



中国科学院大学

University of Chinese Academy of Sciences

研究生学位论文开题报告

报告题目_____融合知识指导和数据驱动_____

_____的视觉显著性检测_____

学生姓名_____崔哲_____学号_____201828008629005_____

指导教师_____苏荔_____职称_____副教授_____

学位类别_____工学硕士_____

学科专业_____计算机应用技术_____

研究方向_____计算机视觉_____

研究所(院系)_____计算机科学与技术学院_____

填表日期_____2019/12/24_____

中国科学院大学

1. 选题的背景及意义

随着现今图像数据的爆炸式增长，人类所接收的信息量呈指数级增长，“一视同仁”的数字图像处理机制不仅速度慢，而且浪费资源。人类可以很容易地判断出图像中的显著区域，不管图像的环境背景如何复杂，都能实时地注意到图像中的重要部分，选择出图像的子集进行深度处理，以此减少场景的复杂度，这种选择部分区域来进行注意的机制叫做视觉注意机制。视觉注意机制启迪我们用计算机视觉模拟人类视觉，提取出图像中的显著区域或者感兴趣的部分进行集中处理。图像的显著区域检测属于视觉显著性分析范畴，其将视觉注意机制引入到图像处理中，通过筛选出人眼认为的重要区域，从而减少图像数据处理量。

视觉显著性检测研究主要分为两个方向：人眼关注点检测和显著性物体检测，两者的任务不同，如图 1-1 所示，本文研究的方向为显著性物体检测中的二值分割。

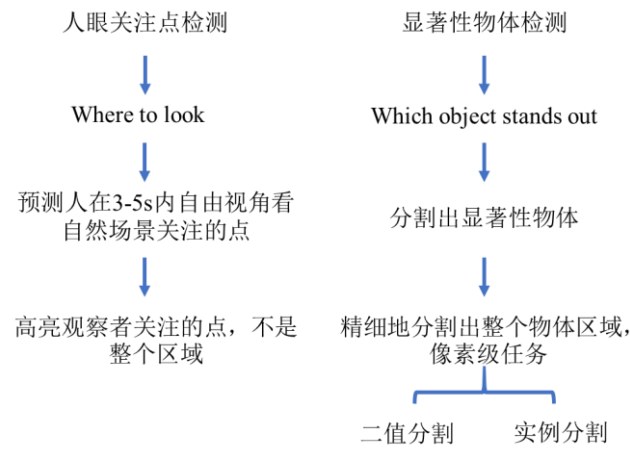


图 1-1 人眼关注点检测 vs 显著性物体检测

视觉显著性描述的是在一幅场景中一个物体吸引视觉注意的能力，这种能力源自该物体与周围事物的迥异性，由观察者的主观经验引起^[1]。显著度检测的目标是检测图像中的显著目标，滤除冗余的背景信息，从而只关注人类视觉感兴趣的图像区域，如图 1-2 给出了图像显著度检测的示例。



图 1-2 图像显著度检测示例

应用需求推动研究发展，随着信息技术的快速发展，如何设计及利用计算机处理这些以爆炸式速度增长的图像及视频信息，对于弥补人与计算机在视觉理解中的差距具有重要的研究意义和应用价值，显著度检测已成为计算机视觉领域的重要研究课题。可靠的显著度检测可以在没有先验知识的情况下筛选出图像中的重要信息，在降低图像内容理解和场景分析复杂度方面具有重要意义。因此，图像的显著性区域检测是许多计算机视觉任务中的一个重要步骤，包括图像分割、目标识别、自适应压缩、内容感知的图像编辑和图像检索等，高效的视觉显著度检测能大大提升这些任务的整体效率。

2. 国内外本学科领域的发展现状与趋势

作为计算机视觉中的一个重要问题，图像中的显著目标检测(SOD)近年来得到了越来越多的研究。越来越多的国内外研究机构、大学开始重视对图像显著性的研究，并不断取得突破性的成果。

