研究与开发

基于图像描述算法的离线盲人视觉辅助系统

陈悦[1]，郭宇[1,2]，谢圆琰[1]，米振强[1]
（1. 北京科技大学计算机与通信工程学院，北京 100083；
2. 北京科技大学顺德研究生院，广东 佛山 528399）

摘　要：针对现有盲人视觉辅助设备存在的不便，探讨了基于模型剪枝的图像描述模型在便携式移动设备上运行的方法。回顾了图像描述模型和剪枝模型技术，重点提出了一种针对图像描述模型的改进剪枝算法。结果表明，在保证准确性的前提下，剪枝后的图像描述模型可以大幅降低工作时的处理时间和消耗的电源容量，能够随时随地快速准确地对环境信息进行描述及语音播报。

关键词：视觉辅助系统；图像描述模型；模型压缩和加速；模型剪枝算法
中图分类号：TP391
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2022014

Offline visual aid system for the blind based on image captioning

CHEN Yue[1], GUO Yu[1,2], XIE Yuanyan[1], MI Zhenqiang[1]
1. School of Computer & Communication, University of Science and Technology Beijing, Beijing 100083, China
2. Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China

Abstract: In view of the inconveniences of existing visual aid systems for the blind, the method of running the image captioning model on portable mobile devices based on model pruning was discussed. Model pruning techniques and image captioning models were reviewed. An improved model pruning algorithm for image captioning model was proposed. Experimental results show that, on the premise of ensuring accuracy, the image captioning model after pruning can greatly reduce processing time and power consumption capacity, and can quickly and accurately describe environmental information and voice broadcast anytime and anywhere.

Key words: visual assisted system, image captioning model, model compression and acceleration, model pruning algorithm

0　引言

视觉障碍群体是残疾人群中容易被忽略的庞大人群，眼睛的缺陷让他们无法通过视觉系统感知外界的信息，从而给日常生活和出行带来极大不便。现阶段，视觉辅助设备给盲人的生活带来了一些便利[1]。但现有的盲人辅助工具或多或少存在着价格昂贵、辅助功能有限、交互性差、无法离线使用等缺点。基于此，本文提出了一种搭载在低成本便携设备中基于图像描述算法的离线

收稿日期：2021−09−17；修回日期：2021−11−19

盲人视觉辅助系统。

图像描述能够利用语言描述图像内容。2014 年提出 m-RNN 模型[2]和 NIC 模型[3]后，图像描述任务相较于基于检索的模型产生了较大进步，在此之后的研究结合目标检测等高等级语义信息实现高层次视觉任务[4-5]、结合场景图实现细粒度可控的图像描述模型[6-7]、生成独特精确且有信息量的图像描述[8-9]等方面对其改进。

随着卷积神经网络[2-9]的发展、模型精度的不断提升，神经网络计算量越来越大的同时还伴随了大量的冗余。这造成了实现深度学习网络模型要么需要具备强大计算能力的设备，要么需要能够传输大量数据的网络。这对于实现能够随身携带、帮助视觉障碍人群提供日常服务的小型移动设备带来了巨大的挑战：一是小型移动设备无法完成大型深度学习网络计算量，二是人们不能保证自己时时都处于能够传输大量数据的网络环境中。基于此，在保证模型准确率的同时尽可能降低模型的复杂度成为了一个热门研究课题，从剪枝[10-11]、量化[12-13]、蒸馏[14-15]、低秩分解[16-17]、加法网络[18-20]等方面实现模型压缩，已经被广泛应用在各种模型上。

本文将在文献[10]的方法上做出改进，对典型的图像描述模型进行剪枝压缩，在确保图像描述精度的同时提高其运行速度，减少其工作消耗，并将其部署在低成本便携式移动设备上，盲人通过拍摄照片便可以收听到含有实时周围环境信息描述的语音播报。不同的模型剪枝算法也能够实现本文所实现的功能，但正如前文所述本文主要提出一种利用图像描述模型为盲人提供视觉辅助的方法，并采用模型剪枝的方法解决当前图像理解等神经网络模型计算量大，难以部署在低成本移动设备的问题，对不同模型剪枝的实现效果不在本文的重点考虑范围之内，因此对于不同压缩算法达到的压缩程度以及精度不做深入探讨。

1 图像描述模型的相关介绍

本文所采用的图像描述算法的整体框架为编码－解码（encoder-decoder）模型[21]。编码－解码模型原本用来解决自然语言处理中的序列到序列（sequence-to-sequence，seq2seq）问题，如自然语言翻译、文章摘要、问答系统等，编码－解码模型如图 1 所示。其中，编码－解码模型在最常见的自然语言翻译模型中，编码端和解码端使用的都是循环神经网络（recurrent neural network，RNN）模型，一种语言的输入通过 RNN 的编码部分，生成一个语义编码信息 C，之后经过 RNN 的解码部分输出为另一种语言。将图像描述模型嵌入到编码－解码模型架构中后，编码部分使用的是卷积神经网络，解码部分使用的是循环神经网络。

