arXiv:2303.08318v1 [cs.MM] 15 Mar 2023

通过联合建模社交影响和标签关系进行微视频标注

Xiao Wang
scz.wangxiao@gmail.com 山东大学

Tian Gan*
gantian@sdu.edu.cn
山东大学

Yinwei Wei
weiyinwei@hotmail.com 新加坡国立大学

Jianlong Wu*
jlwuhust@gmail.com 哈尔滨工业大学（深圳） & 山东大学

Dai Meng
daimeng@kuaishou.com 快手科技

Liqiang Nie
nieliqiang@gmail.com 哈尔滨工业大学（深圳）

摘要

过去十年间，微视频在各类用户生成内容平台上迅速涌现。根据我们的统计，约有 85.7% 的微视频缺少标注。本文聚焦于为微视频添加标签。现有方法大多侧重分析视频内容，忽视了用户的社交影响和标签关系。同时，现有的标签关系构建方法要么性能不足，要么标签覆盖率较低。为联合建模社交影响和标签关系，我们将微视频标签任务表述为在所构建异构网络中的链接预测问题。具体而言，我们以半监督方式构建标签关系（以标签本体表示）。随后，我们结合标签关系、视频-标签标注和用户关注关系来构建网络。之后，通过行为传播建模以及视觉与语言知识聚合，得到更优的视频和标签表示。最后，在该视频-标签网络中计算每个微视频与所有候选标签之间的语义相似度。在三个垂直领域的工业数据集上进行的大量实验表明，与若干最先进的基线方法相比，我们的模型具有优越性。

CCS 概念

• **信息系统** →**文档表示**。

关键词

微视频标签；行为传播；本体构建

1 引言

过去十年间，用户生成内容（UGC）平台上的微视频蓬勃发展。为加强搜索和推荐等应用，[25, 34, 35], 标签被广泛用于概括微视频。考虑到用户上传微视频时可能不会添加足够的标签，因此，标签标注已成为用户生成内容平台运营团队的一项繁重日常工作。根据对快手平台收集的 6 亿多个视频进行的统计，其中约有 85.7% 的视频

完全没有标签。为了以尽可能少的人工投入生成标签，自动微视频标注受到工业界和学术界的广泛关注。大多数现有方法将视频标签视为一个多标签分类问题。



图 1：社交影响与标签关系示例。社交影响：用户 A 是用户 B 的关注者，并模仿用户 B 创作带有相似标签的视频（以斜体表示）。

基础方法利用视频内容和描述中的信息进行标签标注 [15, 36]。更复杂的方法则利用额外信息来辅助标注，例如标签图 [3, 9, 14, 38], 查询日志 [21], 用户行为 [2],和用户档案 [11, 33]。

然而，用户的社交影响和标签关系鲜少受到讨论，而它们本可以发挥关键作用。一方面，用户的社交影响至关重要，因为用户可能会模仿其社交网络邻居来创作相似的微视频，尤其是在用户生成内容平台上，因为用户既是内容创作者，也是内容消费者。以图 1 为例，用户 B 创作了一段关于“草莓蛋糕”的微视频，随后其关注者用户 A 模仿这段视频，创作了一段带有几乎相同标签集合的类似视频。社交网络中的这种模仿现象，即**行为传播** [1], ，若能得到恰当建模，便可提升标签预测性能。另一方面，标签关系体现为**标签本体**，它通常表示为由标签构成的有向无环图（DAG），其中标签之间存在是_子主题_的关系。标签本体从三个方面提供外部知识，促进标签预测：(i) 标签依赖关系 [32, 38], (ii) 自顶向下知识迁移 [11], 以及 (iii) 自底向上语义抽象 [22]。

为纳入用户的社交影响和标签关系，我们构建了一个异构用户-视频-标签网络。随后利用这一异构图神经网络，获得更好的视频和标签表示。因此，我们能够衡量视频与标签之间的相似度，从而得到标签标注结果。上述方法仍面临以下挑战：**C1：行为传播建模**。我们需要提出一种有效的模型，为行为传播建模过滤无关信息，因为并非所有微视频都是通过模仿产生的。**C2：视觉语言知识聚合**。学习标签表示有两个知识来源：视觉

*通讯作者：Tian Gan 和 Jianlong Wu。

知识来自视频，语言知识则来自标签本体中的相关标签。现有聚合方法 [32, 38] 采用向量拼接或注意力机制等直接聚合方法，但由于忽略了共有知识中的冗余，因此效果不理想。具体而言，拼接会重复共有知识，而注意力机制往往过度关注独有知识 [13, 19, 28]。**C3：标签本体构建**。大多数方法 [11, 32] 从现有开放知识库中获取所需的标签本体。然而，由于用户生成内容平台上的视频内容变化极快，开放知识库只能覆盖部分标签。其他方法 [3, 6] 尝试根据标签统计构建标签本体。这些方法基于规则，因此准确性不足。

为解决 **C3**，我们基于人工设计特征，采用半监督分类构建标签本体。随后，我们设计了一种对抗聚合图 Transformer 网络（**RADAR**），在整个图上传播信息，以更好地表示每个微视频和标签节点。RADAR 是一种异构图神经网络（GNN），由门控图变换器（GGT）和对抗聚合网络（AAN）组成。GGT 旨在通过聚合邻居的信息并过滤无关信息来解决 **C1**。为应对**C2**，我们采用 AAN 聚合视觉知识和语言知识，去除冗余信息，同时保留互补信息。最后，我们计算给定微视频与所有候选标签之间的语义相似度，并将其作为预测结果。为了验证模型的合理性，我们在大规模真实世界数据集上进行了广泛实验。实验结果证明了 GGT 和 ANN 的有效性，而且我们的方法始终优于多种最先进的基线方法。

总而言之，本研究的贡献主要有以下三点：

- 我们在一个 UGC 平台上构建了一个包含 45 万条微视频及其创作者社交网络的数据集。实验表明，社交网络有助于视频标签。据我们所知，这是首项针对微视频标签对用户社交影响进行建模的研究。
- 我们提出了一种半监督方式来构建标签本体，只需少量人工投入，效果便优于现有方法。
- 我们设计了 RADAR，这是一种由 GGT 和 AAN 组成的异构图神经网络，能够以端到端方式联合建模用户的社交影响和标签关系。在真实世界数据集上，RADAR 的表现显著优于当前最先进的方法。我们的代码已发布[1]。

2 相关工作

2.1 视频标签

视频标签旨在找出能够描述视频核心内容的关键词。根据所使用的信息，所有方法可分为两类。**基础方法**仅利用视频内容中的信息：文本、视觉和音频信息。这类方法又分为两种子类型。关键词短语提取方法从视频描述中提取标签 [23, 24, 37],，并根据其与核心语义的相关性对标签进行排序 [10, 24]。多标签分类方法则从预定义词集中为视频分配标签 [4, 15, 36]。**复杂方法**利用视频内容之外的额外信息，例如结合知识库 [3, 9, 14, 38], 和查询日志构建的标签图 [21], 用户行为 [2], 和用户档案 [11, 33]。在上述所有额外信息中，标签图最受关注。通常，标签图是根据标签共现 [3, 32, 38] 或共有知识库 [11, 12, 32]构建的。然而，前者只能提供统计关系，而后者的标签覆盖率较低。与标签图不同，社交网络信息（具体而言是用户关注关系）尚未应用于视频标签。

