论文调研

2025 年 6 月 21 日

1 Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model

该论文发表在计算语言学协会年会，论文的核心贡献在于提出了一种新颖的多模态层次融合模型（Hierarchical Fusion Model），旨在提升 Twitter 平台上讽刺检测的准确性。这一模型巧妙地整合了文本、图像以及图像属性三种模态的数据，以更全面地捕捉和理解推文中潜在的讽刺意图。并且论文发表了一个新的数据集，用于多模态 Twitter 讽刺检测

1.1 论文中处理各个模态的模型

论文提取图像特征的方法：

1. 预训练和微调的 ResNet 模型：使用预训练的 ResNet 模型来提取图像的特征，论文中提出应更换模型最后一层全连接层以适应新任务。
2. 区域特征提取：将输入图像调整为 448 × 448 大小，并分成 14 × 14 个区域。每个区域通过 ResNet 模型提取特征，得到 14 × 14 个区域特征向量，这些向量被称为原始图像向量。
3. 图像引导向量：将所有区域特征向量进行平均池化，得到一个图像引导向量，该向量用于指导后续的特征融合。公式如下：

$$\mathbf{v}_{image} = \frac{\sum_{i=1}^{N_r} \mathbf{v}_{region_i}}{N_r} \tag{1}$$

N_r 指图像块的数量。

图像属性模态处理

1. 属性预测：使用另一个预训练和微调的 ResNet 模型来预测每张图像的 5 个属性。这些属性是图像的高级语义概念。

2. 属性嵌入：将预测的属性词通过 GloVe 词向量进行嵌入，得到每个属性的嵌入向量，这些向量被称为原始属性向量。

3. 属性引导向量：通过一个两层神经网络计算每个属性的注意力权重，然后对属性嵌入向量进行加权平均，得到属性引导向量。公式如下:

$$\alpha_i = \mathbf{W}_2 \cdot \tanh(\mathbf{W}_1 \cdot \mathbf{e}^{(a_i)} + \mathbf{b}_1) + \mathbf{b}_2 \tag{2}$$

$$\alpha = softmax(\alpha) \tag{3}$$

$$\mathbf{v}_{attr} = \sum_{i=1}^{N_a} \alpha_i \mathbf{e}^{(a_i)} \tag{4}$$

其中 $\mathrm{e}^{(a_i)}$ 为原始属性向量， v_{attr} 为最终的属性指导向量

文本模态处理

1. 使用双向长短期记忆网络（Bi-LSTM）来提取文本的特征。

2. Bi-LSTM 的每个时间步的前向和后向隐藏状态进行拼接，得到原始文本向量。

3. 文本引导向量：将所有时间步的原始文本向量进行平均，得到文本引导向量

1.2 模态融合

1. Early Fusion: 将图像属性模态生成的属性指导向量 v_{attr} 通过非线性变换后，作为 Bi-LSTM 的初始状态。

$$[\mathbf{h}_{f0}; \mathbf{h}_{b0}; \mathbf{c}_{f0}; \mathbf{c}_{b0}] = ReLU(\mathbf{W} \cdot \mathbf{v}_{attr} + \mathbf{b}) \tag{5}$$

2. Representation Fusion: 利用图像、文本和属性三种模态的信息，更准确地建模它们之间的关系。通过结合低级的原始向量和高级的指导向

量，重新构造每种模态的特征向量，使其能够更好地反映模态间的相互作用。

$$\alpha_{(i)}^{mn} = W_{mn}^{2} \cdot \tanh\left(W_{mn}^{1} \cdot \left[\mathbf{X}_{m}^{(i)}; \mathbf{v}_{n}\right] + \mathbf{b}_{mn}^{1}\right) + \mathbf{b}_{mn}^{2} \tag{6}$$

该公式计算模态 m 的第 i 个原始向量 $\mathrm{X}_{m}^{(i)}$ 在模态 n 的指导向量 v_{n} 指导下的引导权重。公式将两个向量拼接使模型能够同时考虑当前模态的特征和其他模态的信息，两次线性变换将其映射到一个新的空间，曲正切函数引入了非线性。

$$\alpha^{mn} = softmax(\alpha^{mn}) \tag{7}$$

这个公式对计算得到的引导权重 a^{mn} 进行归一化处理。

$$\alpha_{(i)}^{m} = \frac{\sum_{n \in \{text,image,attr\}} \alpha_{(i)}^{mn}}{3} \tag{8}$$

