Spike+L**。这是我们的方法，只有滞后校准步骤，没有考虑到内容浓度。
- **钉子+L+E+T**。这就是我们的完整模型。

视频亮点总结的基准

对于亮点总结，我们提供了我们的方法与五个基准的比较。

- **SumBasic.**专门利用频率进行总结构建的总结（Nenkova & Vanderwende, 2005）。
- **潜在语义分析（LSA）。**基于奇异值构成（SVD）的文本总结，用于潜在的主题识别（Steinberger & Jezek, 2004）。
- **LexRank**。基于图的总结，根据句子图中的特征向量中心度概念计算句子的重要性（Erkan & Radev, 2004）。
- **KL-Divergence**.基于摘要和源语料库之间KL-分歧的最小化，使用贪婪搜索进行总结（Haghighi & Vanderwende, 2009）。
- **卢恩法**。启发式总结法，考虑到词频和句子在文章中的位置（Luhn, 1958）。

6.4 实验结果

在本节中，我们报告了亮点检测和亮点总结的实验结果。

亮点检测的结果

在我们的亮点检测模型中，切割一个词链l_{89n}的阈值被设定为11个词，概念映射的阈值为ϕMNO:49P设置为0.5，概念映射的阈值op_n设置为15，而参数λ用于控制情绪和概念集中的平衡，设置为0.9。第7节提供了一个参数分析。

表4是我们的方法和基准的不同组合的精确度、召回率和F1指标的比较。我们的完整模型(Spike+L+E+T)在所有指标上都优于其他基准。Random选择和统一选择的精度和召回率都很低，因为它们没有包含任何结构或内容信息。尖峰选择有相当大的改善，因为它利用了一个镜头的评论强度。然而，并非所有评论密集的镜头都是亮点。例如，在一个视频的开头和结尾的评论，通常是作为一种礼貌的高容量的问候和告别。另外，尖峰选择法通常将亮点集中在有大量评论的连续镜头上，而我们的方法可以跳到和分散到其他不那么密集但在情感或概念上集中的镜头。这可以从Spike+E+T的表现中看出。

我们还观察到，单独的滞后校正（Spike+L）大大改善了Spike-选择的性能，部分证实了我们的假设，即滞后校正在时间同步评论相关任务中很重要。

	精度		F-1
随机选择	0.1578	0.1587	0.1567
统一-选择	0.1775	0.1830	0.1797
尖峰选拔	0.2594	0.2167	0.2321
钉子+E+T	0.2796	0.2357	0.2500
穗+L	**0.3125**	0.2690	0.2829
钉子+L+E+T	0.3099	**0.3071**	**0.3066**

表4.亮点检测方法的比较。

亮点归纳的结果

在我们的亮点总结模型中，情感偏差b_{O8M52M}@被设定为0.3。

对1-gram BLEU, ROUGE的比较。我们的方法和基准的F1值见表5。我们的方法优于所有其他方法，特别是在ROUGE-1上。LSA的BLEU最低，主要是因为LSA在统计上偏向于长句和多词句，但是这些句子在时间同步的情况下没有代表性。

	BLEU-1	ROUGE-1	F1-1
LSA	0.2382	0.4855	0.3196
淘宝网	0.2854	0.3898	0.3295
KL-分歧	0.3162	0.3848	0.3471
卢恩	0.2770	0.4970	0.3557
LexRank	0.3045	0.4325	0.3574
我们的方法	**0.3333**	**0.6006**	**0.4287**

表5.亮点总结方法的比较（1克）。

词。SumBasic方法的表现也很差，因为它单独考虑了语义相关的词，而不像我们的方法那样使用概念而不是词。

表6列出了我们的方法和基准在2-gram BLUE, ROUGE和F1上的比较。我们的方法也优于所有其他方法。

从结果来看，我们认为在对时间同步的评论文本进行总结之前，进行滞后校准以及概念和情感映射是至关重要的。滞后校正将亲身经历的评论缩减到原来的镜头，防止不准确的亮点检测。概念和情感映射之所以有效，是因为时间同步的评论通常很短（平均7.2个字），评论的意义通常集中在一两个 "中心词 "上。