2.2 异构图神经网络

异构网络是包含多种元路径的图。异构图神经网络被广泛用于为这类网络推导出更好的节点表示 [5, 17, 39, 40]。例如，Wang 等人 [31] 提出了一种基于两级注意力的方法。具体而言，节点级注意力衡量每条元路径中邻居的重要性，而语义级注意力衡量不同元路径的重要性。Hu 等人[8] 提出了一种方法，使用节点和边类型相关参数来参数化每条元路径的图变换器网络。Lv 等人 [20] 提出了一种方法，证明简单的图注意力网络结合三种广为人知的技术，可以胜过此前所有模型。

3 问题定义

我们将微视频标签任务表述为视频-标签网络 $G(\mathrm{V}, \mathrm{E})$ 上的链接预测问题。V 是由视频和标签组成的节点集合，E 是边集合。请注意，基于图数据的视频标签任务也可以表述为知识图谱嵌入问题 [9]。然而，这种表述仅支持传导式学习，无法处理新视频。

对于每个节点 $u \in \mathrm{V}$，我们将其初始表示记为 $\mathrm{h}^0(u) \in \mathbb{R}^d$，其中 d 为隐藏维度的大小。对于标签节点，h^0 由词嵌入与可学习嵌入相加得到。对于视频节点，h^0 由聚合后的帧嵌入得到。给定 h^0 和 G，我们的 RADAR 模型——一个 L层异构图神经网络——可以推导出更优的节点表示 $\mathrm{h}^L \in \mathbb{R}^d$：

$$\mathrm{h}^L = \mathrm{RADAR}\left(\mathrm{h}^0, G\right). \tag{1}$$

基于该表示，我们预测置信度分数 $\hat{y}(v, t) \in \mathbb{R}$，即微视频 $v \in \mathrm{Vvideo}$ 带有标签 $t \in \mathrm{Vtag}$ 的置信度：

$$\hat{y}(v,t) = \mathrm{Sigmoid}\left(\mathrm{h}^L(v) \cdot \left[\mathrm{h}^L(t)\right]^T\right). \tag{2}$$

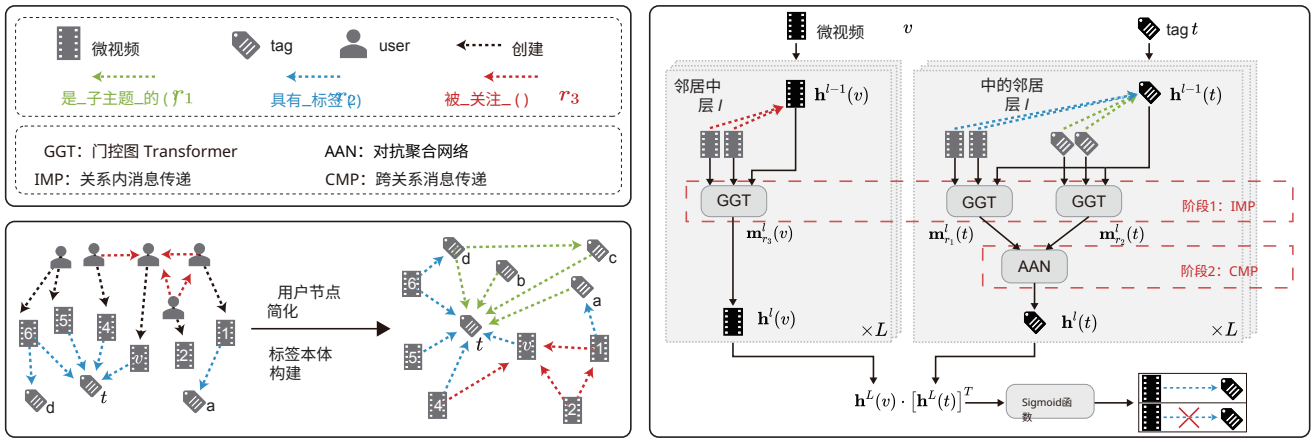
4 方法

我们的整体框架如图 2 所示。首先，我们提出一种半监督分类方法来构建标签本体，以解决本体构建难题。随后，我们构建异构视频-标签网络，以联合建模社交影响和标签本体。之后，我们设计对抗聚合门控图变换器网络（RADAR），以推导出更优的视频和标签表示。RADAR 的两个组件分别解决行为传播建模和视觉-语言知识聚合难题。最后，我们估计微视频与所有标签之间的语义相似度，并将其作为预测结果。

4.1 标签本体构建

在本小节中，我们构建标签本体，它是由标签和是_子主题_的 关系组成的图。它天然是一个有向无环图（DAG），因为由是_子主题_的 关系构成的环

[1]https://github.com/SCZwangxiao/RADAR-MM2022.git



(a) 视频-标签网络构建 (b) RADAR 架构

图 2：方法概览。(a) 根据原始数据构建视频-标签网络，原始数据由用户及其标注了标签的微视频组成。(b) 我们的 RADARnetwork 在每一层中的图神经网络传播分为两个阶段。

会导致悖论。本体构建包括两个步骤：子主题发现和 DAG 构建，如图 3 所示。在子主题发现过程中，我们基于人工设计特征，使用半监督分类来发现标签之间的子主题关系。这些特征源自视频标签标注中的标签统计，因此能够覆盖所有标签。此步骤将生成一个 DAG 违反图。随后，我们施加图约束，得到一个有向无环图，作为最终的标签本体。

4.1.1 子主题发现。我们假设如果两个标签存在子主题关系，它们至少会在一个微视频中共现。基于这一假设，我们将子主题发现表述为一个针对共现标签的二分类问题。具体而言，对于每一对共现标签 $(u,\ v)$，$u,\ v \in \text{Vtag}$，我们预测其子主题得分 $r(u,\ v)$:

$$r(u,v) = g\left(\mathrm{k}(u,v)\right), \tag{3}$$

$$r(u,v) = \begin{cases} 1 & \text{if } v \text{ is a subtopic of } u, \\ 0 & \text{if no relationship between } u \text{ and } v, \end{cases} \tag{4}$$

其中，$g(\cdot)$ 可以是任意分类器，$\mathrm{k}(u,\ v) \in \mathbb{R}^{d_r}$ 是我们手工设计的特征，d_r 是特征的维度。这些特征旨在反映标签之间的以下两种属性:

语义重叠 指两个标签之间含义的重叠。如果两个标签 u 和 v 存在子主题关系，它们必然存在语义重叠。我们使用点互信息 $\mathrm{PMI}(u,v) \in \mathbb{R}$来衡量语义重叠:

$$\mathrm{PMI}(u,v) = \log\left(\frac{p(u,v)}{p(u)p(v)}\right), \tag{5}$$

以及点式 Kullback–Leibler 散度 $\mathrm{PKL}(u,v) \in \mathbb{R}$:

$$\mathrm{PKL}(u,v) = p(u,v)\log\left(\frac{p(u,v)}{p(u)p(v)}\right), \tag{6}$$

其中，$p(u,v)$表示 u,v出现在同一条微视频中的概率，而 $p(u)$ $p(v)$分别表示 u,v的出现概率。与 PMI 相比，由于额外考虑了 $p(u,v)$，PKL 较少偏向于出现频率较低的标签。

语义广度是指标签含义的广泛程度。语义越广，表示其在标签本体中的层级越高（例如，“食物”的语义比“蛋糕”更广）。

我们使用两个特征来衡量这一点。其中一个是标签转移概率 $p(u|v) \in \mathbb{R}$ 从 v到 u:

$$p(u|v) = \frac{|\text{videos that have tag } u \text{ and } v|}{|\text{videos that have tag } v|}. \tag{7}$$

