基于大规模预训练 CLIP 模型微调的图文检索方法

摘 要

随着互联网的迅速发展，海量数据的涌现使得从大量信息中筛选出有价值内容变得日益重要。信息检索技术因此成为一项关键技术，尤其是在面对真实世界中复杂的多模态数据时。传统的信息检索模型往往只能处理单一模态数据，而现实情况通常更为复杂，涉及文本、图像等多种数据类型。为了解决这一挑战，本文提出了一种基于自然语言处理（NLP）、计算机视觉（CV）技术，以及多模态模型 **CLIP（Contrastive Language-Image Pre-training）**的信息检索模型。该模型特别针对中文数据集进行了优化，通过大量数据的预训练，并在特定比赛数据集上进行微调，显著提升了检索精度，超越了部分传统中文模型的性能。

本文首先对图像与文本数据进行了建模，通过预处理将异构的图像与文本信息转化为结构化的张量形式，使得基于深度学习的模型能更好地对其处理。接下来，我们对问题一的**图像检索文本**（Image-to-Text，I2T）任务与问题二的**文本检索图像**（Text-to-Image，T2I）任务进行了**统一建模**，将两大检索问题转化为**图像与文本的相似度衡量问题**，按照相似度对检索目标进行排序，以具有最高相似度的目标作为检索结果。为了度量图像文本多模态数据的相关程度，我们引入多模态模型 CLIP 的结构，其基于**对比学习（Contrastive Learning）**方法，训练出能够将图像和文本映射到同一嵌入空间的**图像编码器（Image Encoder）**和**文本编码器（Text Encoder）**。这一映射过程便于计算不同模态数据之间的余弦相似度，从而实现高效的图文互检索功能。

具体而言，为了兼顾模型精度与效率，我们实现的 CLIP 模型采用 **ViT-L/14** 模型作为图像编码器，以及 **RoBERTa** 模型作为文本编码器。同时，这些编码器均经过充分预训练，显著加快了后续训练的收敛速度。接下来，我们收集了一系列高质量的公开数据集，构建了一个包含大约 2 亿图像文本对的中文多模态**预训练数据集**，并基于该数据集对所实现的 CLIP 模型进行预训练，得到一个**泛化性能强大的基线模型**。基于对问题的统一建模，我们的基线模型在问题一的 I2T 任务中达到了 **77.77%** 的 R@5 精度，在问题二的 T2I 任务中达到了 **78.88%** 的 R@5 精度。

为了进一步挖掘该模型的潜力，我们针对比赛数据集（容量仅为 5 万），通过**图像裁剪**、**文本翻译**等方式进行数据增强，得到一个容量为 40 万的**增强数据集**。进而，我们使用增强数据集对预训练的 CLIP 基线模型进行微调。经过仅仅 10 个回合的微调，我们的模型便能在问题一的 I2T 任务中达到了 **88.88%** 的 R@5 精度，在问题二的 T2I 任务中达到了 **89.99%** 的 R@5 精度。该结果相较基线模型有着显著地提升，验证了微调策略与数据增强方法的有效性。

综上，本文对图像检索文本任务与文本检索图像任务进行了统一建模，基于 CLIP 框架实现了一个兼顾性能与效率的多模态模型，并整合高质量公开数据集进行预训练得到泛化性能强大的基线模型。进而，我们对比赛数据集进行增强，并对基线模型微调，显著提升了特定任务下的性能表现，对中文信息检索领域具有重要的理论和实践意义。

关键词： 多模态特征融合　图文检索　预训练—微调　对比学习　深度学习

目录

一、 问题描述与假设

1. 问题背景

随着近年来智能终端设备和多媒体社交网络平台的飞速发展，多媒体数据呈现海量增长的趋势，使当今主流的社交网络平台充斥着海量的文本、图像等多模态媒体数据，也使得人们对不同模态数据之间互相检索的需求不断增加。有效的信息检索和分析可以大大提高平台多模态数据的利用率及用户的使用体验，而不同模态间存在显著的语义鸿沟，大大制约了海量多模态数据的分析及有效信息挖掘。因此，在海量的数据中实现跨模态信息的精准检索就成为当今学术界面临的重要挑战。图像和文本作为信息传递过程中常见的两大模态，它们之间的交互检索不仅能有效打破视觉和语言之间的语义鸿沟和分布壁垒，还能促进许多应用的发展，如跨模态检索、图像标注、视觉问答等。