	BLEU-2	ROUGE-2	F1-2
淘宝网	0.1059	0.1771	0.1325
LSA	0.0943	0.2915	0.1425
LexRank	0.1238	0.2351	0.1622
KL-分歧	0.1337	0.2362	0.1707
卢恩	0.1227	0.3176	0.1770
我们的方法	**0.1508**	**0.3909**	**0.2176**

表6.亮点总结方法的比较（2-Gram）。

评论。情感图谱和概念图谱可以有效地防止生成的摘要中出现冗余。

7 参数的影响

7.1 射击长度的影响

我们分析了*镜头长度*对*F*1高光检测得分的影响。首先，从黄金分割的高光镜头长度的分布（图2），我们观察到大多数高光镜头长度位于[0,25]（se-conds）范围内，10秒为模式。因此，我们绘制了所有四个模型在5到23秒的不同镜头长度下的*F*1得分（图3）。

从图3中我们观察到：（1）我们的方法（Spike+L+E+T）在不同的镜头长度下一直优于其他基准；（2）然而，我们的方法相对于Spike方法的优势似乎随着镜头长度的增加而被削弱。这是合理的，因为随着镜头长度的增加，每个镜头中的评论数量也会增加。

镜头的积累。到了一定程度后，不管它包含什么样的情感和话题，有明显更多评论的镜头将被视为高光。然而，情况可能并不总是这样的。在现实中，当评论太少时，完全依靠数量的检测会失败；另一方面，当有大量的评论均匀地分布在镜头中时，峰值可能不是一个好的指标，因为

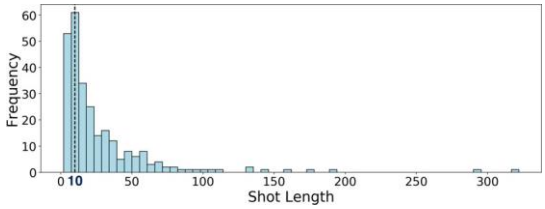

图2.突出黄金标准的射击长度分布。

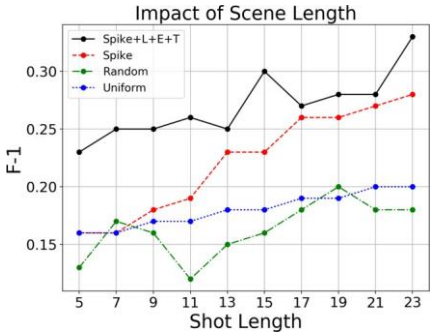

图3.镜头长度对高光检测的F-1得分的影响。

现在每个镜头都有同样大的评论量。此外，现实中的大部分高光都在15秒以内，图3显示我们的方法在这种更精细的水平上可以更准确地检测高光。

7.2 高光检测的参数

我们分析了四个参数对高光检测召回率的影响：词链的最大沉默度ıê·÷，概念映射的阈值jèÿ±ŷé'/，概念映射的邻居数top_n，以及情感和概念浓度的平衡λ（图4）。

从图4中，我们观察到以下情况。(1）当涉及到滞后校准时，似乎有一个最佳的*最大沉默长度*：11秒作为我们的数据集的最长空白延续链。这个值控制了一个词链的紧凑性。(2)
在概念映射中，*与现有概念的最小重合度*控制了概念合并的阈值，越高越好。

阈值越高，两个合并的概念就越相似。当重叠度增加到一定程度（在我们的数据集中为0.5）时，召回率就会增加，而在这一点之后就不会再有改善。(3)
在概念映射中，似乎有一个最佳的搜索*邻居数*（在我们的数据集中为15）。(4)
情感和概念集中度（*lambda*）之间的平衡更倾向于情感方面（在我们的数据集中为0.9）。