直观而言，$p(u|v)$越大，v越可能是 u的子主题。例如，已知标签“蛋糕”是“食物”的子主题，且使用了“蛋糕”，那么也很可能会看到“食物”（即，p(食物|蛋糕)较大）。实践证明，该概率在工业级概念挖掘系统 [16]中十分有效。另一个特征是标签熵 $H(u) \in \mathbb{R}$:

$$H(u) = -\sum_{w} p(u|w)\log(p(u|w)). \tag{8}$$

直观而言，H（u）较高的标签会从其他标签接收更多、更强的流入转移，这表明它在标签本体中处于更高层级。FBVO 也使用 H（u）来推导视觉本体 [6]。

结合上述语义重叠和广度特征，我们可以定义手工设计的特征 $\mathrm{k}(u,\ v)$，其中包含六个基本特征：p（$u|v$）、p（$v|u$）、H（u）、H（v）、PMI（$u,\ v$）、PKL（$u,\ v$），以及两个二阶特征：log（p（$u|v$）/p（$v|u$））、log（H（u）/H（v））。基于这些特征，式（3）中的分类以半监督方式进行。具体而言，标签来自一小部分共现标签对。同时，特征可以根据所有共现标签计算，从而利用未标注数据中的信息。

对于每个标签对（$u,\ v$），我们可以估计 v 是 u 子主题的置信度分数：$\hat{r}$（$u,\ v$）。我们保留置信度分数高于阈值 δ_r 的标签对，并将它们连接成一个有向图。

4.1.2 有向无环图构建。由于并非所有预测都正确，因此必然会出现有向无环图违例，如图 3(b) 所示。因此，我们施加图约束以推导出最终的有向无环图。

如图 3(c) 所示，我们首先按照熵 H降序排列所有标签。考虑到熵越高意味着在标签本体中的层级越高，我们随后仅保留从熵较低的标签指向熵较高标签的是_子主题_的 关系。之后，排序后的标签列表可视为有向无环图的拓扑排序，并可作为我们的标签本体。请注意，可能会存在孤立的连通分量，因为在保留 $\hat{r}(u,\ v)$ $> \delta_r$ 时会过滤掉大量边

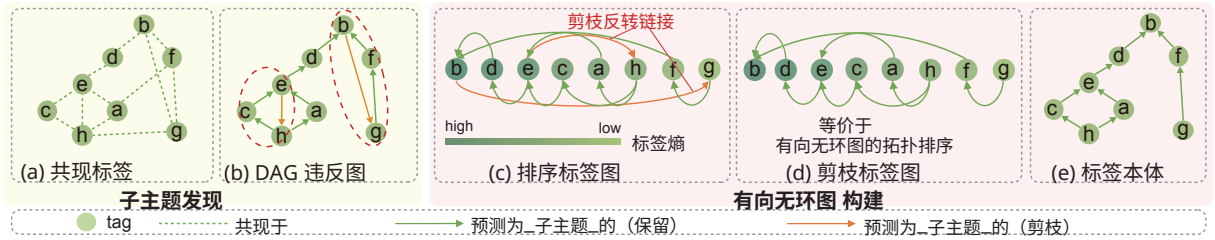


图 3：标签本体构建。(a)(b) 是_子主题_的关系被发现，并形成一个有向无环图违反图。(c) 对所有标签进行排序，然后剪除反向链接。(d)(e) 剪枝标签图等价于一个有向无环图，即我们的标签本体。

在子主题发现中。对于这些连通分量，我们会放宽约束，直到它们不再孤立或 $\hat{r}(u,\ v) < \epsilon_r$，并将其余孤立分量连接到熵最高的标签。

4.2 视频-标签网络

我们构建了一个有向异构视频-标签网络，以纳入社交网络和标签本体，如图 2(a) 所示。

我们的原始数据由用户、微视频和标签组成，并包含两种关系：具有_标签 和被_关注_。后一种关系为行为传播提供了线索。

我们首先通过删除用户节点，并让用户创建的视频继承被_关注_关系 来简化原始数据。具体来说，如果用户$_A$ 被用户$_B$关注，那么用户$_A$的所有视频都会被用户$_B$的所有视频关注。请注意，我们只保留从先前上传的视频指向新上传视频的被_关注_关系，因为只有旧视频才能影响新视频。简化后，我们纳入从构建的标签本体中得到的是_子主题_的 关系，从而得到视频-标签网络。

形式化地，我们将微视频集合记为 Vvideo，将标签集合记为 Vtag。它们构成节点集合 $\mathrm{V} = \mathrm{Vvideo} \cup \mathrm{Vtag}$。为方便起见，我们将三种关系是_子主题_的、具有_标签和被_关注_的 分别记为 r_1，r_2 和 r_3。它们对应的关系集分别为 E_{r_1}、E_{r_2}、E_{r_3}。这些关系集构成边集 $\mathrm{E} = \mathrm{E}_{r_1} \cup \mathrm{E}_{r_2} \cup \mathrm{E}_{r_3}$。

4.3 RADAR 架构

4.3.1 RADAR 总体架构。为了获得更好的视频和标签表示，我们设计了 RADAR，这是一种 L 层图神经网络。考虑到标签节点有两种类型的入边，而视频节点只有一种，我们对它们采用不同的 GNN 传播策略。具体而言，我们将传播分为两个阶段：对视频节点和标签节点进行关系内消息传递，以及仅对标签节点进行跨关系消息聚合。相应地，如图 2(b) 所示，RADAR 中的每个 GNN 层由两个组件构成：**门控图 Transformer**（GGT）和**对抗聚合网络**（AAN）。

在关系内消息传递 阶段，对标签和视频节点均应用 GGT，以获取从其邻居传播而来的消息信息。形式上，视频 v 接收由其 r_3 邻居传递的消息 $\mathrm{m}^l_{r_3}(v) \in \mathbb{R}^d$：

$$\mathrm{m}^l_{r_3}(v) = \mathrm{GGT}\left(\mathrm{h}^{l-1}(v), \left\{\mathrm{h}^{l-1}(s) | s \in \mathcal{N}^l_{r_3}(v)\right\}\right), \tag{9}$$

视频表示通过以下方式更新：

$$\mathrm{h}^l(v) = \mathrm{m}^l_{r_3}(v), \tag{10}$$

其中，上标 l 表示第 l 层，$\mathrm{h}^l(v) \in \mathbb{R}^d$ 是节点 v 的输出表示，$\mathrm{N}^l_{r_3}(v)$ 是节点 v 在 r_3 关系下的邻居。同时，标签 t 分别接收来自其 r_1 和 r_2 邻居传递的消息 $\mathrm{m}^l_{r_1}(t)$、$\mathrm{m}^l_{r_2}(t)$：

$$\begin{cases} \mathrm{m}^l_{r_1}(t) = \mathrm{GGT}\left(\mathrm{h}^{l-1}(t), \left\{\mathrm{h}^{l-1}(s) | s \in \mathcal{N}^l_{r_1}(t)\right\}\right), & (11) \\ \mathrm{m}^l_{r_2}(t) = \mathrm{GGT}\left(\mathrm{h}^{l-1}(t), \left\{\mathrm{h}^{l-1}(s) | s \in \mathcal{N}^l_{r_2}(t)\right\}\right), & (12) \end{cases}$$

其中，$\mathrm{N}^l_{r_1}(t)$，$\mathrm{N}^l_{r_2}(t)$分别是节点 t在 r_1 和 r_2 关系下的邻居。

对于每个视频，邻域消息携带了有关是_关注_的邻居对其社交影响的信息。由于并非所有视频都是通过模仿产生的，在 GGT 中，我们使用门控机制过滤无关信息，以应对行为传播建模挑战。对于每个标签 $t \in \mathrm{Vtag}$，邻域消息分别从是_子主题_的邻居和具有_标签的邻居中获取语言和视觉知识。因此，我们提出 AAN，以在下一阶段解决视觉-语言聚合挑战。