这个公式通过将来自所有其他模态的归一化引导权重进行平均，得到模态 m 的第 i 个原始向量的最终重建权重。

$$\mathbf{v}_{m} = \sum_{i=1}^{L_m} \alpha_{(i)}^{m} \mathbf{X}_{m}^{(i)} \tag{9}$$

这个公式利用最终的重建权重对模态 m 的所有原始向量进行加权求和，生成模态 m 的重建特征向量。

3. Modality Fusion：将来自不同模态（文本、图像和属性）的特征向量融合在一起，形成一个固定长度的融合向量。

$$\mathbf{v}'_{m} = \tanh(\mathbf{W}_{m3} \cdot \mathbf{v}_{m} + \mathbf{b}_{m3}) \tag{10}$$

通过公式将上一步各个模态的输出结果 v_{m} 通过线性层转化成长度一样的向量。

$$\tilde{\alpha}_{m} = \mathbf{W}_{m2} \cdot \tanh(\mathbf{W}_{m1} \cdot \mathbf{v}_{m} + \mathbf{b}_{m1}) + \mathbf{b}_{m2} \tag{11}$$

$$\tilde{\alpha} = softmax(\tilde{\alpha}) \tag{12}$$

计算每个模态的注意力权重

$$\mathbf{v}_{fused} = \sum_{m \in \{text,image,attr\}} \tilde{\alpha}_{m} \mathbf{v}'_{m} \tag{13}$$

求的最终融合向量，最终输入两层全连接神经网络作为分类层得出预测结果，损失函数为交叉熵。

2 Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection

论文发表于 Findings of the Association for Computational Linguistics: EMNLP 2020，提出了一种新颖的基于 BERT 架构的模型，该模型能够同时考虑文本内部和跨模态（文本与图像之间）的不一致性，解决了现有模型未充分考虑讽刺表达中不一致性特征的问题。具体方法是设计了跨模态注意力机制，通过文本-图像匹配层捕捉文本和图像之间的不一致性。引入了文本内部的不一致性分析：利用标签（hashtags）与文本之间的矛盾来捕捉文本内部的不一致性。

2.1 论文中处理各个模态的模型

1. 文本模态的处理：在论文中作者将文本模态分为正文和标签，二者都以词序列的形式输入模型:

$$X = \{x_1, x_2, \cdots, x_N\} \tag{14}$$

每个词 x_i 被表示为一个 d 维向量。论文中作者选择基础型 bert 处理文本模态。

2. 图像模态处理：给定一张图像 I，首先将其大小调整为 224×224 像素。这是 ResNet 模型的标准输入尺寸。
之后使用 ResNet-152 模型提取图像特征，去掉模型最后一层全连接层，取最后一层卷积层的输出：

$$ResNet(I) = \{r_i \mid r_i \in R^{2048}, i = 1, 2, \cdots, 49\} \tag{15}$$

最后用一个线性变化把图像特征投影到和文本特征相同的维度：

$$\mathbf{G} = \mathbf{W}_v \cdot ResNet(I) \tag{16}$$

2.2 inter-modality Attention

从自注意力机制中得到灵感，其核心思想是通过计算序列中每个 token 对之间的相关性，生成一个加权表示，从而捕捉序列中的长距离依赖关系。

作者发现跨模态不一致信息可以表现为多模态特征之间的互动。自注意力机制的一般公式如下：

$$Output = softmax\left(\frac{(XW_Q)(XW_K)^{\mathrm{T}}}{\sqrt{d_k}}\right)(XW_V) \tag{17}$$

公式中的 X 代表着输入序列，分别经过三个线性变化后映射到查询、键和值空间：

1. 查询（Query）：表示关注的当前元素，用于与其他元素进行比较。
2. 键（Key）：表示其他元素的特征，用于与查询进行比较
3. 值（Value）：表示其他元素的值，用于生成最终的加权表示