7.3 亮点总结的参数

我们还分析了*情感偏差*b±êèï¥èµ对ROGUE-1和ROGUE-
2的影响，用于高亮总结。结果在图5中描述。
从图5中，我们观察到，当涉及到高亮总结时，情感发挥了适度的作用（情感偏差=0.3）。这比它在高光检测任务中的作用要小，因为在高光检测任务中，情感的集中比概念的集中要重要得多。

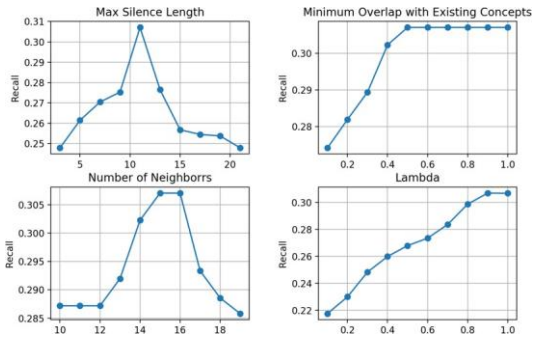

图4.高光检测参数的影响。

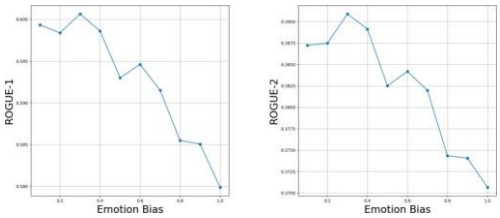

图5.高亮总结的参数的影响。

8 总结

在本文中，我们提出了一个新颖的无监督框架，用于基于众包时间同步评论的视频亮点检测和总结。对于亮点检测，我们开发了一种滞后校准技术，可以缩小滞后的在概念映射的词汇链的基础上，将评论返回到其原始场景。此外，通过对每个镜头中的评论强度和概念情感的集中度进行评分来检测视频亮点。对于高光部分的总结，我们提出了一个两级的SumBasic，在每次迭 代 选择句子时同时更新单词和概念的概率。在未来，我们计划为高光检测整合多种信息来源，如视频元数据、观众资料，以及通过视频处理的多种模式的低级特征。

参考文献

Barzilay, R., & Elhadad, M. (1999).使用词汇链进行文本总结。Advances in auto-matic text summarization, 111-121.

Ehren, R. (2017).字面意思还是习惯用语？使用词嵌入识别德语多词表达的单次出现的阅读。计算语言学协会欧洲分会第15届会议学生研究研讨会论文集，103-112。

Ekman, P. (1992).关于基本情感的论证。认知与情感, 6(3-4), 169-200.

Erkan, G., & Radev, D. R. (2004).Lexrank:基于图的词汇中心性作为文本总结的显著性。Journal of Artificial Intelligence Research, 22, 457-479.

Goldberg, Y., & Levy, O. (2014). word2vec explained:衍生出Mikolov等人的负采样词嵌入方法。 arXiv预印本arXiv:1402.3722。

Haghighi, A., & Vanderwende, L. (2009).探索多文档总结的内容模型。在《人类语言技术》会议上发表的论文。2009年计算语言学协会北美分会的年会。

Hanjalic, A., & Xu, L.-Q.(2005).情感视频内容表示和建模.IEEE transactions on multimedia, 7(1), 143-154.

Hirst, G., & St-Onge, D. (1998).词汇链作为语境的代表，用于检测和纠正误导性语言。WordNet:WordNet: An electronic lexical database, 305, 305-332.

Ikeda, A., Kobayashi, A., Sakaji, H., & Masuyama, S. (2015).nico nico douga上的评论分类，用于基于参考内容的注释。在基于网络的信息系统（NBiS），2015年第18届国际会议上发表的论文。

Lin, C., Lin, C., Li, J., Wang, D., Chen, Y., & Li, T.（2012）。从MI-CROBOGs生成事件故事情节。在第21届ACM国际信息与知识管理会议上发表的论文。

Lin, C.-Y.(2004).Rouge:一个用于自动评估摘要的软件包.论文发表于文本总结的分支。ACL-04研讨会论文集.

