**自然语言处理大作业——多模态谣言核查系统**

**社交网络多模态虚假信息核查系统的研究与实现——最终报告**

**1. 引言**

**1.1 项目背景与研究意义**

在当今高度互联的社交网络时代，图文并茂的多模态信息已成为主流传播形式。然而，这也为虚假信息（谣言）的快速、广泛散播提供了温床，对社会信任、公共安全乃至个人认知均构成严重威胁。传统的谣言核查依赖领域专家，耗时费力，难以应对海量信息的冲击。因此，研究并实现一个能够自动、高效、精准地对多模态信息进行真实性核查的系统，不仅是技术上的挑战，更具有极其重要的现实意义与社会价值。

**1.2 研究目标与主要任务**

本项目旨在设计并实现一个先进的多模态虚假信息核查系统，能够深度理解图文内容，并结合外部证据进行综合研判。

核心研究目标包括：

* **多模态特征提取**：利用前沿的深度学习模型，从查询信息（图像与文本）及其证据材料中，精准提取深层次的视觉与语义特征。
* **核心算法设计**：设计并实现一种创新的特征交互与融合机制，有效建模查询信息与海量证据之间的复杂关联，特别是同构模态间的关联性。
* **系统实现与优化**：基于PaddlePaddle框架，完成整个系统的代码实现、调试、训练与评估，并探索优化策略以提升系统效率与性能。
* **性能验证**：在标准数据集上对模型进行充分训练和评估，验证所提出方案的有效性，并达成高精度的分类目标（本项目中使用宏F1分数作为核心指标）。
* **Web在线评测平台**：构建一个Web端在线评测的平台，通过爬取网络上的证据信息实现对待检测谣言信息的实时检测

**1.3 主要工作与贡献**

本项目成功完成了从系统设计、代码实现到实验验证的全流程工作，主要贡献体现在：

1. **实现了一个高效的数据预处理管道**：创新性地设计并实现了**图像特征缓存机制**，将耗时的图像特征提取过程从训练循环中剥离，变为一次性的预处理步骤，极大地提升了实验迭代的效率。
2. **构建了基于双重注意力机制的核心模型**：提出并实现了一个**双重注意力（Dual-Attention）模型**，能够分别在文本和图像两个模态内部，动态地衡量查询信息与各证据信息的相关性，从而聚焦于最关键的证据进行判断。
3. **完成了系统的端到端实现与验证**：基于PaddlePaddle框架，完整实现了数据处理、模型训练、评估和预测的全套流程，并在数据集上取得了 **F1分数达 0.9372** 的优异性能，验证了技术方案的可行性与先进性。

**2. 系统设计与技术方案**

**2.1 数据集分析与预处理**

该系统使用的数据集结构清晰，旨在通过“查询-证据”对比的方式进行事实核查。数据集被划分为训练、验证和测试三部分。

* **数据集规模**：
  + **训练集**：11,184 个样本
  + **验证集**：1,309 个样本
  + **测试集**：1,129 个样本
* **样本结构**：每个样本都由一个**查询 (Query)** 和一组相关的**证据 (Evidence)** 构成。
  + **查询**：由一张图片（位于img文件夹）及其配对的文本描述组成。
  + **证据**：证据从两个维度进行搜集，以提供全面的参考信息。
    1. **正向文本检索**：源于 img\_html\_news 文件夹。这是通过查询文本在网络上搜索到的证据，其信息（如网页标题、摘要、图片路径）被记录在 direct\_annotation.json 文件中。
    2. **反向图像检索**：源于 inverse\_search 文件夹。这是通过查询图片进行“以图搜图”找到的证据，其信息（如匹配的网页标题）被记录在 inverse\_annotation.json 文件中。