由公式可以看出，当前关注的元素通过与键求点积后对值求加权和可以得到其对其他所有元素的相关性。
在论文中，文本特征 H 被用作查询（Query），而图像特征 G 被用作键（Key）和值（Value），即可求得文本特征对图像特征的相关性：

$$ATT_i(H,G) = softmax\left(\frac{[W_i^Q H]^{\mathrm{T}}[W_i^K G]}{\sqrt{d_k}}\right)[W_i^V G]^{\mathrm{T}} \tag{18}$$

将 h 个注意力头的输出进行拼接，然后通过一个线性变换得到最终的跨模态注意力输出：

$$MATT(H,G) = [ATT_1(H,G), \cdots, ATT_h(H,G)]W_o \tag{19}$$

得到注意力输出后进行两次残差融合和归一化，实现了信息的融合，确保了原始输入信息能够直接传递到下一层。层归一化 LN 对融合后的特征进行归一化处理，稳定了训练过程，防止梯度消失或爆炸。

$$Z = LN(H + MATT(H,G)) \tag{20}$$

$$TIM(H,G) = LN(Z + MLP(Z)) \tag{21}$$

以上过程还要重复几次，论文中实验发现重复三次后不一致性差异最大。基于 BERT 的模型中，输入序列的开始位置会添加一个特殊标记 [CLS]，它在序列中占据第一个位置。TIM3(H, G) 是经过 3 次文本-图像匹配层处理后的输出，包含了文本和图像之间的交互信息。从这个输出中取 [CLS] token 对应的编码结果，就可以得到 HG（图像特征和文本特征的相关性）。

2.3 Intra-modality Attention

论文提出利用亲和矩阵用于建模文本特征和标签特征之间的交互关系，公式如下：

$$C = \tanh(H^{\mathrm{T}} W b^{\mathrm{T}}) \tag{22}$$

公式通过建模文本特征和标签特征之间的交互关系，来捕捉文本内部的不一致性。

计算出亲和矩阵 C 后，通过按列进行最大池化（max-pooling）操作来生成权重向量 a：

$$a = column - wisemax - pooling(C) \tag{23}$$

最终的文本内部不一致性表示 HT 通过将权重向量 a 与标签特征矩阵 T 相乘得到：

$$H^{\mathrm{T}} = a^{\mathrm{T}} \tag{24}$$

2.4 预测

将文本内部不一致性表示 $\mathrm{H^T}$ 和跨模态不一致性表示 $\mathrm{H^G}$ 拼接在一起，拼接后的特征向量通过一个线性层（全连接层）来降低维度，将其映射到与类别数量相同的维度，线性层的输出通过 Softmax 函数转换为概率分布。公式如下：

$$\hat{y} = Softmax(W[H_G : H_T] + b) \tag{25}$$

3 Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement

论文发表在 Conference on Empirical Methods in Natural Language Processing 2022。现有的大多数技术仅关注文本输入与其配图之间原子层面的不一致，而忽略了更复杂的组合形式，论文提出同时探索文本和图像模态的原子级和组合级语义一致性，更全面地捕捉文本和图像之间的复杂关系，提升讽刺检测的准确性。并且通过利用预训练模型（如 CLIP 和 GPT-2）生成的图像标题等外部知识，为模型提供额外的语义信息，进一步提升性能。

3.1 论文中处理各个模态的模型

1. 文本特征提取：使用 BERT 模型为每个词生成特征向量，然后通过一个多层感知机（MLP）进一步处理这些特征向量，MLP 通过多层的线性变换和非线性激活函数，将 BERT 提取的高维词特征向量映射到一个低维空间。最终结果为：

$$T = [t_1, t_2, \cdots, t_n] \tag{26}$$

2. 图像特征提取：将图像大小调整为 224×224，将图像分割成 r 个 patches（小块），并将这些 patches 重塑成一个序列，将这些 patches 输入图像编码器以获取每个 patch 的视觉特征。最后也要通过一个 MLP 层，最终结果如下：

$$I = [i_1, i_2, \cdots, i_r] \tag{27}$$

3. 外部知识特征提取：将图像的大小调整为 224 × 224，利用预训练的 CLIP 模型对图像进行编码，提取图像的语义特征。将经过映射后的图像特征作为初始输入，馈送到微调后的 GPT-2 模型中。GPT-2 模型根据图像特征生成描述图像的文本标题。该标题包含了图像的语义信息，可以作为图像的外部知识。因为生成的标题为文本形式，所以对其再进行图像特征提取的步骤。