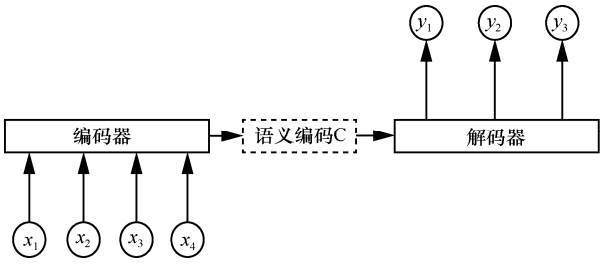


图 1 编码－解码模型

注意力（attention）机制类似于人眼的注意机制[22]，能够随着解码的进行改变对局部的注意力。在编码－解码模型的基础上加入软注意力（soft attention）机制，可以生成更合理的单词。加入软注意力机制后的图像描述整体框架如图 2 所示。

- 编码端：利用 VGG16 模型提取图像特征，本文只利用模型的卷积层，经卷积层提取之后最终形成注释向量（annotation vector）。
- 解码端：为了避免训练时的梯度消失现象，本文使用长短期记忆（long short-term memory，LSTM）网络[23]代替 RNN：

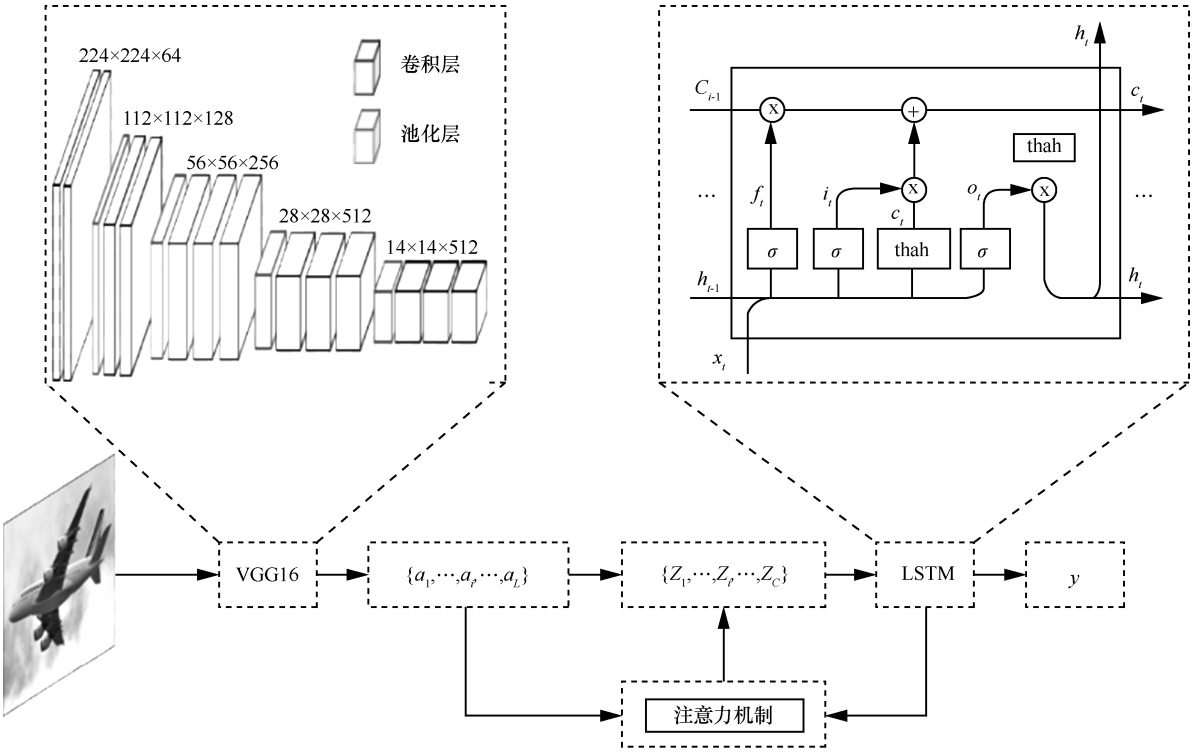


图 2　加入软注意力机制后的图像描述整体框架

$$\begin{cases} f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\ c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t * \tanh(c_t) \end{cases} \tag{1}$$