**2.2 系统总体架构**

本系统设计遵循清晰的分层架构，确保了流程的模块化与高效性。整体架构如下图所示，包含输入层、特征提取层、特征交互与融合层及分类输出层。

* **输入层**：接收一个待核查的样本，包含“查询图片”和“查询文本”，以及通过正向（文本搜）和反向（以图搜图）检索获得的多份“证据图片”和“证据文本”。
* **特征提取层**：采用预训练模型对多模态数据进行编码。
  + **图像特征**：使用 **ResNet-50** 深度卷积网络提取所有图像的2048维视觉特征向量。
  + **文本特征**：使用 **ERNIE-M** 多语言模型提取所有文本的768维语义特征向量。
* **特征交互与融合层**：这是系统的核心创新所在，通过双重注意力机制建模查询与证据的关联。
* **分类输出层**：将最终融合的特征向量输入分类器，输出**“真实”、“虚假”或“无法核实”**的三分类预测结果。

**2.3 核心模型算法：双重注意力机制**

为精准判断查询信息与海量证据的相关性，我们设计并实现了NetWork核心模型模块，其关键在于**双重注意力机制 (Dual Attention Mechanism)**。

1. **文本注意力模块** (self.attention\_text):
   * **输入**: 查询文本的特征向量 (作为Query)，以及所有证据文本的特征向量 (作为Key和Value)。
   * **功能**: 该模块能够计算查询文本与每一条证据文本之间的语义相似度，并据此分配不同的注意力权重。这使得模型能够自动聚焦于内容上最相关的几条文本证据，忽略无关或干扰信息，最终生成一个高度浓缩且相关的文本证据表示。
2. **图像注意力模块** (self.attention\_image):
   * **输入**: 查询图片的特征向量 (作为Query)，以及所有证据图片的特征向量 (作为Key和Value)。
   * **功能**: 与文本模态类似，该模块在视觉层面上进行关联性判断，使模型能够重点关注与查询图片在场景、物体、风格上最相似的证据图片，并生成一个加权融合后的视觉证据表示。
3. **最终融合与分类**:
   * 模型将四个关键部分的特征向量进行拼接（Concatenate）：**查询文本特征, 注意力加权后的文本证据特征, 查询图像特征, 注意力加权后的图像证据特征**。
   * 这个强大的、信息丰富的融合向量，最终被送入一个由两个全连接层构成的分类器中，进行最终的分类判决。

**3. 项目迭代与优化历程**

本项目的开发遵循了“先构建、后优化”的迭代思路。我们首先搭建了一个功能完整的原型系统，以验证核心算法的可行性；随后，针对原型暴露出的性能瓶颈和评估方法上的不足，进行了针对性的、系统性的优化，最终形成了当前的高效版本。

**3.1 第一阶段：原型系统构建 (基于 main.py)**

在项目初期，我们的核心目标是快速搭建一个可运行的基线（Baseline）系统，验证我们所设计的“查询-证据”对比以及双重注意力机制的有效性。

* **数据处理方案**:
  + 我们设计了 NewsContextDatasetEmbs 类来处理数据。
  + 此方案采用实时处理（On-the-fly）的模式：在每次训练迭代需要数据时，程序都会从硬盘读取原始图片，然后进行缩放、裁剪等一系列变换操作。
* **核心模型与算法**:
  + 我们构建了初版的 NetWork 模型，其内部**集成了ResNet图像编码器** (self.resnet)。
  + 这意味着，在模型每次进行前向传播时，都需要将批次中的所有原始证据图片**实时送入ResNet模块进行特征提取**。
  + 尽管如此，模型的核心思想——即使用ERNIE-M处理文本，并设计**双重注意力机制**来融合图文证据——在这一阶段就已确定并实现。
* **训练与评估策略**:
  + 训练方面，我们采用了标准的AdamW优化器和带预热的学习率衰减策略。
  + 在模型评估和选择上，我们以**准确率**作为衡量模型性能和保存最佳版本的主要依据。
* **遇到的核心瓶颈**:
  + 原型系统虽然功能完整，但很快暴露出了致命的**性能瓶颈**：**训练效率极其低下**，在训练的过程中，即使是使用的云端服务器的NVIDIA4090显卡也需要至少7-8h才能跑完一个epoch，这让我们意识到我们处理数据集的方式存在问题。
  + **问题根源在于**：在项目初期构建的原型系统中，我们很快发现其训练效率极其低下，其根源在于一个复合型的**数据供给流程瓶颈**。这个瓶颈并非由单一因素造成，而是由磁盘I/O、CPU计算和冗余操作共同导致的。