关于显著度的研究是从生物研究发展而来，早期 Koch 等人^[2]通过对人类视觉系统的研究，在认知心理学方面将有显著性的特征进行了整合，提出了非常有影响力的生物启发模型，即视觉注意理论，但并未具体实现。

图 2-1 简要展示了显著性检测方法的发展历史，与其他计算机视觉任务相比，显著性检测的历史相对较短，可以追溯到先驱者 Liu^[3]和 Achanta^[4]的工作。大多数传统的非深度学习方法都是基于低级特征，并依赖于某些启发式方法(例如，颜色对比，背景先验等)。

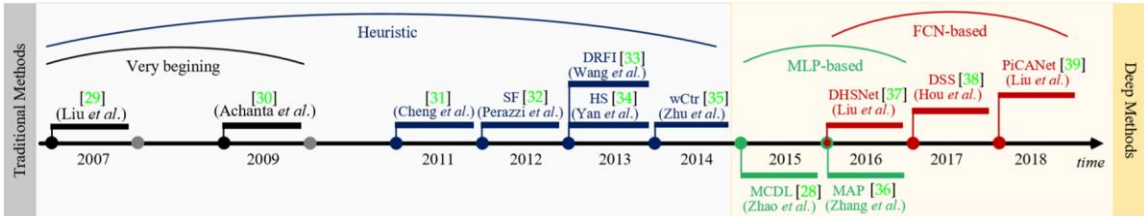


图 2-1 显著性检测的简要发展历史^[5]

随着深度学习技术在计算机视觉领域取得的巨大成功，自 2015 年以来，越来越多的基于深度学习的显著性检测方法应运而生。早期的深度显著性检测模型通常利用多层感知器(Multi-layer Perceptron, 简称 MLP)来预测从每个图像处理单元中提取的深度特征的显著性得分。后来，一种更加有效和高效的形式，即基于全卷积网络(Fully Convolutional Network, 简称 FCN)的网络成为了显著性检测网络结构的主流。

2.1 基于 MLP 的方法

基于 MLP 的显著性检测方法通常为图像的每个处理单元提取深层特征，以训练 MLP 分类器进行显著性得分预测，如图 2-2 所示。根据图像处理单元的不

同可以分为基于 Super-pixel/patch 的方法和基于 Object proposal 的方法。

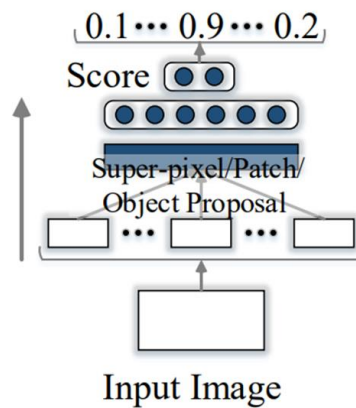


图 2-2 显著性检测的简要发展历史^[5]

Zhao 等人^[6] 于 2015 年设计了一个包含多个上下文的深度模型来捕获对象显著性，其网络模型如图 2-3 所示。全局上下文用于在全图像中建模显著性，而局部上下文用于在细节区域中进行有效预测。将全局上下文和局部上下文集成到多上下文深度学习框架中进行显著性检测，并对全局和局部上下文建模进行联合优化。