3.2 Atomic-Level Congruity Modeling

论文使用多头交叉注意力机制来对齐两种不同模态的方法，公式如下：

$$head_i = softmax\left(\frac{(TW_i^q)^\top (IW_i^k)}{\sqrt{d/h}}\right)(IW_i^v) \tag{28}$$

该公式的意义是通过将文本作为查询，模型能够根据文本中的每个词去图像中寻找相关的部分。作者也试过用图像做查询，结果效果没有文本好。
将多个注意力头的输出进行拼接，并通过多层感知机（MLP）和残差连接更新文本表示：

$$\tilde{T} = norm(T + MLP([head_1 \| head_2 \| \cdots \| head_h])) \tag{29}$$

和前面的论文提到的一样，上述步骤要多进行几次，论文里认为三次效果最好。最后计算文本和图像之间的最终原子级一致性分数：

$$Q_a = \frac{1}{\sqrt{d}} \tilde{T} I^\top \tag{30}$$

$$s_a = softmax(\tilde{T}W_a + b_a)^\top Q_a \tag{31}$$

内积计算可以衡量两个向量之间的相似性，两个模态的表示矩阵通过内积可以得到文本中每个词与图像中每个 patch 的相似度。之后通过线性层计算每个词的重要性分数，再用 softmax 函数对重要性分数归一化，对相似性矩阵进行加权求和，得到最终的原子级一致性分数。

3.3 Composition-Level Congruity Modeling

论文通过构建文本图和视觉图并使用图注意力网络（GAT）来捕捉文本和图像的复杂结构关系，学习组合级特征，并计算组合级一致性分数，以衡量文本和图像在更深层次上的匹配程度。

1. 构建文本图和视觉图：对于文本图，作者通过 spacy 工具提取两个词之间存在依赖关系，如果词之间存在依赖关系则二者之间有一条边。对于视觉图，图像的 patch 视为图的节点，根据几何邻接关系连接相邻的节点。

2. 使用图注意力网络对文本和视觉模态中的图进行建模，以文本模态为例：

$$\alpha_{i,j}^l = \frac{\exp(LeakyReLU(v^{l\top}[\Theta^l t_i^l \| \Theta^l t_j^l]))}{\sum_{k\in N(i)\cup\{i\}} \exp(LeakyReLU(v^{l\top}[\Theta^l t_i^l \| \Theta^l t_k^l]))} \tag{32}$$

该公式计算第 l 层两个节点之间的注意力系数，即第 l 层节点 j 对节点 i 的影响程度。通过特征变换与拼接、注意力分数计算以及归一化，衡量节点间的相关性并确定信息传播的权重。

$$t_i^{l+1} = \sum_{j\in N(i)\cup\{i\}} \alpha_{i,j}^l \Theta^l t_j^l \tag{33}$$

该公式通过注意力系数加权邻居节点特征并求和，同时保留自身特征，来更新目标节点的特征，从而形成包含上下文信息的组合级特征表示。经过特定的层数后得到最终文本模态的组合级嵌入

$$\hat{T} = [t_1^{L\mathrm{T}}, t_2^{L\mathrm{T}}, \ldots, t_r^{L\mathrm{T}}] \tag{34}$$

视觉模态同理，经过特定层数后：

$$\hat{I} = [i_1^{L_\mathrm{I}}, i_2^{L_\mathrm{I}}, \ldots, i_n^{L_\mathrm{I}}] \tag{35}$$