首先，最主要的瓶颈是**磁盘I/O速度与GPU计算速度的严重不匹配**。GPU完成一个批次（batch）的复杂数学运算可能仅需几毫秒，但CPU从硬盘中读取下一个批次的原始图片文件到内存却需要耗费长得多的时间。这导致了“GPU饥饿”现象：强大的GPU在完成任务后，大部分时间处于空闲状态，等待着CPU和硬盘缓慢地“喂给”它下一批数据，使得整体训练时间被大大拉长。

其次，**实时的图像变换加剧了CPU的负担**，构成了次级瓶颈。对每张图片进行缩放、裁剪和归一化等操作都需要消耗CPU资源。当CPU忙于处理这些计算时，它就无法高效地调度数据读取任务，这进一步延迟了数据进入GPU的时间，使得数据供给管道更加拥堵。

最后，从算法流程设计的角度看，最大的问题在于**大量的重复计算**。对于同一张图片，其通过固定的ResNet模型提取出的特征向量是恒定不变的。然而，在原型系统中，这一本应只计算一次的特征，却在每一个训练周期（Epoch）中都被反复地、冗余地计算，这造成了宝贵的GPU算力的巨大浪费。

**3.2 第二阶段：性能优化与模型精调 (基于 main2.ipynb)**

为了解决第一阶段遇到的效率瓶颈，并进一步提升模型的科学性，我们进行了大刀阔斧的改进，形成了最终的优化版本。

* **关键改进一：引入图像特征缓存机制**
  + **动机**：将一次性计算与重复性计算分离，彻底解决特征提取的效率瓶颈。
  + **实现**：我们重写了数据处理模块，设计了全新的 FeatureCachedNewsContextDataset 类。该模块的核心是一个**预处理（Pre-processing）**流程：在正式训练开始前，程序会完整地遍历一次数据集，利用ResNet-50将所有图片（查询图片和证据图片）的特征一次性计算完毕，并以高效的 .pkl 格式保存到磁盘。
  + **成果**：在后续的所有训练过程中，程序不再接触原始图片，而是直接从硬盘高速加载这些预先计算好的特征向量。这一改动，使得数据加载速度得到了**数量级的提升**，让系统能够进行快速的实验和迭代。
* **关键改进二：模型结构轻量化与职责分离**
  + **动机**：配合特征缓存机制，简化模型结构，使其更专注于核心任务。
  + **实现**：我们将沉重的 **ResNet 模块从 NetWork 模型的定义中彻底移除**。优化后的 NetWork 模型不再负责图像特征提取，它的输入直接就是预先计算好的特征向量。
  + **成果**：模型结构变得更加轻量、清晰，其核心职责被聚焦于“多模态特征的融合与分类”，符合“高内聚、低耦合”的设计原则。
* **关键改进三：评估指标的科学化**
  + **动机**：采用比准确率更全面、更鲁棒的指标来评估和选择模型。
  + **实现**：我们将模型选择的“金标准”从**准确率**切换为了**宏F1分数（Macro F1-Score）**。F1分数能够同时兼顾模型的精确率和召回率，在处理可能存在类别不平衡的数据时，是比准确率更科学、更可靠的评估工具。
  + **成果**：这确保了我们最终保存和用于测试的模型，是在综合性能上表现最优的模型，而不仅仅是“猜对”的样本多。

**3.3 优化成果总结**

|  |  |  |  |
| --- | --- | --- | --- |
| **对比维度** | **第一阶段：原型系统 (main.py)** | **第二阶段：优化系统 (main2.ipynb)** | **优化成效** |
| **数据处理** | 实时读取、实时变换、实时提取特征 | **一次性预处理，特征缓存** | ✅ **训练效率数量级提升** |
| **模型结构** | NetWork 内含ResNet，结构臃肿 | NetWork 移除ResNet，结构轻量 | ✅ **职责清晰，聚焦核心任务** |
| **评估指标** | **准确率 (Accuracy)** | **宏F1分数 (Macro F1-Score)** | ✅ **模型选择更科学、更鲁棒** |