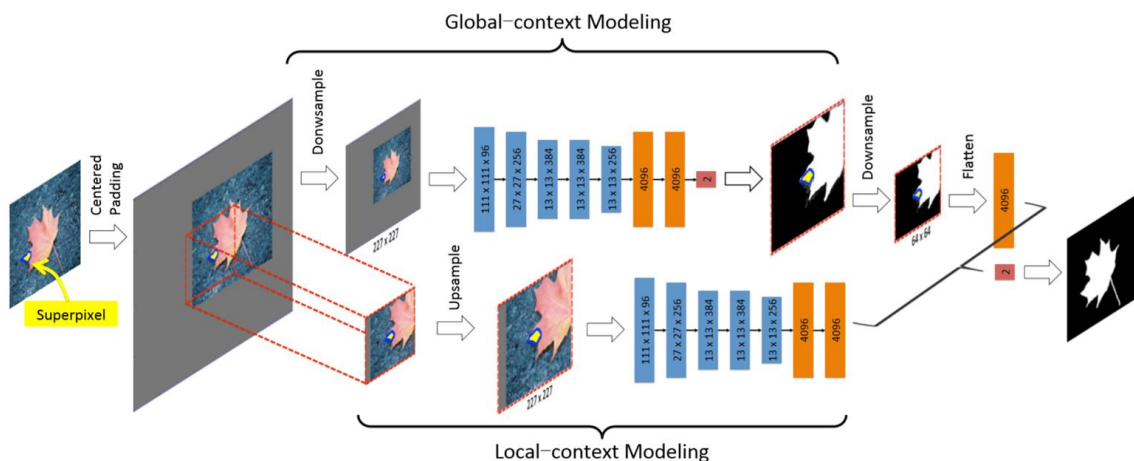


图 2-3 网络模型^[6]

基于 MLP 的方法存在以下问题：

- (1) 无法利用整体空间信息；
- (2) 耗时(需要对每一个 patch 送入网络进行处理)；
- (3) 每一个处理单元的所有像素共享相同的显著性预测得分。

2.2 基于 FCN 的方法

尽管基于 MLP 的显著性检测模型的性能优于以前的非深度学习显著性检测模型，但它们无法捕获关键的空间信息，并且由于它们需要逐一处理所有子单元而非常耗时。受到 FCN 在语义分割方面的巨大成功的启发，最新的显著性检测模型将图像分类模型(如 VGGNet 和 ResNet)修改为全卷积模型。这样，这些显著

性检测模型将受益于端到端的空间显著性表示学习，并且可以在单个前馈过程中有效地预测显著性图。典型的网络结构可以分为五类：单工作流网络 (Single-stream network)、多工作流网络 (Multi-stream network)、多分支融合网络 (Side-fusion network)、自底而上/自顶而下网络 (Bottom-up/top-down network) 和多任务网络 (Branched network)，如图 2-4 所示。

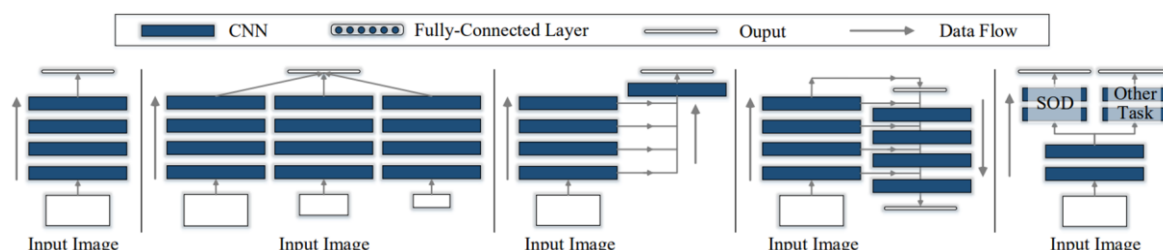


图 2-4 基于 FCN 的方法网络结构分类^[5]

2.2.1 单工作流网络

Wang 等人^[7]于 2016 年提出的显著性检测模型是单工作流网络结构的代表，其网络模型如图 2-5 所示。在每一个时间步中，通过 rfcn 将输入的图像和显著性先验图都向前推进，得到预测的显著性映射，进而作为下一个时间步的显著性先验图。第一个时间步中的先验映射通过显著性先验来初始化。