论文中指出两层性能最优。
当文本图不可靠时，通过拼接组合级特征和句子嵌入 c 来增强文本特征，其中 c 是 $\tilde{T}$ 中词嵌入的加权和：

$$c = softmax(TW_c + b_c)^\top \tilde{T} \tag{36}$$

3. 生成一致性分数：首先用点积计算组合级文本特征和组合级图像特征的相似度矩阵：

$$Q_p = \sqrt{\frac{1}{d}}([\hat{T} \parallel c]\hat{I}^\top) \tag{37}$$

之后计算组合级一致性分数：

$$s_p = softmax([\hat{T} \parallel c]W_p + b_p)^\top Q_p \tag{38}$$

3.4 Knowledge Enhancement

也是计算一个原子级一致分数和组合级一致分数，只是把视觉模态换成了标题模态，最终得到 s_k^p 和 s_k^a

3.5 预测

作者先每个图像 patch 对于讽刺检测的重要性：

$$p_v = softmax(IW_v + b_v) \tag{39}$$

之后计算标题模态每个词对于讽刺的重要性：

$$p_k = softmax(kW_k + b_k) \tag{40}$$

最后将原子级和组合级一致性分数与图像 patch 的重要性结合，并通过 softmax 函数生成最终的预测结果，同时支持整合外部知识以增强模型性能：

$$y' = softmax(W_y^k[p_v \odot s_a \parallel p_v \odot s_p \parallel p^k \odot s_k^a \parallel p^k \odot s_k^p] + b_k^v) \tag{41}$$

4 Modeling inter-modal incongruous sentiment expressions for multi-modal sarcasm detection

论文发表在 Neurocomputing，提出了一种用于多模态反讽检测的深度跨模态映射图卷积网络（DCMG）新方法，主要贡献包括：设计有效的跨模态映射网络，通过两两映射文本和图像特征向量并引入共享掩码参数，充分捕捉模态间特征的共性和差异；利用外部的形容词-名词对（ANPS）知识构建跨相关图，结合图卷积网络（GCN）和基于检索的注意力机制，精准提取多模态中的关键反讽线索，有效提升反讽检测的准确性，多个实验验证了该方法相较于现有技术的优势。

4.1 如何处理各个模态

1. 文本模态：使用 Roberta-BASE 模型进行文本特征提取。
2. 图像模态：使用 CLIP 模型中的 ViT-B-32 模型进行图像特征提取。
3. 外部形容词名词对：使用 SentiBank 工具包来提取 ANPs。SentiBank 是一个用于提取图像中语义信息的工具包，能够识别图像中的物体和场景，并生成描述这些物体和场景的形容词-名词对。该工具可以根据 anps 与图像的相关性进行排序，论文从中选择前五个 anps。

4.2 Cross-modal mapping

论文使用特征映射、权重调整和归一化处理的方法，实现文本和图像特征的有效融合，为多模态反讽检测提供更全面、准确的特征表示。特征映射是该论文中用于多模态融合的关键步骤，通过将不同模态的特征向量映射到一个共享的空间，捕捉它们之间的关联和差异。（其实就是调整线性变换的权重矩阵使得各个模态映射到同一空间）具体过程如下：

1. 首先用两层线性变换将一个模态的特征向量映射到另一个模态的空间：

$$V_{t,i}^{(k)} = \omega_{t,i2} \cdot \tanh(\omega_{t,i1} \cdot x_t^{(k)} + b_{t,i1}) + b_{t,i2} \tag{42}$$

   $\mathrm{x}_t^{(k)}$ 是从文本或图像中提取的第 k 个原始特征向量，通过与训练过的权重矩阵相乘即可映射到其他模态的空间。论文中解释这些权重矩阵

封装了将模态 t 的特征转换为模态 i 特征所需的知识。

2. 特征归一化：

$$V_{t,i} = softmax(V_{t,i}^{(k)} W_{t,i}, \lambda^{\top}) \tag{43}$$

简单的归一化，但是作者在这里加入了共享掩码参数 ，通过在归一时加入一个相同的数字可以对映射后的特征权重进行加权，以强调或抑制某些特征，从而捕捉模态之间的共享信息。

3. 特征融合：

$$V_f n = W_s \sum_{k=1}^{K} \frac{V_{t,i}^{(k)}}{2} \tag{44}$$

对经过映射和归一化处理后的特征向量进行加权求和，以得到最终的融合特征。

4.3 Cross-correlation graph

提出了一种融合文本与图像特征的跨相关图构建方法，通过提取图像的形容词-名词对（ANPs）语义信息并结合文本特征，利用情感线索准确捕捉多模态情感关系，从而提升反讽检测性能。为了方便分析，将每个 ANP 中的形容词和名词分别表示为 ai 和 ni。

1. 首先用余弦定理计算文本词和 ANPS 的名词部分的相似度：

$$\cos - sim(t_i, n_j) = \frac{\sum_{n_{i,j}=1}(t_i \times n_j)}{\sum_{n_i=1} h_t \times \sum_{n_j=1} n_j} \tag{45}$$