在跨关系消息聚合阶段，仅对标签节点应用 AAN，以聚合子主题消息 $\mathrm{m}^l_{r_1}(t)$ 和视频消息 $\mathrm{m}^l_{r_2}(t)$。对于每个标签 $t \in \mathrm{Vtag}$，其表示更新如下：

$$\mathrm{h}^l(t) = \mathrm{AAN}\left(\mathrm{m}^l_{r_1}(t), \mathrm{m}^l_{r_2}(t)\right). \tag{13}$$

在每个图神经网络层中，按照上述两个阶段更新节点表示后，我们得到最终表示 h^L。

4.3.2 门控图 *Transformer*。 GGT 的目标是推导出在第 l 层从邻居传播而来的消息信息 $\mathrm{m}^l_r(t) \in \mathbb{R}^d$，其中 r 是邻居关系。如图 4(a) 所示，我们采用图 Transformer 聚合邻域信息，并使用门控残差网络过滤无关信息。具体而言，我们将中心节点记为目标节点 t，将邻居节点记为源节点 $s \in \mathrm{N}^l_r(t)$。GGT 由目标节点类型 a_t、源节点类型 a_s 和邻居关系 r 参数化。

受 Transformer架构 [8, 27], 的启发，我们首先将目标节点 t 作为查询 $\mathrm{q}^l_r(t) \in \mathbb{R}^d$，并将源节点 $s \in \mathrm{N}^l_r(t)$

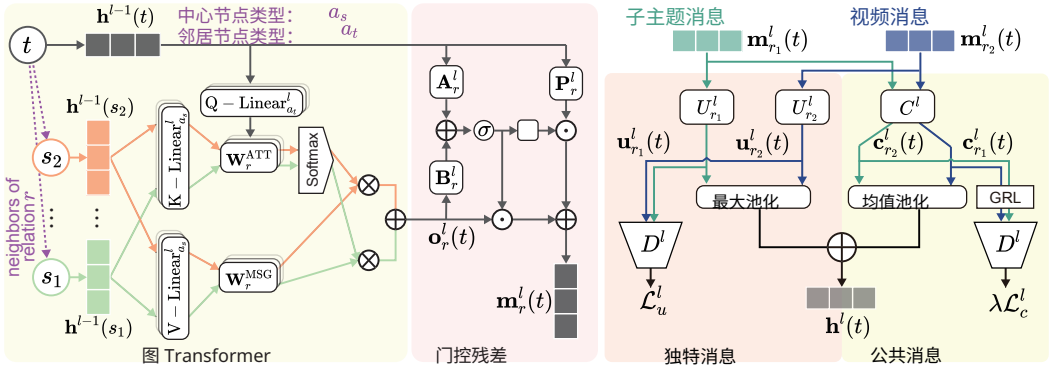


图 4：(a) GGT 使用图 Transformer 聚合邻居信息，并使用门控残差网络过滤无关信息。(b) ANN 将子主题消息和视频消息分离为共有信息和独特信息，并对其进行聚合。

同时作为键 $\mathrm{k}_r^l(s) \in \mathbb{R}^d$ 和值 $\mathrm{v}_r^l(s) \in \mathbb{R}^d$:

$$\begin{cases} \mathrm{q}_r^l(t) = \text{Q-Linear}_{a_t}^l\left(\mathrm{h}^{l-1}(t)\right), & (14) \\ \mathrm{k}_r^l(s) = \text{K-Linear}_{a_s}^l\left(\mathrm{h}^{l-1}(s)\right), & (15) \\ \mathrm{v}_r^l(s) = \text{V-Linear}_{a_s}^l\left(\mathrm{h}^{l-1}(s)\right)\mathrm{W}_r^{\text{MSG}}, & (16) \end{cases}$$

其中，$\text{Q-Linear}_{a_t}^l$、$\text{K-Linear}_{a_s}^l$、$\text{V-Linear}_{a_s}^l$ 是对应特定节点类型的线性变换层，而 $\text{WMSG}_r \in \mathbb{R}^{d\times d}$ 是关系 r 的消息矩阵。

然后，我们计算查询与键之间的点积，并应用 Softmax 来衡量源节点 s 相对于目标节点 t 的重要性，如 $\alpha_r^l(t, s) \in \mathbb{R}$:

$$\alpha_r^l(t,s) = \frac{exp\left(\mathrm{q}_r^l(t)\mathrm{W}_r^{\text{ATT}}\mathrm{k}_r^l(s)^T/\sqrt{d}\right)}{\sum_{w\in\mathcal{N}_r^l(t)} exp\left(\mathrm{q}_r^l(t)\mathrm{W}_r^{\text{ATT}}\mathrm{k}_r^l(w)^T/\sqrt{d}\right)}, \quad (17)$$

其中，$\text{WATT}_r \in \mathbb{R}^{d\times d}$ 是基于边的投影矩阵。请注意，Softmax 在同一关系 r的源节点范围内计算，我们将其称为分离注意力。这不同于现有图 Transformer [8]中的互注意力，后者在所有关系的邻居范围内进行归一化。我们采用分离注意力，因为视频-标签网络中的不同关系差异很大，无法将它们放在一起加权。

随后，我们可以通过对值进行加权求和，得到图 Transformer 的输出 $\mathrm{o}^l(t) \in \mathbb{R}^d$:

$$\mathrm{o}_r^l(t) = \sum_{s\in\mathcal{N}_r^l(t)} \alpha_r^l(t,s)\mathrm{v}_r^l(s). \quad (18)$$

最后，我们提出<style id='1'>门控残差网络 来推导邻域消息 $\mathrm{m}_r^l(t)$:

$$\begin{cases} \mathrm{z}_r^l(t) = \text{Sigmoid}\left(\mathrm{A}_r^l\mathrm{h}^{l-1}(t) + \mathrm{B}_r^l\mathrm{o}_r^l(t)\right), & (19) \\ \mathrm{m}_r^l(t) = \mathrm{z}_r^l(t)\odot\mathrm{o}_r^l(t) + \left(1-\mathrm{z}_r^l(t)\right)\odot\left(\mathrm{P}_r^l\mathrm{h}^{l-1}(t)\right), & (20) \end{cases}$$

其中，$\mathrm{z}_r^l(t) \in \mathbb{R}^d$ 是门控值，$\odot$ 表示逐元素乘法，A_r^l、B_r^l、$\mathrm{P}_r^l \in \mathbb{R}^{d\times d}$ 是关于第 (l) 层中关系 r的线性变换矩阵。门控机制通常会过滤无关信息。对于是_接着_由邻居，当视频并非通过模仿创建时，它会抑制这些邻居。对于是_子主题_的 邻居，它只保留一般语义，作为对子主题标签的抽象。

为简洁起见，我们省略了 P_r^l 后的 GELU 激活函数、A_r^l、B_r^l 和 P_r^l 后的偏置向量，以及 WMSGr 和 WATTr 中的上标 l。

4.3.3 对抗聚合网络。 AAN 专为标签节点设计，用于分别从子主题消息和视频消息中聚合语言知识和视觉知识。这两种消息都可分为独特信息（仅存在于其中一种消息中）和共有信息（同时存在于两种消息中）。拼接或注意力机制 [31] 等现有方法表现欠佳，因为它们忽略了共有信息中的冗余。具体而言，拼接会重复共有信息，而注意力机制往往会忽略独特信息 [13, 19]。