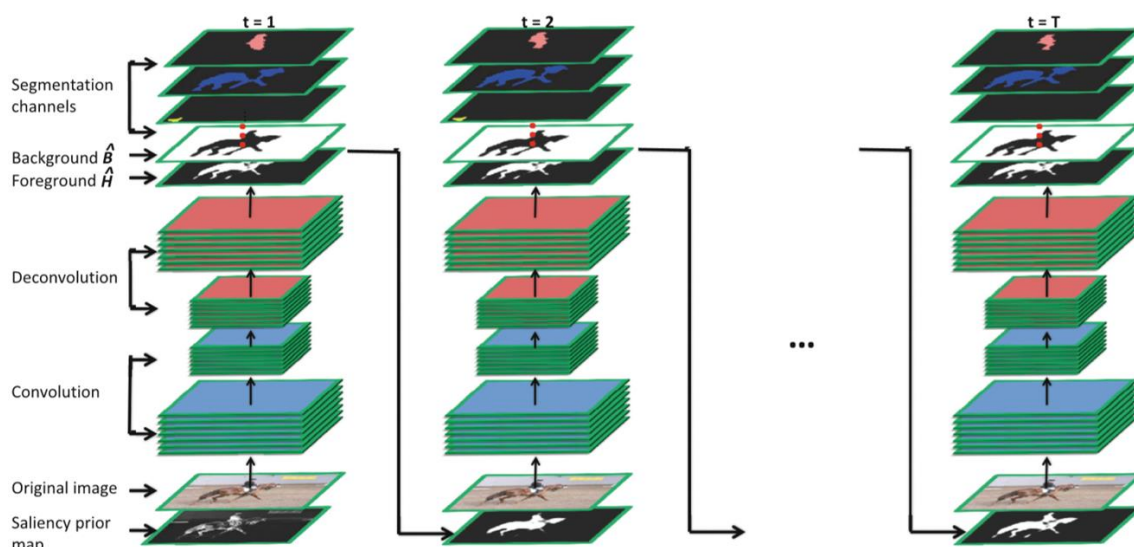


图 2-5 网络模型^[7]

2.2.2 多工作流网络

Wang 等人^[8]于 2017 年提出的显著性检测模型是多工作流网络结构的代表，其网络模型如图 2-6 所示。在第一阶段，深度前馈网络生成一个粗糙的显著性图，如图 2-6 中(d)，其中丢失了很多详细的结构。然后，利用增强网络逐级连续地重

新精细化前面的显著性图，如图 2-6 中(e)，阶段性增强网络有效地将主网络层中的高级语义特征与增强网络中的低级特征的丰富空间信息相结合。

主网络有助于定位显著性对象，而带有空间金字塔模块的增强网络有助于逐渐生成更精细的细节并嵌入全局上下文信息。

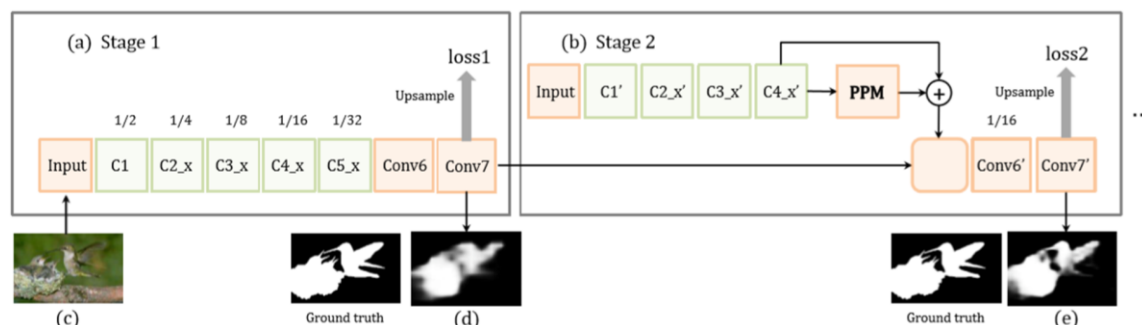


图 2-6 网络模型^[8]