分母部分计算的是文本模态模长和 ANPS 名词模长的乘积。而且这里计算的仅仅是两者之间的相似度。

2. 之后计算文本词与 ANPS 之间的情感差异：

$$\xi_{i,j} = \gamma^{-\omega(t_i)\omega(a_j)\times|\omega(t_i)-\omega(a_j)|} \tag{46}$$

在这里作者用到了 SenticNet 情感词典，里面将所有的词的情感从-1 到 1 进行排序，若词典中没有则情感默认为 0。公式中可以看出当情感一致时差异值较小，当情感不一致时差异值较大。

3. 计算整体的关联权重：

$$A_{i,j} = \cos - sim(t_i, n_j) \times \xi_{i,j} + 1 \tag{47}$$

+1 用于强调跨模态节点的聚合，增强跨模态信息的融合。

这样就求得文本模态与图像模态的跨模态相关图。

4.4 Multi-modal fusion

作者通过将 GCN 与改进的注意力机制相结合，提升了反讽检测的性能，步骤如下：

1. 首先利用之前求到的跨模态相关图捕捉文本和图像特征之间的情感关系，把这些关系赋给原始的模态向量：

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{48}$$

第一层 $\mathrm{H}^{(l)}$ 为模态原始向量， $\tilde{D}$ 为跨模态图的度矩阵。

2. 计算节点 j 和节点 i 之间的相关性：

$$s(h_j, g_i) = v^{\top} \tanh(Wh_j + Ug_i) \tag{49}$$

其中 $\mathrm{G} = \mathrm{Concat}(\mathrm{V}_{text}, V_{fn}, V_{image}) = [g_1, g_2, \ldots, g_n]$ 为前面提到的最终融合特征与文本特征和图像特征矩阵的拼接， h_j 为公式 48 的最后一层输出。

之后归一化相似度可以求得节点 j 对节点 i 的注意力权重：

$$a_{i,j} = softmax(s(h_j, g_i)) = \frac{\exp(s(h_j, g_i))}{\sum_{n=1}^{N}\exp(s(h_n, g_i))} \tag{50}$$

3. 用求得的权重对 G 中向量进行加权运算，得到一个标量值表示相似度：

$$f = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{i,j} \cdot g_j \tag{51}$$

4.5 预测

全连接层和 softmax 函数处理最终的反讽表示 f, 得到反讽选择空间中的概率分布：

$$\hat{y} = softmax(W_o f + b_o) \tag{52}$$

5 Multi-Modal Sarcasm Detection via Graph Convolutional Network and Dynamic Network

该论文发表在论文发表在 “CIKM ’24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management”，论文提出结合图卷积网络（GCN）和动态网络用于多模态反讽检测，同时引入模拟人类思维的外部知识增强方法生成增强文本，实验验证了该方法相较于当前先进方法在多模态反讽检测任务中的有效性。

5.1 如何处理各个模态

1. 文本特征提取：使用 COMET 挖掘出文本中隐含的社会事件影响和人类情感信息， $\text{effect}_i = \{\tilde{w}_{1i}, \tilde{w}_{2i}, \ldots, \tilde{w}_{Mi}\}$ 为时候影响， $\text{react}_i = \{w_{i1}, w_{i2}, \ldots, w_{iM}\}$ 为人类情感因素，这些补充的额外知识都是以词序列的形式输出的，将它们与原始的文本模态拼接：$\text{A\_Text}_i = Text_i \oplus effect_i \oplus react_i$，再交给 Robert 模型处理：

$$T = [t_1, t_2, \ldots, t_m] = RoBERTa(AText) \tag{53}$$

2. 物体级别特征提取：使用 Faster R-CNN 检测图像中的物体并提取其视觉特征 vi、位置特征 pi、物体类别 ci 和物体属性 ai。对于位置和视觉特征论文选择了线性变换求和：

$$f_i = W_v v_i + W_p p_i + b_f \tag{54}$$

对于物体类别和物体属性选择 Robert 模型处理：

$$\tilde{c}_i = RoBERTa(c_i) \tag{55}$$

$$\tilde{a}_i = RoBERTa(a_i) \tag{56}$$

最后拼接：

$$V_o = [[f_1, \tilde{c}_1, \tilde{a}_1]^\top, [f_2, \tilde{c}_2, \tilde{a}_2]^\top, \ldots, [f_k, \tilde{c}_k, \tilde{a}_k]^\top] \tag{57}$$