Lin, K. -S., Lee, A., Yang, Y. -H., Lee, C. -T., & Chen, H.H.(2013).利用音乐情感和人脸特征自动提取戏剧视频的亮点.Neurocomputing, 119, 111-117.

Luhn, H. P.(1958).文献摘要的自动创建。IBM研究与发展杂志，2（2），159-165。

Lv, G., Xu, T., Chen, E., Liu, Q., & Zheng, Y. (2016).阅读视频:基于语义嵌入的众包时间同步视频的时间标签。在AAAI会议上发表的论文。

Morris, J., & Hirst, G.(1991).词汇凝聚力由词典关系作为文本结构的一个指标。计算语言学，17（1），21-48。

Nenkova, A., & Vanderwende, L.(2005).频率对总结的影响。Microsoft Re-search, Redmond, Washington, Tech.MSR- TR-2005, 101.

Ngo, C.-W., Ma, Y.-F., & Zhang, H.-J. (2005).视频通过图形修改进行总结和场景检测。IEEE Transactions on Circuits and Systems for Video Technology, 15（2）, 296-305.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J.(2002).BLEU: a method for automatic evaluation of machine translation.在第40届计算语言学协会年度会议上发表的论文。

Sakaji, H., Kohana, M., Kobayashi, A., & Sakai, H.(2016).通过Nico Nico Douga上的评论估计标签。在基于网络的信息系统（NBiS），2016年第19届国际会议上发表的论文。

Shou, L., Wang, Z., Chen, K., & Chen, G.(2013).Sumblr：对不断变化的推文流进行连续总结。论文发表在第36届ACM SIGIR国际信息研究与发展会议的论文集上。

Sipos, R., Swaminathan, A., Shivaswamy, P., & Joa-chims, T.（2012）。使用子模态词汇覆盖的时态语料库总结。在第21届ACM国际信息与知识管理会议上发表的论文。

Steinberger, J., & Jezek, K.(2004).在文本总结和总结评估中使用潜在语义分析。论文发表于Proc.ISIM'04。

Tran, T. A., Niederée, C., Kanhabua, N., Gadiraju, U., & Anand, A.(2015).平衡新颖性和突出性。自适应学习为高影响事件的时间线总结排列实体。论文在第24届ACM Interna- tional on Conference on Information and Knowledge Management会议上预发。

Wu, B., Zhong, E., Tan, B., Horner, A., & Yang, Q.(2014).众包的时间同步视频标签，使用时间和个性化的主题建模。在第20届ACM SIGKDD国际知识发现和数据挖掘会议上发表的论文。

Xian, Y., Li, J., Zhang, C., & Liao, Z.(2015).视频高光镜头提取与时间同步通信。论文发表在第七届地球规模的移动计算和在线社交网络热点问题国际研讨会上。

Xu, L., & Zhang, C.(2017).衔接视频内容和评论。同步视频描述与众包时间同步评论的时空总结。在第三届AAAI人工智能会议上发表的论文。

Xu, M., Jin, J. S., Luo, S., & Duan, L.(2008).基于唤醒和情绪特征的层次化电影情感内容分析。论文发表于第16届ACM国际多媒体会议论文集。

Xu, M., Luo, S., Jin, J. S., & Park, M.(2009).通过多模态下的中层表征进行有感情的内容分析.论文发表于第一届互联网多媒体计算与服务国际会议论文集。

Yan, R., Kong, L., Huang, C., Wan, X., Li, X., & Zhang, Y.(2011).通过进化的跨时空总结生成时间线。论文在《自然语言处理的经验方法》会议上预发。

Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X., & Zhang, Y.(2011).进化的时间线总结：通过迭代替代的平衡优化框架。在第34届ACM SIGIR信息检索研究与发展国际会议上发表的论文。