在我们的 AAN 中，我们提出将共有信息与独特信息分离，以去除冗余信息，同时保留互补信息。形式化地，对于每个目标标签节点 $t \in$ Vtag，首先将传入消息 $\mathrm{m}_{r_1}^l(t)$ 和 $\mathrm{m}_{r_2}^l(t)$ 分离为公共消息 $\mathrm{c}_{r_1}^l(t) \in \mathbb{R}^d$、$\mathrm{c}_{r_2}^l(t) \in \mathbb{R}^d$:

$$\begin{cases} \mathrm{c}_{r_1}^l(t) = C^l\left(\mathrm{m}_{r_1}^l(t)\right), & (21) \\ \mathrm{c}_{r_2}^l(t) = C^l\left(\mathrm{m}_{r_2}^l(t)\right), & (22) \end{cases}$$

以及独特消息 $\mathrm{u}_{r_1}^l(t)$、$\mathrm{u}_{r_2}^l(t)$:

$$\begin{cases} \mathrm{u}_{r_1}^l(t) = U_{r_1}^l\left(\mathrm{m}_{r_1}^l(t)\right), & (23) \\ \mathrm{u}_{r_2}^l(t) = U_{r_2}^l\left(\mathrm{m}_{r_2}^l(t)\right), & (24) \end{cases}$$

其中，$C^l(\cdot)$、$U_{r_1}^l(\cdot)$ 和 $U_{r_2}^l(\cdot)$ 是用于分别学习共享特征、独有特征以及*is_子主题_of*和*has_*标签关系的线性变换层。

为确保上述分离成功，我们需要使 $\mathrm{c}_{r_1}^l(t)$、$\mathrm{c}_{r_2}^l(t)$ 的分布尽可能接近，同时使 $\mathrm{u}_{r_1}^l(t)$、$\mathrm{u}_{r_2}^l(t)$ 的分布尽可能远离。有两种可选方法。一种是利用互信息衡量分布之间的距离。其缺点在于，互信息的上下界只能通过估计得到。因此，我们采用另一种方法，即对抗训练，它基于

表 1：数据集统计。

数据集	#视频	#Tag	#标签 /视频	#关注者 /视频	#视频 /Tag	#子主题 /Tag
Apparel	245,353	4,817	5.1	468.4	259.8	3.0
化妆品	96,924	3,114	4.0	101.5	124.5	2.2
food	108,206	8,640	4.3	24.6	53.8	2.4

极小极大博弈。我们使用判别器来区分两个独有特征 $\mathrm{u}_{r_1}^l(t)$、$\mathrm{u}_{r_2}^l(t)$，同时使其无法区分公共特征 $\mathrm{c}_{r_1}^l(t)$、$\mathrm{c}_{r_2}^l(t)$。形式化地，我们有：

$$\min_{D^l,U_{r_1}^l,U_{r_2}^l}\max_{C^l}\mathcal{L}_{\text{adv}}^l=\mathcal{L}_u^l+\lambda\mathcal{L}_c^l, \tag{25}$$

$$\begin{cases}\mathcal{L}_u^l=-\log\left(D^l\left(\mathrm{u}_{r_1}^l(t)\right)\right)-\log\left(1-D^l\left(\mathrm{u}_{r_2}^l(t)\right)\right), & (26)\\ \mathcal{L}_c^l=-\log\left(D^l\left(\mathrm{c}_{r_1}^l(t)\right)\right)-\log\left(1-D^l\left(\mathrm{c}_{r_2}^l(t)\right)\right), & (27)\end{cases}$$

其中 D^l 是作为判别器的二元线性分类器，λ 是权衡参数，L_c^l，L_u^l 分别是针对共享特征和独有特征的判别损失。

经过共享-独有分离后，我们直接将它们聚合起来，得到标签节点 t的输出表示 $\mathrm{h}^l(t)$：

$$\begin{cases}\mathrm{c}^l(t)=\text{MeanPool}\left(\mathrm{c}_{r_1}^l(t)+\mathrm{c}_{r_2}^l(t)\right), & (28)\\ \mathrm{u}^l(t)=\text{MaxPool}\left(\mathrm{u}_{r_1}^l(t)+\mathrm{u}_{r_2}^l(t)\right), & (29)\\ \mathrm{h}^l(t)=\frac{1}{2}\left(\mathrm{u}^l(t)+\mathrm{c}^l(t)\right). & (30)\end{cases}$$

4.3.4 训练与推理。我们计算视频与所有标签之间的语义相似度，并将其作为预测结果。具体而言，对于每个视频节点 v 和标签节点 t，节点 v是否具有标签 t的概率估计为 $\hat{s}(v,t)\in\mathbb{R}$：

$$\hat{s}(v,t)=\text{Sigmoid}\left(\mathrm{h}^L(v)\cdot\left(\mathrm{h}^L(t)\right)^T\right). \tag{31}$$

训练期间，我们使用二元交叉熵（BCE）损失L作为标签损失：

$$\mathcal{L}_{\text{tag}}=\text{BCE}\left(y(v,t),\hat{s}(v,t)\right), \tag{32}$$

其中，$y(v,t)$ 表示视频 v是否具有标签 t的真实标注。

我们使用梯度反转层（GRL）[7] 实现公式（25）中的极小极大博弈，以进行端到端训练。最终损失 L 由分类损失和对抗训练损失组成：

$$\mathcal{L}=\mathcal{L}_{\text{tag}}+\sum_{l=1}^{L}\mathcal{L}_{\text{adv}}^l. \tag{33}$$

在推理过程中，新视频没有出向关系，因为它没有标签（没有标签_关系），并且只有先前上传的视频才能影响新上传的视频（不存在由被_关注_的出向关系）。因此，可以预先计算旧节点的节点表示。这样，当新视频加入视频-标签网络时，我们只需计算它的入向被_关注_的消息 $\mathrm{m}_{r_3}^l(t)$，即可获得该视频的表示。因此，我们的模型能够实现归纳推理。

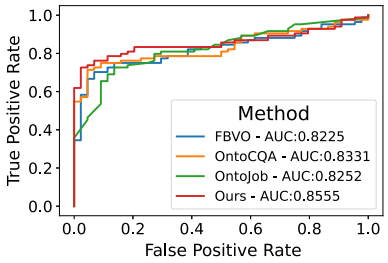


图 5：子主题发现的 ROC 曲线 5 实验

5.1 数据集

5.1.1 数据收集与准备。我们从快手平台的三个垂直领域（服饰、化妆品和食物）获取了大规模、真实世界的微视频数据集。我们从2021年7月至8月的精选页面中随机选取了500,000个微视频。它们的标签来自多种基于人工的信号（例如搜索观看），以及仅利用视频文本描述的在线视频打标签系统。视频创作者之间的关注关系于8月14日收集，我们还过滤掉了机器用户及其视频。我们排除了过短（< 1 秒）或过长（> 60 秒）的微视频，移除了非视觉标签，并筛除了标签少于2个的视频。

为量化数据集质量，我们人工评估了随机抽取的 1,000 个微视频，标签准确率和召回率分别为 91.3% 和 66.2%。我们的数据集分为 3 个部分：训练集（80%）、验证集（10%）和测试集（10%）。

5.1.2 标签本体构建。我们根据第 4.1 节提出的方法构建了标签本体。在 768,798 个共现标签对中，我们标注了 1,000 个。根据子主题发现的精确率和召回率定义两个阈值δ_r，ϵ_r。具体而言，我们选择了精确率为 90% 时的 δ_r，以及召回率为 90 % 时的 ϵ_r。如图 5 所示，我们评估了子主题发现的构建性能，结果表明，我们的方法明显优于基于规则的方法 FBVO [6]、无监督方法 OntoJob [29] 和 OntoCQA [26]。我们构建的本体覆盖了 96.67% 的标签，只有 2.94% 的标签是通过 ϵ_r松弛覆盖的。