图像文本检索指的是输入某一模态的数据（例如图像），通过训练的模型自动检索出与之最相关的另一模态数据（例如文本），它包括两个方向的检索，即基于文本的图像检索和基于图像的文本检索，如图 1 所示。基于文本的图像检索的目的是从数据库中找到与输入句子相匹配的图像作为输出结果；基于图像的文本检索根据输入图片，模型从数据库中自动检索出能够准确描述图片内容的文字。然而，来自图像和来自文本的特征存在固有的数据分布的差异，也被称为模态间的“异构鸿沟”，使得度量图像和文本之间的语义相关性困难重重。

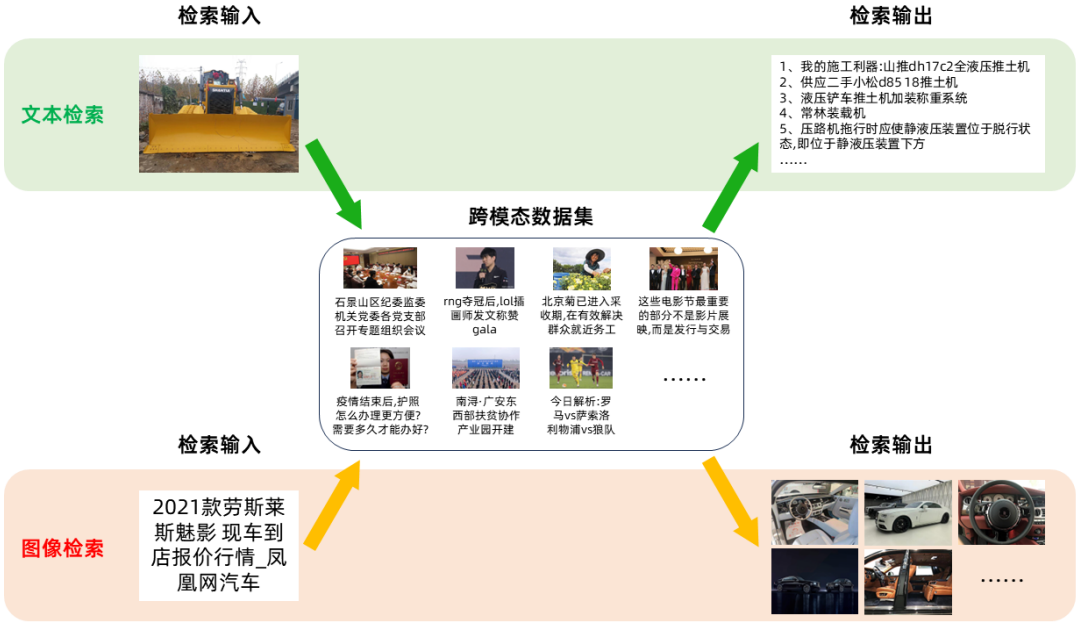

图 1: 图像文本检索

2. 解决问题

本赛题是利用附件 1 的数据集，选择合适方法进行图像和文本的特征提取，基于提取的特征数据，建立适用于**图像检索**的多模态特征融合模型和算法，以及建立适用于**文本检索**的多模态特征融合模型和算法。基于建立的“多模态特征融合的图像文本检索”模型，完成以下两个任务，并提交相关材料。

(1) **图像检索文本**

基于图像检索的模型和算法，利用附件 2 中“word_test.csv”文件的文本信息，对附件 2 的 ImageData 文件夹的图像进行图像检索，并罗列检索相似度较高的前五张图像，将结果存放在“result1.csv”文件中（模板文件详见附件 4 的 result1.csv）。其中，ImageData 文件夹中的图像 ID 详见附件 2 的“image_data.csv”文件。

(2) **文本检索图像**

基于文本检索的模型和算法，利用附件 3 中“image_test.csv”文件提及的图像 ID，对附件 3 的“word_data.csv”文件进行文本检索，并罗列检索相似度较高的前五条文本，将结果存放在“result2.csv”文件中（模板文件见附件 4 的 result2.csv）。其中，“image_test.csv”文件提及的图像 ID，对应的图像数据可在附件 3 的 ImageData 文件夹中获取。