通过这一系列的迭代与优化，我们不仅解决的了项目初期的性能瓶颈，更重要的是建立了一套高效、科学的实验流程，为最终取得优异的模型性能奠定了坚实的基础。

**4. 实验与结果分析**

**4.1 实验环境与设置**

* **硬件平台**: 本地端 NVIDIA 4070 Laptop GPU，云端租借 NVIDIA 4090 GPU。
* **软件框架**:
  + 编程语言: Python 9
  + 深度学习框架: PaddlePaddle, PaddleNLP 10
  + 核心模型: ERNIE-M, ResNet-50 11
  + 数据处理: NumPy, Pandas, PIL 12

**4.2 评估指标**

考虑到谣言分类任务中可能存在的类别不平衡问题，本项目最终选择**宏F1分数 (Macro F1-Score)** 作为模型性能的核心评估指标，它能更公允地评价模型在所有类别上的综合表现。同时，也监控准确率（Accuracy）作为参考。

**4.3 实验结果与分析**

通过在优化后的代码 (main2.ipynb) 上进行10个周期的训练，系统表现出优异的学习能力和性能。

* **最终性能**: 在验证集上，模型最终达到的**最佳F1分数为 0.9372**，对应的准确率为 0.92513，证明了本系统设计与算法的有效性。
* **训练过程分析**:
  + 从训练日志可以看出，模型的F1分数随着训练的进行稳步提升，从第一个周期结束时的 0.7535 上升至第九个周期结束时的峰值 0.9372。
  + 这清晰地表明，模型通过双重注意力机制，成功地学会了如何从复杂的证据中提取并利用有效信息来辨别谣言。
  + 系统成功实现了基于验证集F1分数的最佳模型保存机制，确保了最终用于测试的模型是泛化能力最强的版本。

**5. 系统功能实现与在线评测平台**

为验证本研究提出模型的有效性，并提供一个直观、可交互的应用范例，我们基于Gradio框架，设计并实现了一个功能完备的“多模态谣言核查在线评测平台”。该平台不仅封装了核心的谣言检测算法，更通过双模式设计，兼顾了算法的全面性验证与系统的实用性及容错能力。

**5.1 评测平台概述**

评测平台的核心设计思想是提供两种逻辑独立但功能互补的验证模式，以应对不同的使用场景：

* **模式A (多模态综合验证)**: 旨在最完整、最严谨地复现并验证核心算法在处理多源、多模态证据链时的综合性能。此模式要求用户提供一个完备的数据案例包，用于对模型进行全面的压力测试。
* **模式B (实时/手动文本验证)**: 旨在模拟真实世界应用中的两种常见场景：一是需要系统自动从互联网上搜集证据进行快速响应；二是在网络受限或需要使用特定证据集时，系统应具备手动处理能力。此模式突出了系统的端到端自动化能力与在复杂环境下的鲁棒性。