5.1.3 数据集统计。我们在表 1 中展示了数据集的统计信息。它们各不相同的统计特征使后续实验产生了不同结果。服装数据集的视觉信息最丰富（每个标签对应的视频数量最多）；食品数据集的社交网络和标签本体密度最低（每个视频的关注者数量最少，且每个标签的子主题数量位居倒数第二）。

5.2 实验设置

5.2.1 实验细节。标签词嵌入取自 BERT 的最后一个隐藏层。视频帧嵌入由以 SwinTransformer 为骨干网络的 TSN 提取。受 GPU 内存限制，我们按照 Deep Graph Library [30] 的做法，在每个图神经网络层中为每种关系采样 4 个邻居。层数 L 设为 2。批量大小设为 1024，并使用学习率为 0.0005 的 AdamW 优化器。特征丢弃和边丢弃率均设为 0.2。为稳定对抗训练，我们遵循 Ganin 和 Lempitsky[7] 的方法，将 λ 从

表2：与基线方法的性能比较。

模型	服装数据集					化妆品数据集					食品数据集				
	mAP	P@1	P@3	R@5	R@10	mAP	P@1	P@3	R@5	R@10	mAP	P@1	P@3	R@5	R@10
基础方法	59.89	89.22	74.97	57.12	64.68	46.33	74.30	45.49	47.23	56.06	45.98	60.13	47.78	50.57	59.60
ML-GCN	61.49	91.24	76.75	57.97	65.53	48.03	77.20	46.42	48.15	57.02	47.54	63.94	49.65	51.75	59.97
NeXtVLAD	63.14	91.90	77.44	58.86	66.98	49.78	77.93	47.93	49.58	58.92	48.49	64.92	50.51	52.43	60.93
CMA	63.04	91.51	77.29	58.85	66.99	50.24	78.60	48.29	50.04	59.41	49.28	65.54	50.91	53.09	61.65
MALL-CNN	63.31	91.67	77.95	59.07	67.20	50.46	78.93	48.25	50.15	59.83	49.58	66.00	50.89	53.32	61.99
TagGNN	63.44	93.15	78.23	59.13	66.82	52.39	84.80	50.32	51.64	60.37	50.92	74.49	53.58	53.09	60.05
HAN	61.83	92.17	77.06	58.00	65.76	51.72	84.87	49.54	51.31	59.99	50.76	75.32	53.53	53.02	59.79
HGT	63.85	92.35	78.25	59.43	67.58	54.50	86.62	51.78	53.61	62.62	51.56	76.22	54.38	53.53	60.04
SimpleHGN	63.16	91.80	77.99	59.09	67.10	53.68	86.28	51.17	52.80	61.62	52.76	77.27	55.21	54.19	61.33
Ours	**65.05**	**93.74**	**79.07**	**59.89**	**68.22**	**56.55**	**88.16**	**53.18**	**54.98**	**64.55**	**56.93**	**82.16**	**58.94**	**57.11**	**63.99**

0 按照以下计划调整为 λ_0： $\lambda = \lambda_0 (2/(1 + exp(-\gamma p)) - 1)$，其中 p 在训练过程中从 0 线性变化到 1。所有数据集的 γ 和 λ_0 分别设为 20 和 0.0005。所有基线方法均采用相同的设置。

5.2.2 评估指标。我们采用了多标签分类中三种常见的评估指标：平均精度均值（mAP）、P@K 精确率和 R@K 召回率。P@K 和 R@K 衡量前 K 个标签的预测性能，而 mAP 衡量所有标签的整体排序结果。

5.3 基线方法

为了验证我们的 RADAR 模型，我们将其与以下基线方法进行了比较：

• 视频标签预测方法：ML-GCN [3], NeXtVLAD [15, 36], CMA [38], 和 MALL-CNN [14]。我们将它们与我们的方法进行比较，以验证引入额外社交网络和标签图信息的有效性。需要注意的是，NeXtVLAD、CMA 和 MALL-CNN 以序列级特征 $\mathbf{h} \in \mathbb{R}^{F \times d}$ 作为输入，并使用额外的聚合模块，而非我们方法中的语义池化。

• 异质图神经网络方法：HAN [31],TagGNN [21], HGT[8], 和 Simple-HGN [20] 是异质图神经网络领域的四种代表性方法，其中 Simple-HGN 是最先进方法。我们将 RADAR 与它们进行比较，以验证 RADAR 模型的有效性。

5.4 性能比较

与基线方法相比，我们的结果汇总于表2。通过分析该表，我们得出了以下观察结果：

• 异质图神经网络方法的表现通常优于视频标签基线方法。两者的主要区别在于，基线方法更好地利用了视频帧特征，而图神经网络方法则利用了标签本体和额外的社交网络。性能提升表明，对于微视频，融入社交影响和标签关系比对帧特征进行建模更有效。

• 我们提出的模型 RADAR 始终优于所有异质图神经网络方法。与表现最佳的基线方法相比，RADAR 在服饰、化妆品和食品数据集上的平均精度均值分别相对提升了 1.9%、3.8% 和 7.9%。这一提升证明了 RADAR 在视频标签任务中的优越性。

• Simple-HGN 是唯一一个在不同数据集上相关排名有所差异的异构图神经网络基线方法。值得注意的是，它的表现在食物数据集中最佳，而在服装数据集的所有异构图神经网络基线方法中则位列倒数第二。巧合的是，这一相关排名与各数据集的社交网络和标签本体密度相反。*is_followed_by* 和 *is_subtopic_of* 关系越密集，Simple-HGN 的表现就越差。考虑到 Simple-HGN 是通用异构图神经网络领域的最先进方法，我们将这一现象归因于视频标签与以往任务有很大不同这一事实。在视频标签任务中，来自帧的视频特征与来自语言模型的标签特征是存在较大语义差距的两种模态。这一现象也证明了我们的对抗聚合网络的必要性，该网络能够聚合视觉和语言信息。

5.5 消融实验

本节开展了多项实验，以进一步分析我们模型的有效性，结果如表3所示。具体而言，我们首先探究了视频-标签网络中不同关系的作用，随后分析了各个组件的有效性，包括门控图变换器（GGT）和对抗聚合网络（AAN）。

5.5.1 不同关系的消融实验。我们将模型与以下变体进行了比较：1）**去除后续关系**，移除*is_followed_by*关系；2）**去除子主题关系**，移除*is_subtopic_of* 关系；3）**同时去除子主题关系和后续关系**，移除这两种关系。

在三个数据集中，**去除后续关系**的结果始终低于**去除子主题关系**的结果，这表明社交网络信息的贡献大于标签本体。

请注意，如果从完整模型中移除*is_followed_by*，性能下降幅度比从**去除子主题关系**的模型中移除*is_followed_by* 更大。换个角度看，当*is_subtopic_of* 关系存在时，加入*is_followed_by*关系带来的提升更大。因此，这两种关系的组合并非两者效果的简单相加。这表明两种关系包含互补信息，因此，有必要在 AAN 模块中分离共有信息和特有信息。