3. 评估指标

图像文本检索包括两个具体的任务，即文本检索（Image-to-Text，I2T），即针对查询图像找到相关句子；以及图像检索（Text-to-Image，T2I），即给定查询语句检索符合文本描述的图像。为了与现有方法公平地进行比较，在文本检索问题和图像检索问题中都采用了广泛使用的评价指标：召回率 Recall at K（R@K）。R@K 定义为查询结果中真实结果（Ground Truth）排序在前 K 的比率，通常 K 可取值为 1、5 和 10，计算公式如式 (1) 所示。

$$R@K = \frac{\text{Matched}_{\text{top-}K}}{\text{GroundTruth}_{\text{total}}} \tag{1}$$

其中，$\text{GroundTruth}_{\text{total}}$ 表示真实匹配结果出现的总次数，$\text{Matched}_{\text{top-}K}$ 表示在排序前 K 个输出结果中出现匹配样本的次数。$R@K$ 反映了在图像检索和文本检索中模型输出前 K 个结果中正确结果出现的比例。本赛题的评价标准设定 $K = 5$，即评价标准为 $R@5$。

4. 基本假设

为了构建图像文本双向检索模型，我们做出如下合理的假设：

1. 训练数据集中的图像文本匹配关系正确可靠；
2. 训练集与测试集中的图像文本对的具有一致的数据分布；
3. 测试集中，每幅图像都存在与之匹配的文本，每条文本都存在与之匹配的图像。

二、 问题建模

1. 图像建模

在计算机视觉（Computer Vision，CV）领域，常用的图像表示方法是使用张量。张量是多维数组的扩展，可以表示高维数据。对于图像彩色，我们使用三维张量 $x \in \mathbb{R}^{H\times W\times C}$ 描述。其中，H 表示高度，W 表示宽度，C 表示通道数（对于彩色图像，通道数通常为 3）。

由于题目数据中的图像具有不同的长宽比、分辨率，不利于模型统一处理。于是，我们按照以下规则，对所有图像进行预处理：

1. 对于所有长宽比小于 2:1 的图像，将其拉伸为 1:1，使用双立方插值法（Bicubic Interpolation）下采样至 224×224 分辨率。
2. 对于所有长宽比大于 2:1 的图像，截断其长边，仅保留长宽比小于 2:1 的部分，再按照规则 1. 进行处理。

至此，我们可以将所有图像的分辨率处理为 224×224，进而使用四维张量 $X \in \mathbb{R}^{N\times H\times W\times C}$ 表示整个数据集，其中 N 为图像数量，$H = W = 224$，$C = 3$。

2. 文本建模

在自然语言处理（Natural Language Processing，NLP）领域，文本被视作一个由单词组成的序列。为了便于表达，将所有可能出现的单词汇集成一张表，称为词汇表（Vocabulary），其中每个单词对应一个唯一的序号（Index）。

为了使用深度学习模型学习单词的语义，我们需要将每个词语用一个固定长度的向量表示，分为稀疏表示（如 One-hot 编码）和分布式表示（如 Word2Vec）。由于稀疏表示的诸多弊端，这里我们采用单词的分布式表示。分布式表示将词转化为一个定长（设为 D_{emb}）、稠密并且互相存在语义关系的向量。此处的存在语义关系可以理解为：分布相似的词，是具有相同的语义的。

如此一来，一切文本都能被映射为一个由定长词向量组成的序列。然而，文本中单词的数量或多或少，因此单词序列的长度无法确定，这是不利于语言建模的。为了解决这个问题，常用的方法是指定一个最大序列长度（设为 L），然后按以下规则处理不同长度的文本：

1. 对于单词数量小于 L 的文本，在其后方填充若干特殊的单词“<pad>”，使其长度达到 L。
2. 对于单词数量超过 L 的文本，舍弃第 L 个单词后的内容。

于是，我们可以将所有文本处理为长度为 L 的单词序列，其中每个单词被表示为一个 D_{emb} 维向量。也就说说，一段文本可以被表示为一个形状为 $L \times D_{\text{emb}}$ 的矩阵。进而，我们对数据集中所有文本进行处理，得到一个三维张量 $Y \in \mathbb{R}^{M\times L\times D_{\text{emb}}}$，其中 M 是文本数量。

3. 图文检索建模

为了进行图文检索（包括图像检索文本、文本检索图像），关键是定义一个匹配度函数。该函数的输入为一幅图像以及一段文本，输出为图像与文本的内容匹配程度，即 $\text{Match}\,(Image, Text) \in [-1, 1]$。其数值大小表示匹配程度，-1 表示完全不匹配，1 表示完全匹配。