2.2.3 多分支融合网络

Hou 等人^[9]于 2017 年提出的显著性检测模型是多分支融合网络结构的代表，其网络模型如图 2-7 所示。现有的基于 FCN 的显著性检测模型没有明确地处理空间尺度问题。Holistically Nested Edge Detector(简称 HED)在边缘检测中提供了跳跃层结构(skip-layer)处理空间尺度问题，但 HED 在显著性检测上性能并不明显。文章基于 HED 的结构，采用 short-connection 连接 skip-layer，在 VGGNet 的每一个 stage 的最后一个卷积层连接 side output 层，并将高层信息通过 short connection 连接到低层，然后分别与 ground-truth 计算交叉熵损失。最后将所有输出进行融合，融合后的输出再与 ground-truth 计算交叉熵损失，有效地结合了多层信息。高层信息帮助低层信息定位显著性物体低层信息细化显著性物体空间信息。

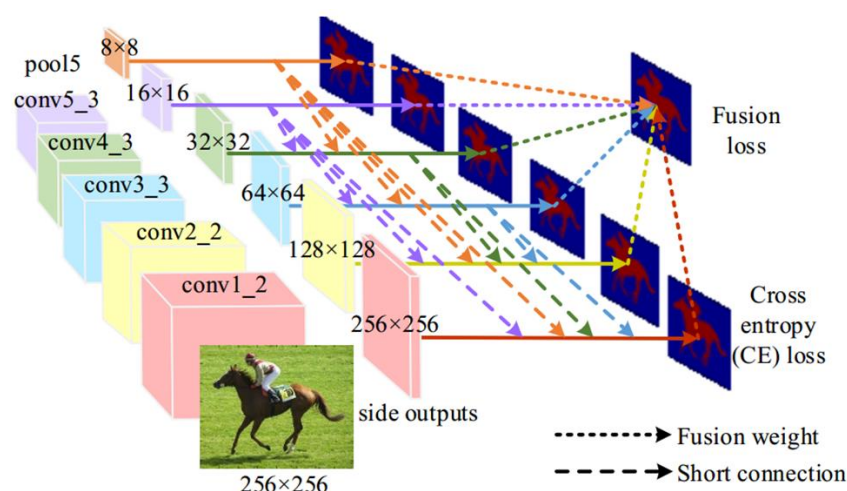


图 2-7 网络模型^[9]

2.2.4 自底而上/自顶而下网络

Chen 等人^[10]于 2018 年提出的显著性检测模型是自底而上/自顶而下网络结构的代表,其网络模型如图 2-8 所示。尽管学习显著性优化的残差细节是自然和直接的,但是如果没有额外的监督,网络很难准确地捕捉到它们。大多数现有的深度显著性模型都是对图像分类网络进行微调,因此,在残差学习过程中,微调的网络将无意识地集中于具有高响应值的区域,从而难以捕获残差细节,例如物体边缘和其他未检测到的目标。为了解决这一问题,提出了反向注意(reverse attention)以自顶向下的方式引导侧输出残差学习。通过从侧输出特征中去除当前预测的显著区域,网络最终可以发现丢失的目标部分和细节,从而获得高分辨率和高精度。

该方法首先在最深层生成一个语义清晰但分辨率较低的粗糙显著图,通过侧边输出特征中删除当前预测的显著区域(当前预测从其较深的层上采样得到),引导整个网络依次发现互补目标区域和细节。这种自上而下的擦除方式最终可以将粗分辨率和低分辨率预测重新完善为一个完整的、高分辨率的显著性图。