5.5.2 GGT 的消融实验。我们将 GGT 中的门控残差网络替换为残差网络，构成变体**不含门控残差**；并将分离注意力替换为互注意力，构成变体**采用互注意力**。与完整模型相比，**不含门控残差** 和**采用互注意力**的性能

表3：关于两种关系、门控图 Transformer 和对抗聚合网络有效性的消融研究。

模型	服装数据集					化妆品数据集					食品数据集				
	平均精度均值	P@1	P@3	R@5	R@10	平均精度均值	P@1	P@3	R@5	R@10	平均精度均值	P@1	P@3	R@5	R@10
不含后续关系	62.28	91.49	76.97	58.49	66.34	48.68	77.71	47.11	48.72	58.00	47.37	63.89	49.87	51.78	60.26
不含子主题	63.60	93.56	78.43	59.17	66.96	53.17	86.23	50.93	52.43	60.98	53.27	79.39	56.37	54.67	60.74
不含子主题和后续关系	62.17	91.45	76.97	58.39	66.18	48.33	77.35	46.88	48.59	57.57	46.48	63.33	48.82	51.38	59.35
不含门控残差	64.20	93.32	78.54	59.52	67.66	55.20	86.77	52.06	54.06	63.27	55.65	80.90	57.74	56.18	62.94
采用互注意力	64.12	93.34	78.55	59.44	67.61	55.32	87.35	52.14	54.10	63.35	55.69	81.00	57.66	56.34	63.24
采用拼接	63.82	93.34	78.50	59.24	67.16	53.48	86.57	51.20	52.58	61.32	52.64	78.25	55.67	54.20	60.50
采用注意力机制	64.76	93.48	78.97	59.78	67.89	55.27	88.01	52.30	53.89	62.86	55.47	81.64	58.13	56.11	62.45
采用 LMF	64.36	93.25	78.98	59.63	67.72	55.00	87.51	52.25	53.67	62.42	54.85	81.24	57.64	55.69	61.71
不含 L_{adv}	63.87	93.21	78.60	59.38	67.28	54.31	86.57	51.83	53.47	62.26	56.57	82.08	58.78	56.88	63.44
采用双向边	65.26	93.94	79.13	60.09	68.35	56.86	88.06	53.34	55.12	64.65	57.19	82.27	58.98	57.35	64.32
Ours	65.05	93.74	79.07	59.89	68.22	56.55	88.16	53.18	54.98	64.55	56.93	82.16	58.94	57.11	63.99

在所有数据集上都持续下降。这表明，在对行为传播进行建模时，过滤无关信息非常重要，而分离注意力是在模态差异下的一种有效设计。

5.5.3 AAN 消融实验。为了验证我们提出的 AAN 的效果，我们通过引入四种变体进行了一系列实验：1）**使用拼接**，用向量拼接和线性变换层取代 AAN，这是最简单的多模态聚合方法；2）**使用注意力**，在 [31] 中用注意力模块取代 AAN；3）**使用 LMF**，用低秩多模态融合（LMF） [18], 取代 AAN，这是效果最好的多模态向量聚合方法；4）**不使用** L_{adv}，移除对抗损失 L_{adv}。

从表3可以看出，我们的方法始终优于所有多模态向量聚合方法。性能提升并非源于参数数量增加，因为**使用 LMF**的参数数量最多。与**不使用** L_{adv}相比，我们认为性能提升归功于视觉和语言消息中共性信息与独特信息的分离。

5.5.4 双向边。我们还探究了双向边的影响。具体而言，在变体 **含双向边**中，我们向网络添加了 是_上位主题_的和 拥有_视频关系。由于视频节点有两种类型的入站消息，我们也对其应用了 AAN。请注意，我们没有添加 关注关系，因为我们希望新视频在归纳学习时不会向原始图发送消息。尽管此变体略微提升了性能，但考虑到额外的计算负担，我们没有采用这种设置。

5.6 敏感性分析

我们研究了不同超参数设置对性能的影响。我们探究了两个主要参数：图神经网络层的数量 L 以及独特损失的最大权重 λ_0。

如图6所示，由于过平滑现象在图神经网络中很常见，当模型包含两层图神经网络层时，性能最佳。当 λ_0 约为 0.0005 至 0.001 时，模型性能最佳，这表明独特损失与共有损失之间达到了适当的平衡。请注意，$\lambda_0 = 0$ 与**w/o** L_{adv}不同。尽管两者都不包含共有信息，但在后一种变体中，判别器 D^l 被移除了，因此它也不包含独特信息。我们观察到，**w/o** L_{adv} 会导致

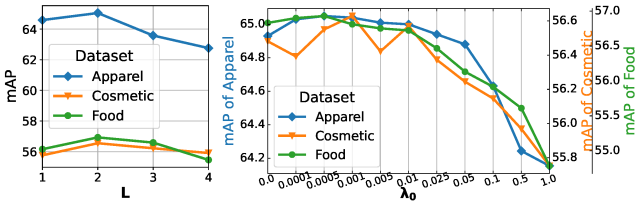


图 6：敏感性分析

与设置 $\lambda_0 = 0$相比，性能下降幅度更大。这表明独特信息更为重要。

6 结论与未来工作

在本研究中，我们将社交影响和标签关系整合到用户-视频-标签网络中，以提升微视频标签效果。为构建标签关系，我们以半监督方式，基于标签统计和图约束建立了标签本体。为获得视频-标签网络中更好的节点表示，我们提出了 RADAR，这是一种由门控图 Transformer 和对抗聚合网络组成的异构图神经网络。在三个数据集上开展的大量实验结果证明了 RADAR 的优越性及其两个组件的有效性。

未来，我们将推广该方法，以处理更复杂的标签知识结构。对于行为传播建模，我们计划对用户社交影响的时间动态进行建模。

7 致谢

本研究得到国家自然科学基金资助，项目编号：62176137、U1936203 和 62006140；山东省自然科学基金资助，项目编号：ZR2020QF106；北京智源人工智能研究院（BAAI）；快手科技。