3. 图像全局级别特征提取：用 vit 模型处理图像：

$$V = [v_1, v_2, \ldots, v_r] \tag{58}$$

5.2 Local Graph Convolutional Network

论文构建文本图、图像图和跨模态图，并利用图卷积网络（GCN）来捕捉多模态数据中的语义关系和不一致特征，从而有效识别反讽表达。

1. 文本模态图：使用 spacy 工具捕捉到词语之间的依赖关系，有依赖则建立一条边：

$$t_{i,j} = \{1\ , if D(t_i, t_j) = 1, i, j \in [1, m]0, otherwise \tag{59}$$

2. 图像模态图：同一物体内部节点之间建立一条权重为 1 的边，当两个节点都为类别节点时，边的权重为两个模块重叠的面积与总面积比值。

$$A_v[i,j] = \{1\ ,\ \ i \bmod 3 = j \bmod 3,\ \ i, j \in [1, 3k] S_{i,j},\ \ i \bmod 3 = 1\ j \bmod 3 = 1,\ \ i, j \in [1, 3k]0, \tag{60}$$

3. 跨模态图: 节点涵盖所有文本词和图像物体特征: $\mathrm{v}_i^c\ (i \in [1, m + 3k])$，但是用到了知识图谱 ConceptNet，如果文本词和图像物体之间存在关系，则边的权重为 1。

在构建三个图后，论文使用多层图卷积网络（GCN）架构学习模态间和模态内的不一致表达，每一层 GCN 通过构建文本模态层、图像模态层和跨模态层来捕捉不同模态之间的交互和内部特征：

$$G_t^l = ReLU(\tilde{A}_t G_c^{l-1} W_t^l + b_t^l) \tag{61}$$

$$G_v^l = ReLU(\tilde{A}_v G_t^l W_v^l + b_v^l) \tag{62}$$

$$G_c^l = ReLU(\tilde{A}_c G_v^l W_c^l + b_c^l) \tag{63}$$

其中 G_x^l 为第 l 层协作 GCN 处理后对应图的节点表示，x 为文本，图像和跨模态图。 G_t^0 为初始输入文本表示与图像物体级特征的组合，即：$\mathrm{G}_c^1 = \{t_1, t_2, \ldots, t_m, f_1, \tilde{c}_1, \tilde{a}_1, \ldots, f_k, \tilde{c}_k, \tilde{a}_k\}$。公式中的 $\tilde{A}_x$ 是之前求到的三个模特图邻接矩阵的归一化。公式中可以看出，作者使用跨模态节点信息更新文本节点，用文本节点信息更新图像节点，最后使用图像节点信息更新跨模态节点，最终实现多模态信息的融合。

经过计算，得到了协作 GCN 层的最终输出 $\mathrm{G}_c^L = \{g_1, g_2, \ldots, g_{m+3k}\}$，以及原始跨模态节点信息： $\mathrm{H} = \{\mathrm{v}_c^1, v_c^2, \ldots, v_c^{m+3k}\}$。论文提出可以用融合了其

他模态信息的最终输出计算原始跨模态节点信息之间的相关性，并得到注意力分数：

$$\alpha_i = softmax\left(\sum_{j=1}^{m+3k} v_j^T g_j\right) \tag{64}$$

用求得的注意力分数对初始节点做加权和：

$$f_G = \sum_{i=1}^{m+3k} \alpha_i h_i \tag{65}$$

最后用一般方法求得一个概率分布：

$$\mathbf{p}_G = softmax(\mathbf{W}_G \mathbf{f}_G + \mathbf{b}_G) \tag{66}$$

5.3 Global Dynamic Network (GDN)

论文中提到的一个很新的概念，实际上就是多头注意力机制，只不过论文中的多头注意力机制每进行两次循环会将结果输入到一个 MLP 中进行处理。最终还得到一个概率分布 P_D。

5.4 最终预测

结合前面得到的两个概率分布计算最终预测：

$$\hat{Y} = argmax(\alpha \times \mathbf{p}_G + (1-\alpha) \times \mathbf{p}_D) \tag{67}$$

论文中没有对参数 α 进行解释。

5.4.1 三级标题

这是三级标题下的正文内容。