**5.2 模式A：多模态证据综合验证**

此模式是本研究核心算法——双重注意力多模态融合模型的旗舰演示。

* **功能定位**：作为对模型理论性能的最终验证，此模式用于处理包含**文本检索**与**图片反向检索**双路证据的复杂案例。
* **工作流程与实现**：
  1. **数据输入**: 用户需通过界面上传一个完整的案例，包括：一份**查询图片**、相应的**查询文本**、由文本检索生成的direct\_annotation.json文件、由图片反向检索生成的inverse\_annotation.json文件，以及direct\_annotation.json中引用的所有本地证据图片文件。
  2. **后端处理**: 后端run\_mode\_a函数接收到全部文件后，首先并行解析两个JSON文件，提取出所有的文本证据（网页标题、摘要等）和direct\_annotation.json中定义的图片证据路径。随后，它将图片路径与用户上传的图片文件进行匹配，加载图像数据。最终，将整合后的查询图文、多源文本证据列表、多张证据图片送入run\_prediction核心推理函数。
  3. **结果呈现**: 系统界面会将所有解析出的证据在一个结构化的DataFrame中进行展示，并明确标注每条证据的**来源**（“文本检索”或“图片反向检索”），同时展示证据图片（若有）和证据文本，为用户提供了清晰的数据溯源。模型的最终分类结果（如“✅ 真实”或“❌ 虚假”或“❓ 不确定”）会在界面的醒目位置单独显示。
* **评测意义**: 此模式的成功运行，证明了我们所构建的模型具备处理和理解异构、多源、多模态信息的复杂能力，是整个系统技术先进性的直接体现。

**5.3 模式B：实时/手动文本证据验证**

此模式的设计更侧重于实用性和灵活性，通过一个单选框提供两种子模式。

* **子模式一：实时在线检索**
  + **功能定位**：模拟系统在真实应用中，从零开始自动完成“**证据搜集 -> 模型推理**”的端到端流程。
  + **工作流程与实现**：用户输入查询图文后，run\_mode\_b\_live函数被触发。它首先调用集成的collect\_evidence\_for\_inference模块，该模块以查询文本为关键词，通过**搜狗搜索引擎API**实时抓取网络上相关的网页信息，并从中提取标题、摘要等作为纯文本证据，最后将这些实时证据送入模型进行判断。此外，该模式还集成了**保存功能**，允许用户将某次成功的实时检索结果（包括生成的JSON文件和查询图片）永久保存到本地，以备后续复现和分析。
* **子模式二：手动输入文本证据**
  + **功能定位**：作为系统的**容错和扩展机制**。
  + **工作流程与实现**：当在线检索功能因网络问题不可用，或研究人员希望使用特定的、非网络来源的文本证据进行验证时，可切换至此模式。用户在文本框中输入多条文本证据（每行一条），run\_mode\_b\_manual函数会直接解析这些文本，并连同查询图文一起送入模型进行推理。该设计**解耦了数据获取与模型推理两个环节**，确保了即便在无网络或数据获取模块失效的情况下，核心的验证功能依然可用，极大地提升了系统的健壮性。

通过以上双模式、三通道的精心设计，本在线评测平台不仅是一个功能强大的模型演示工具，更是一个结构合理、具备高度实用价值和鲁棒性的系统原型。

**6. 结论与展望**

**6.1 工作总结**

本项目成功研究并实现了一个高效、精准的社交网络多模态虚假信息核查系统。通过采用**ERNIE-M**和**ResNet-50**作为强大的特征提取器，并创新性地设计了**双重注意力机制**对图文证据进行加权融合，本系统能够深度挖掘和利用多模态数据间的内在联系。实验结果表明，该系统在目标数据集上取得了**高达0.9372的F1分数**，充分验证了所提出技术方案的先进性和有效性。项目过程中对数据处理流程的**特征缓存优化**，不仅是工程上的巨大改进，也为后续的科研实验提供了宝贵经验。

**6.2 未来展望**

基于本项目的坚实基础，未来的研究可向更深、更广的维度拓展：

* **模型深度优化**: 探索如跨模态注意力（Cross-Modal Attention）等更复杂的融合策略，让图文信息在更早的阶段进行交互。同时，可尝试更先进的骨干网络（如ViT系列模型）。
* **错误案例分析**: 对模型预测失败的案例进行深度剖析，归纳模型的弱点（例如，对特定类型谣言不敏感），并针对性地进行模型改进。
* **模型可解释性研究**: 引入可视化技术（如注意力热力图），直观展示模型在决策时究竟“看”了哪些文本关键词和图像区域，以增强系统的透明度和可信度。
* **系统应用与部署**: 在模型性能稳定后，可将其封装为API服务，探索其在社交媒体内容审核、新闻事实核查等真实场景下的应用潜力，为构建更清朗的网络空间贡献技术力量。