对于图像集合 $X \in \mathbb{R}^{N\times H\times W\times C}$，以及文本集合 $Y \in \mathbb{R}^{M\times L\times D_{\text{emb}}}$，可以得到一个匹配度矩阵 $\text{Score} \in \mathbb{R}^{N\times M}$ 表示所有“图像—文本”对的匹配情况。具体而言，Score 的定义如公式 (2) 所示。

$$\text{Score}\left[i,\ j\right] = \text{Match}\left(X_i,\ Y_j\right),\ 1 \le i \le N,\ 1 \le j \le M. \tag{2}$$

(1)　图像检索文本

在图像检索文本（Image-to-Text，I2T）任务中，我们需要为每幅图像寻找与其匹配程度最高的 K 段文本。而每幅图像与 Score 矩阵中的一行所对应，为了实现该目的，我们沿着行方向对 Score 矩阵进行 ArgSort 操作，使每行的文本按照与每幅图像的匹配程度排序，并以检索形式呈现。接下来，取检索矩阵的前 K 列，得到矩阵 $\text{RowTop} \in \mathbb{R}^{N\times K}$，如公式 (3) 所示，其中 [...] 表示子矩阵检索操作。

$$\text{RowTop} = \text{ArgSort}\left(\text{Score},\ dim = 1\right)\left[:,\ :K\right] \tag{3}$$

此时，RowTop 的第 i 行对应与图像 X_i 匹配程度最高的 K 段文本的位置，则 I2T 任务的结果如公式 (4) 所示，其中 {...} 表示集合。

$$\text{I2T}\left(X_i\right) = \left\{Y_j \mid j \in \text{RowTop}\left[i,\ :\right]\right\} \tag{4}$$

(2)　文本检索图像

类似的，在文本检索图像（Text-to-Image，T2I）任务中，我们需要为每段文本寻找与其匹配程度最高的 K 幅图像。而每段文本与 Score 矩阵中的一列所对应，为了实现该目的，我们沿着列方向对 Score 矩阵进行 ArgSort 操作，使每列的图像按照与每段文本的匹配程度排序，并以检索形式呈现。接下来，取检索矩阵的前 K 行，得到矩阵 $\text{ColTop} \in \mathbb{R}^{K\times M}$，如公式 (5) 所示。

$$\text{ColTop} = \text{ArgSort}\left(\text{Score},\ dim = 0\right)\left[:K,\ :\right] \tag{5}$$

此时，ColTop 的第 j 列对应与文本 Y_j 匹配程度最高的 K 幅图像的位置，则 T2I 任务的结果如公式 (6) 所示。

$$\text{T2I}\left(Y_j\right) = \left\{X_i \mid i \in \text{ColTop}\left[:,\ j\right]\right\} \tag{6}$$

根据所建立的模型，我们只需实现 $\text{Match}\left(Image, Text\right)$ 函数，得到图像与文本的匹配度，即可完成 I2T 任务与 T2I 任务。下面，我们将聚焦于该函数的实现。

三、 CLIP 模型

CLIP（Contrastive Language-Image Pre-training）是一个跨模态学习模型 [Radford et al., 2021]，由 OpenAI 在 2021 年提出。CLIP 模型的核心思想是通过对比学习的方式，将图像和文本映射到

同一个嵌入空间中，使得语义上相关的图像和文本在该空间中更接近。

1. 对比学习

对比学习（Contrastive Learning）是一种自监督学习方法 [Liu et al., 2021]，它通过比较数据的不同变体或不同数据对来学习数据的表示。对比学习的核心思想是：相似的样本在表示空间中应该接近，而不相似的样本应该远离。这种方法通常用于学习数据的低维表示，使其能够捕捉数据的本质特征。

对比学习的实施通常包括以下几个方面：

1. **正负样本的定义**：在对比学习中，图片特征和文本特征构成特征矩阵，该矩阵中图文相匹配为正样本，不匹配为负样本，因此特征矩阵的对角线元素均为正样本，其他元素为负样本。
2. **相似度计算**：使用余弦相似度来表示特征之间的相似度，A、B 矩阵的余弦相似度可由公式 (7) 描述。

$$\text{Cosine-Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} \tag{7}$$

3. **损失函数**：定义一个损失函数来训练模型，使得正样本对的相似度高于负样本对的相似度。典型的损失函数包括三元组损失（Triplet Loss）、对比损失（Contrastive Loss）和交叉熵损失等。