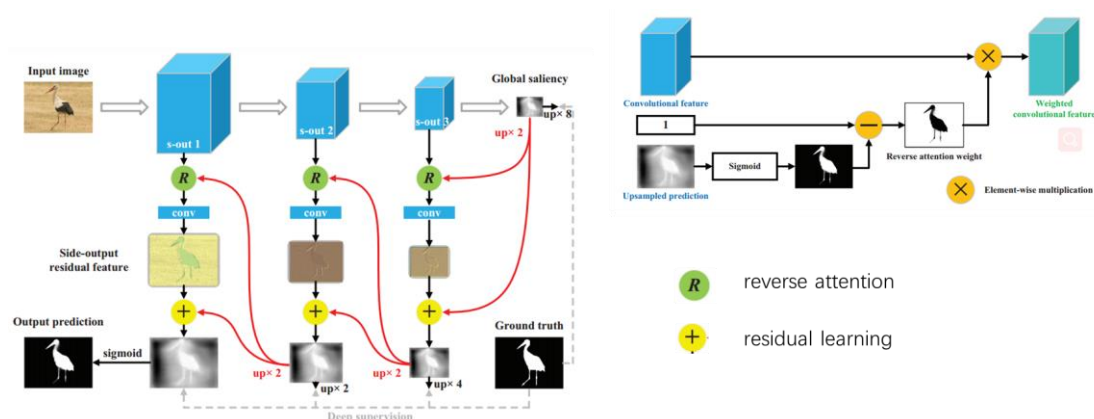


图 2-8 网络模型^[10]

2.2.5 多任务网络

人类的视觉注意力,作为一种直观的先验知识,更加符合人类视觉系统处理视觉信息的生理过程,人眼定位通常与场景中显著对象的位置相关。然而,只有少数方法试图同时解决眼动点检测和物体显著性检测。

Kruthiventi 等人^[11]于 2016 年提出的显著性检测模型是多任务网络结构的代表,其网络模型如图 2-8 所示。本文提出了一种深度卷积神经网络(CNN),它能够在一个单一的框架中预测眼睛的视觉变化和分割显著的物体。文章提出了一个具有分支结构的全卷积网络,同时用于眼动点预测和显著对象分割。共享网络层由 6 个卷积块组成,受 VGG16 启发,前 5 个卷积块采用小卷积核(3*3)以减少参数量。前 5 个卷积块参数由 VGG16 进行初始化,在 VGG16 网络中,每经过一

个卷积块特征图尺寸便减半，由于像素级预测任务最后的预测图的尺寸同样重要，所以在 conv4 和 conv5 中保持特征图尺寸为原图的 1/8 不变。在原 VGG16 网络中，conv5 中卷积层所处理的特征图为原图的 1/16，本文通过空洞卷积来处理这种尺寸的不匹配问题，保证每个卷积块的感受域不变。

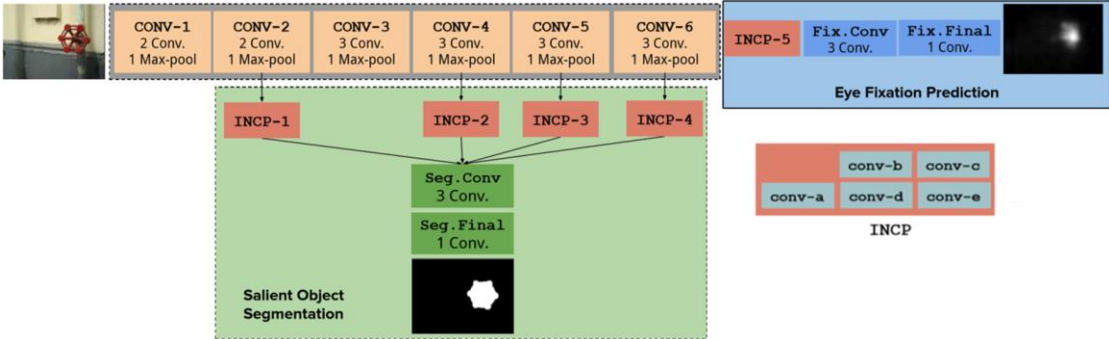


图 2-9 网络模型^[11]

3. 课题主要研究内容、预期目标

尽管得益于深度学习技术，显著目标检测已经取得了重大进展，但要准确检测杂乱场景中的显著物体仍然是一项巨大的挑战利用图片描述指导显著性检测：对比不仅表现在视觉线索的差异上，而且还涉及到高层次的认知和理解。为了更好地检测语义显著的对象，高层次的语义特征变得非常重要。为了解决这个问题，模型需要学习针对显著对象的语义特征，例如对象类别，属性和语义上下文。但是，现有的显著物体检测网络仅在像素级注释上进行训练，而没有对更高级别语义信息的监督。