参考文献

[1] Damon Centola. 2010. 在线社交网络实验中的行为传播。《科学》329 (2010), 1194–1197. [2] Xu Chen, Changying Du, Xiuqiang He, and Jun Wang. 2020. JIT2R：物品标签标注与基于标签的推荐联合框架。载于《国际ACM SIGIR 信息检索研究与发展会议》论文集。ACM, 1681–1684. [3] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. 基于图卷积网络的多标签图像识别。载于《IEEE计算机视觉与模式识别会议》论文集。IEEE, 5177–5186.[4] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. 2020. 大规模整体视频理解。载于《欧洲计算机视觉会议》论文集。Springer, 593–610.[5] Yujie Fan, Mingxuan Ju, Shifu Hou, Yanfang Ye, Wenqiang Wan, Kui Wang, Yinming Mei, and Qi Xiong. 2021. 异构时序图Transformer：用于安卓恶意软件检测的智能系统。载于《ACM SIGKDD知识发现与数据挖掘会议》论文集。ACM, 2831–2839.[6] Quan Fang, Changsheng Xu, Jitao Sang, M. Shamim Hossain, and Ahmed Ghoneim. 2016. 基于大众分类法的视觉本体构建及其应用。《IEEE多媒体汇刊》18 (2016), 702–713. [7] Yaroslav Ganin and Victor S. Lempitsky. 2015. 通过反向传播实现无监督领域自适应。载于《国际机器学习会议》论文集。JMLR.org, 1180–1189. [8] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. 异构图Transformer。载于《国际万维网会议》论文集。ACM, 2704–2710. [9] Di Jin, Zhongang Qi, Yingmin Luo, and Ying Shan. 2021. TransFusion：通过基于转换的知识嵌入进行视频标签推断的多模态融合。载于《ACM国际多媒体会议》论文集。ACM, 1093–1101.[10] Denis Kotkov, Alexandr Maslov, and Mats Neovius. 2021. 重新审视标签相关性预测问题。载于《国际 ACM SIGIR 信息检索研究与发展会议》论文集。ACM, 1768–1772.[11] Mengmeng Li, Tian Gan, Meng Liu, Zhiyong Cheng, Jianhua Yin, and Liqiang Nie. 2019. 基于图卷积网络的微视频长尾话题标签推荐。载于《ACM国际信息与知识管理会议》论文集。ACM, 509–518. [12] Xirong Li, Shuai Liao, Weiyu Lan, Xiaoyong Du, and Gang Yang. 2015. 通过层次语义嵌入实现零样本图像标签标注。载于《国际 ACM SIGIR 信息检索研究与发展会议》论文集。ACM, 879–882. [13] Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, and Bo Zheng. 2020. 用于点击率预测的对抗式多模态表示学习。载于《国际万维网会议》论文集。ACM, 827–836. [14] Xuewei Li, Hongjun Wu, Mengzhu Li, and Hongzhe Liu. 2022. 通过耦合注意力多实例学习与标签关系图实现多标签视频分类。《模式识别快报》156 (2022), 53–59. [15] Rongcheng Lin, Jing Xiao, and Jianping Fan. 2018. NeXtVLAD：一种用于聚合帧级特征的大规模视频分类高效神经网络。载于《欧洲计算机视觉会议研讨会》论文集。Springer, 206–218.[16] Bang Liu, Weidong Guo, Di Niu, Chaoyue Wang, Shunnan Xu, Jinghong Lin, Kunfeng Lai, and Yu Xu. 2019. 腾讯面向查询与文档理解的以用户为中心的概念挖掘系统。载于《ACM SIGKDD国际知识发现与数据挖掘会议》论文集。ACM, 1831–1841.[17] Can Liu, Li Sun, Xiang Ao, Jinghua Feng, Qing He, and Hao Yang. 2021. 用于欺诈交易检测的意图感知异构图注意力网络。载于《ACM SIGKDD知识发现与数据挖掘会议》论文集。ACM, 3280–3288.[18] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. 基于模态特定因子的高效低秩多模态融合。载于《计算语言学协会年会》论文集。ACL, 2247–2256. [19] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. 2020. 基于共享-特有特征迁移的跨模态行人重识别。载于《IEEE计算机视觉与模式识别会议》论文集。IEEE, 13376–13386.[20] Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. 2021. 我们真的取得了很大进展吗？重新审视、基准评测并改进异构图神经网络。载于《ACM SIGKDD知识发现与数据挖掘会议》论文集。ACM, 1150–1160. [21] Kelong Mao, Xi Xiao, Jieming Zhu, Biao Lu, Ruiming Tang, and Xiuqiang He. 2020. 面向信息检索的物品标签标注：一种三部图神经网络方法。在国际 *ACM SIGIR* 信息检索研究与发展会议论文集中。ACM, 2327–2336。[22] Liqiang Nie、Yongqi Li、Fuli Feng、Xuemeng Song、Meng Wang 和 Yinglong Wang。2020年。通过联合问题-主题嵌入学习进行大规模问题标签标注。*ACM*信息系统汇刊 38 (2020)，20:1–20:23。[23] Jingbo Shang、Jialu Liu、Meng Jiang、Xiang Ren、Clare R. Voss 和 Jiawei Han。2018年。从海量文本语料库中自动挖掘短语。*IEEE*知识与数据工程汇刊 30 (2018)，1825–1837。[24] Si Sun、Zhenghao Liu、Chenyan Xiong、Zhiyuan Liu 和 Jie Bao。2021年。在开放域关键短语抽取中捕捉全局信息量。在自然语言处理与中文计算论文集中。Springer，275–287。[25] Teng Sun、Chun Wang、Xuemeng Song、Fuli Feng 和 Liqiang Nie。2022年。联合建模个性化语言风格和情感的响应生成。*ACM*多媒体计算、通信与应用汇刊 *(TOMM)* 18 (2022)，1–20。[26] K. Suryamukhi、P. D. Vivekananda 和 Manish Singh。2021年。在社区问答网站中挖掘标签关系。在概念建模——*ER*国际会议中。Springer，345–355。[27] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。注意力机制就是一切。在神经信息处理系统会议论文集中。MIT Press，5998–6008。[28] Sunny Verma、Chen Wang、Liming Zhu 和 Wei Liu。2019年。DeepCU：整合多模态情感分析中的共性与独特潜在信息。在国际人工智能联合会议论文集中。ijcai.org，3627–3634。[29] Jarno Vrolijk、Stefan T. Mol、Christian Weber、Mohammadreza Tavakoli、Gabor Kismihok 和 Mauro Pelucchi。2022年。OntoJob：从劳动力市场数据中自动学习本体。在语义计算国际会议 *(ICSC)*中。IEEE，195–200。[30] Minjie Wang、Da Zheng、Zihao Ye、Quan Gan、Mufei Li、Xiang Song、Jinjing Zhou、Chao Ma、Lingfan Yu、Yu Gai、Tianjun Xiao、Tong He、George Karypis、Jinyang Li 和 Zheng Zhang。2019年。Deep Graph Library：一个以图为中心、高性能的图神经网络软件包。*arXiv*预印本 *arXiv:1909.01315* (2019)。[31] Xiao Wang、Houye Ji、Chuan Shi、Bai Wang、Yanfang Ye、Peng Cui 和 Philip S. Yu。2019年。异构图注意力网络。在国际万维网会议论文集中。ACM，2022–2032。[32] Ya Wang、Dongliang He、Fu Li、Xiang Long、Zhichao Zhou、Jinwen Ma 和 Shilei Wen。2020年。通过标签图叠加进行多标签分类。在*AAAI*人工智能会议论文集中。AAAI Press，12265–12272。[33] Yinwei Wei、Zhiyong Cheng、Xuzheng Yu、Zhou Zhao、Lei Zhu 和 Liqiang Nie。2019年。面向微视频的个性化话题标签推荐。在*ACM*国际多媒体会议论文集中。ACM，1446–1454。[34] Yinwei Wei、Xiang Wang、Weili Guan、Liqiang Nie、Zhouchen Lin 和 Bao-quan Chen。2019年。面向微视频理解的神经多模态协同学习。*IEEE*图像处理汇刊 29 (2019)，1–14。[35] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He 和 Tat-Seng Chua。2020年。结合隐式反馈的多媒体推荐图精炼卷积网络。在第*28*届*ACM*国际多媒体会议论文集中。ACM，3541–3549。[36] Hao Wu、Jiajie Wang、Yuanzhe Gu、Peisen Zhao 和 Zhonglin Zu。2021年。多模态广告视频标签挑战赛解决方案。在*ACM*国际多媒体会议论文集中。ACM，4808–4812。[37] Wenmain Yang、Kun Wang、Na Ruan、Wenyuan Gao、Weijia Jia、Wei Zhao、Nan Liu 和 Yunyong Zhang。2019年。使用语义关联图进行时间同步的视频标签抽取。*ACM*数据知识发现汇刊 13 (2019)，37:1–37:24。[38] Renchun You、Zhiyao Guo、Lei Cui、Xiang Long、Yingze Bao 和 Shilei Wen。2020年。用于多标签分类的语义图嵌入跨模态注意力。在*AAAI*人工智能会议论文集中。AAAI Press，12709–12716。[39] Le Zhang、Ding Zhou、Hengshu Zhu、Tong Xu、Rui Zha、Enhong Chen 和 Hui Xiong。2021年。用于职业流动预测的注意力异构图嵌入。在*ACM SIGKDD*知识发现与数据挖掘会议论文集中。ACM，2192–2201。[40] Shengming Zhang、Hao Zhong、Zixuan Yuan 和 Hui Xiong。2021年。用于预测高潜力早期初创企业的可扩展异构图神经网络。在*ACM SIGKDD*知识发现与数据挖掘会议论文集中。ACM，2202–2211。