随着移动设备、社交网络和自媒体平台的快速 发展，近些年文本、图像、视频和音频等多媒体数据 呈现爆炸式增长。最新数据显示，Youtube每分钟有长达300 h的视频上传；Flickr、Snapchat等图片 分享网站每天有上亿的图片上传；Facebook、Twitter 等社交媒体网站上，用户每天都分享着海量的文本、图片和视频等多媒体数据。这些不同类型的数 据通常都是描述同一个物体或同一个事件。如网页通常由文字、图像或视频来共同表达同一个事件 或主题。面对如此巨大而相互关联的多媒体数据，用户迫切希望能够利用其中一种模态(如文本)同时检索到与之相关的其他模态结果(如图像、视频和音频等)，即跨模态检索。

首先我们先从基于内容的图像检索作为一个铺垫。图像检索有基于文本的检索和基于内容的检索两种，如果是基于语义的检索的话，在检索之前需要对海量的图片进行语义属性的标注，这种标注有主观性偏差，时间成本很高，而且语义属性也不能完全表达图像中的包含的丰富的信息，检索效果是有限的。基于内容的检索 ，“以图搜图”就有他独特的优势。百度识图，google搜图，淘宝，京东都已经支持用图搜图的形式，它是按照图像特征的描述来搜索，然后通过比较特征间的相似性，按照相似性大小来排序给出最终的检索结果。

一个很直观的例子就是，比如我们在电视剧、网络上看到一款衣服很喜欢，又不知道在哪里能买到，在淘宝上基于关键词的去搜索，可能并不能搜到同款，这时候，图像检索就发挥起了作用，用户只需要将衣服的照片直接输入，就快速准确的检索到相应的店家。就像ppt中的例子所示。

我们知道，网络上充斥着来自不同数据源的多模态多媒体数据；因此，就需要能够适应不同模态的信息检索系统。比如说，我们去故宫旅游，我们对故宫的瓷器感兴趣，那么我们就可以拍下来，输入到跨模态检索系统中，输入图片，输出跟这个瓷器相关的文本描述。

以上图中两个例子来说明跨媒体检索，如Text2Img和Img2Text。输入一张图像，检索出描述这个图像相关文本，输入一个文本，检索出与文本描述相关的图像场景

跨模态检索中不同模态数据之间呈现底层特征异构，高层语义相关的特点。对于同一类别的不同模态数据，文本特征通常通过字典向量、word2vec、glove等来表示，而图像则通过SIFT、cnn特征 等视觉特征来表示，它们分别处于完全不同的特征空间，却表示同一个语义主题，因此不同模态之间的数据及模型具备异质性。我们将wiki数据集中文本特征和图像特征，用tsne聚类可视化显示出来，由图中特征空间的分布结果可以看出，图像和文本数据的特征表达具有本质性的差异，因而无法直接度量它们之间的相似度。

所以跨模态检索的主要问题是如何度量不同模态之间的相似性，关键就在于如何关联不同模态，难点就是跨越语义鸿沟。

所以目前的研究主要是通过对不同模态之间的关系进行建模，常见的方法是公共子空间学习，为不同模态学习具有代表性的特征表示，同时，建模相同数据在不同模态之间的相关性，然后在共同潜在空间中进行相似性度量，从而实现跨模态的检索。从图中可以看出，对于原始特征，图片有图片所在的特征空间，文本有文本所在的特征空间，通过公共子空间的学习，使得，相关的文本和图像在公共子空间中，更接近，不相关的文本和图像对，距离更远