因此，利用图片描述作为辅助语义任务，以提高复杂场景中的显著目标检测效果。图像描述和显著性检测之间的联系已经在图像描述领域中进行了探索。一些工作利用显著性检测使网络关注显著区域来辅助图像描述，这些工作假设图像描述中提到的对象在很大程度上与显著对象保持一致并相关。基于相同的假设，图像描述任务可以为显著对象检测提供丰富的语义监督。如图 3-1，从描述 “A group of people sitting on a boat in the water” 中，我们可以获得有关显著物体的类别，属性和运动的整体知识。



A group of people sitting on a boat in the water.

图 3-1 图像描述示例^[12]

4. 拟采用的研究方法、技术路线、实验方案及其可行性分析

拟采用的模型结构框架如图 4-1 所示，模型由一个共享主网络 ResNet 和两个分支网络（分别为图像描述网络和显著目标检测网络）组成。对于输入图像，使用共享 ResNet 来提取多级特征，使用基于 LSTM 的图像描述模型生成该图像的描述，然后使用注意机制合并每个单词的 hidden state，以获取描述嵌入特征向量(Caption Embedded Vector，简称 CEV)。将描述嵌入特征向量 CEV 和多级特征进行合并到显著目标检测网络中获得最终显著图。

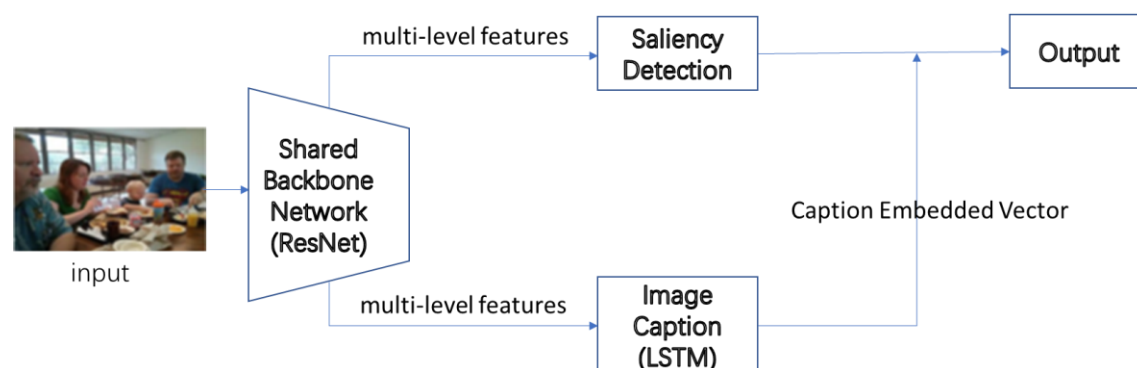


图 4-1 模型结构框架

图像描述网络以图像作为输入并生成描述。为了从描述获取对象级语义知识，使用 LSTM 的 hidden state 来表示每个生成单词的编码特征。考虑到并非标题中的每个单词都与描述重要对象相关，采用一种文本关注机制来加权每个单词的重要性。然后，可以通过对 LSTM 的 hidden state 进行加权合并来获得描述嵌入特征向量 CEV。