2. 模型结构

RoBERTa[Cui et al., 2021]

ViT[Dosovitskiy et al., 2020]

CLIP 模型的核心思想是通过学习图像和文本之间的匹配关系来提高模型的性能。具体来说，CLIP 模型包含两个主要组成部分：一个用于处理图像的 CNN 模型或 ViT 模型，和一个用于处理文本的 BERT 模型。这两个组件都被训练成能够将输入的信息映射到相同的嵌入空间中，并使得相似的图像和文本在嵌入空间中的距离更近。图 2 演示了 CLIP 模型的结构。

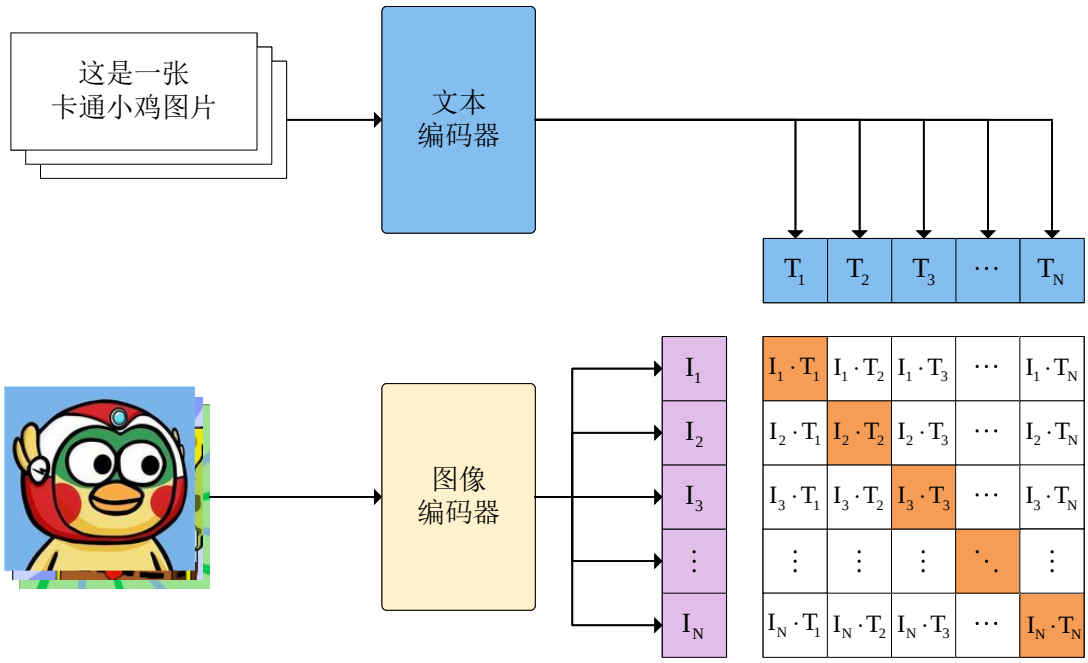

图 2: CLIP 模型结构

3.　学习目标

4.　下游任务

四、　模型预训练

1.　数据集构建

2.　“零样本”效果

五、　模型微调

1.　数据集构建

2.　微调效果

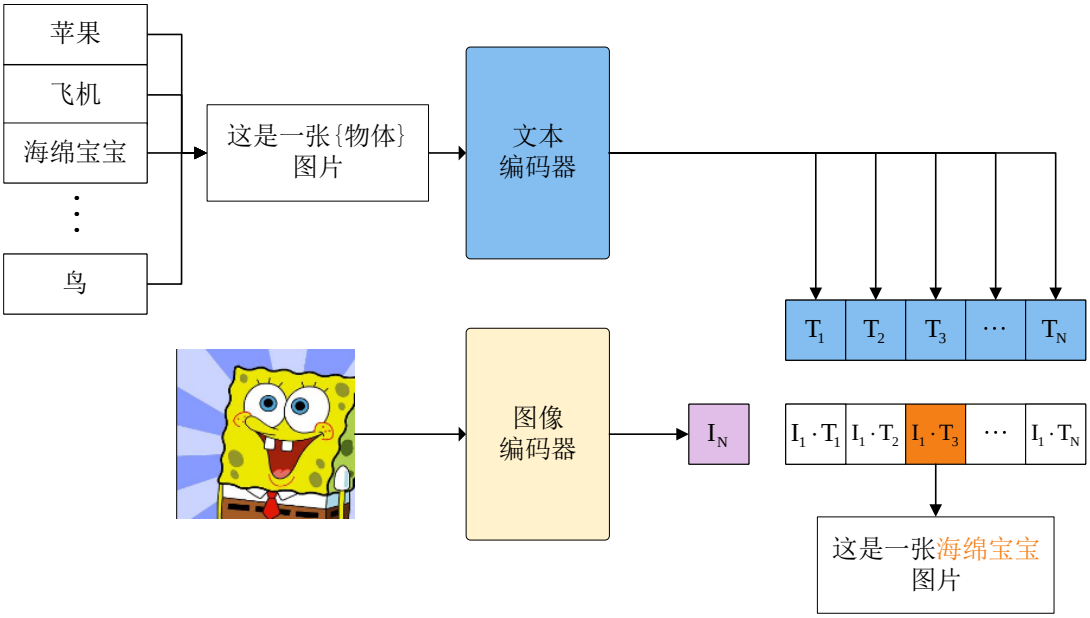

图 3: CLIP 模型推理——图像检索文本

参考文献

[Cui et al., 2021] Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 248–255. Ieee.

[Diederik, 2014] Diederik, P. K. (2014). Adam: A method for stochastic optimization. *(No Title)*.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent.*