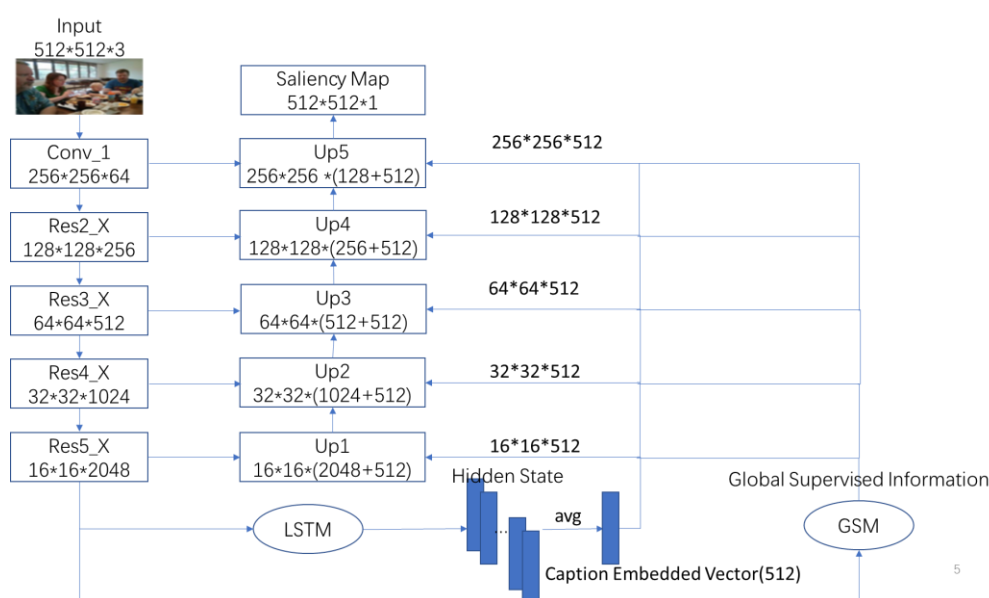


图 4-2 网络模型






















显著目标检测网络采用自底而上/自顶而下网络结构，将描述嵌入特征向量 CEV 与 ResNet 提取的多级视觉特征进行集成，显著目标检测网络利用图像描述

网络生成的描述嵌入特征向量 CEV 来捕获场景的高级语义信息，以识别显著物体，显著目标检测网络和图像描述网络在训练期间共同优化，整个网络的结构如图 4-2 所示。

5. 已有研究基础与所需的研究条件

已完成图像描述网络和显著目标检测网络的搭建工作，显著目标检测网络取得了预期的结果，图像描述网络还没有调优完成。

数据集	MAE	Max-Fmeasure	Max-Emeasure	S-Measure
DUTS-TEST	0.0396	0.8678	0.9039	0.8792
ECSSD	0.0399	0.9379	0.9396	0.9174
PASCAL	0.0762	0.8628	0.8731	0.8448
HKUIS	0.0333	0.9272	0.9412	0.9117
SED1	0.0538	0.9229	0.9122	0.8928
SED2	0.0927	0.7993	0.8054	0.7901

Input							
gt							
Output							

6. 研究工作计划与进度安排

时间	任务
2019年12月-2020年2月	阅读相关参考文献，完善研究方案。
2020年3月-2020年6月	完成图像描述网络，将各模块进行整合调优，准备论文投稿与专利申请。
2020年7月-2020年12月	模型完善，准备中期答辩。
2021年1月-2021年7月	根据中期验收结果对系统进行进一步完善，撰写毕业论文以及准备相应的材料，准备毕业答辩。

7. 参考文献

- [1] Treisman M, Gelade G. A Feature-Integration Theory of Attention[M]. Oxford University Press, 2012:117-119.
- [2] Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry[J]. Human Neurobiology, 1987(4):219-227.
- [3] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.
- [4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1597–1604.
- [5] Wang W, Lai Q, Fu H, et al. Salient Object Detection in the Deep Learning Era: An In-Depth Survey[J]. arXiv preprint arXiv:1904.09146, 2019.
- [6] Zhao R, Ouyang W, Li H, et al. Saliency detection by multi-context deep learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1265-1274.
- [7] Wang L, Wang L, Lu H, et al. Saliency detection with recurrent fully convolutional networks[C]//European conference on computer vision. Springer, Cham, 2016: 825-841.
- [8] Wang T, Borji A, Zhang L, et al. A stagewise refinement model for detecting salient objects in images[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4019-4028.
- [9] Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3203-3212.
- [10] Chen S, Tan X, Wang B, et al. Reverse attention for salient object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 234-250.
- [11] Kruthiventi S S S, Gudisa V, Dholakiya J H, et al. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5781-5790.