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778.

[Krizhevsky et al., 2017] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

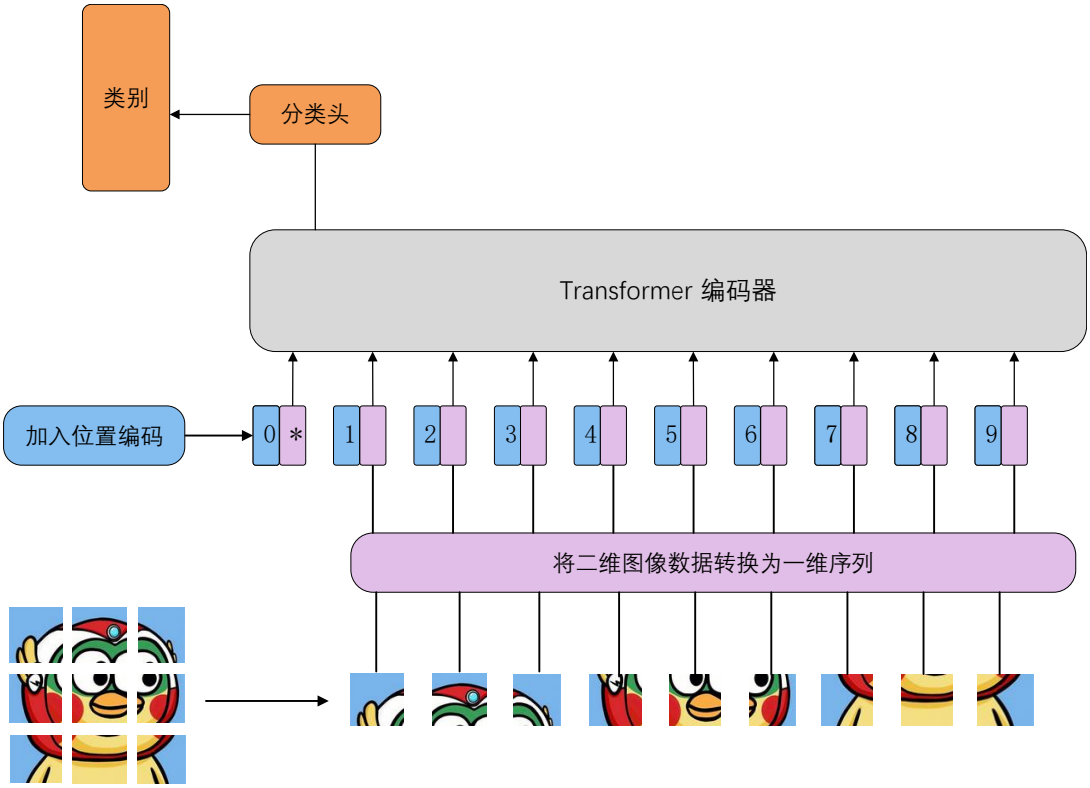

图 4: 图像文本检索

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Proc. Adv. Neural Inf. Process. Syst.*, 2.

[Liu et al., 2021] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876.

[Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Proc. Adv. Neural Inf. Process. Syst.*, 30.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612.

[Zhao et al., 2016] Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imag.*, 3(1):47–57.