其中，式（1）中的σ为 sigmoid 激活函数，细胞状态（cell state）c_t和隐藏状态（hidden state）h_t的初始状态分别由两个独立的多层感知机 $f_{\text{init},c}$、$f_{\text{init},h}$计算得来：

$$c_0 = f_{\text{init},c}\left(\frac{1}{L}\sum_{i=1}^{L} a_i\right) \tag{2}$$

$$h_0 = f_{\text{init},h}\left(\frac{1}{L}\sum_{i=1}^{L} a_i\right) \tag{3}$$

加入注意力机制后的图像描述模型与普通图像描述模型的不同在于，上下文向量（context vector）$\{z_1,\cdots,z_t,\cdots,z_c\}$需要由注释向量 a_i 和注意力机制共同决定。$\{z_1,\cdots z_t,\cdots,z_c\}$是根据某个特定的局部图像信息而产生的上下文向量，注释向量 a_i 会产生一个权重 a_{ti}，在注意力机制中，权重 a_{ti} 是在 t 时刻图像区域 a_i 输入 LSTM 中所占的比重。权重 a_{ti} 由注释向量 a_i 和长短期记忆网络中的隐藏状态 h_{t-1} 之间的相关性计算。之后只需要将 a_i 和对应的 a_{ti} 加权求和就可计算上下文向量 z_t。这样，注意力机制就能够对不同的图像区域产生不同的关注度，进而生成更合理的词。

模型的复杂度一般用浮点数运算量（floating point operation，FLOP）衡量，卷积层 FLOP 的计算式[24]为：

$$\text{FLOP} = (2 \times C_i \times K^2 - 1) \times H \times W \times C_{\text{o}} \tag{4}$$

其中，$(2\times C_i\times K^2-1)$表示一次卷积操作的运算量，$(2\times C_i\times K^2-1)\times H\times W\times C$ 表示拓展到整个卷积操作的运算量。

本文在图像描述模型中采用的是经典的神经网络模型 VGG16 的卷积层部分提取图像特征，VGG16 网络结构的大部分运算量来自其卷积层。本文剪枝的主要目标是减少模型的运算量，也即

压缩方法将针对编码部分忽略解码的相关操作。

2 基于图像描述算法的离线盲人视觉辅助系统

2.1 系统框架

针对现存盲人视觉类辅助工具的不足及盲人对周围环境感知的急切需求，本文设计了如图 3 所示的离线盲人视觉辅助系统。对图像描述模型进行剪枝，使得其可以在低成本便携式移动设备中离线处理图像，解决现有视觉辅助设备价格昂贵、依赖网络、交互性不强等问题。该系统以广角相机拍摄的照片作为输入，之后通过剪枝处理的图像描述模型帮助盲人感知周围环境，并利用扬声器将图像描述模型得到的环境描述通过语音的方式播报，从而从听觉辅助视觉的角度帮助视觉障碍人士实现对环境的感知。本文将在下文对上述功能模块进行具体阐述。

2.2 图像描述模型及剪枝

为了向盲人提供生活上的便利，确保本系统能够离线处理图像并在确保图像描述模型准确度的基础上缩短图像描述模型的运行时间、降低图像描述模型的功耗，本文使用模型剪枝方法对图像描述模型进行压缩剪枝。具体过程如下。

（1）评估神经元的重要程度

根据剪枝粒度的不同，神经元可以定义为一个权重连接，也可以定义为整个特征图。理想情况下，无须对神经元的重要性进行评估，只需要采用暴力方法，逐一对卷积层进行裁剪，并观察裁剪之后损失函数在训练集上的变化，变化最小的即最不重要的特征图，也就是最应该被剪掉的特征图，其目的是使被剪枝的模型的代价函数损失最小，代价函数如式（5）所示，对应的公式相关符号物理意义详见表 1。

$$\min_{W'}\left|C(D|W)-C(D|W)\right| \quad \text{s.t} \quad \left\|W'\right\|_0 \leqslant B \qquad (5)$$

当剪枝后的参数集合无限接近于原网络参数，甚至 $W=W'$ 时（说明裁剪的参数完全没有作用），就认为定义的目标函数达到了最小值，也即完成了模型剪枝的操作。可以发现，Oracle 剪枝算法复杂度极高，以经典网络模型 VGG16 为例，对于总计 4 224 个特征图来说，便需要 $2^{4\,224}$ 次计算，无疑增加了模型训练的时间和难度。

为了解决上述问题，可以使用泰勒级数展开[25]近似损失函数的变化。对于所有的特征图 $h=\{z_0^{(1)},z_0^{(2)},\cdots,z_L^{(C)}\}$ 来说，剪掉某一个特征图 h_i

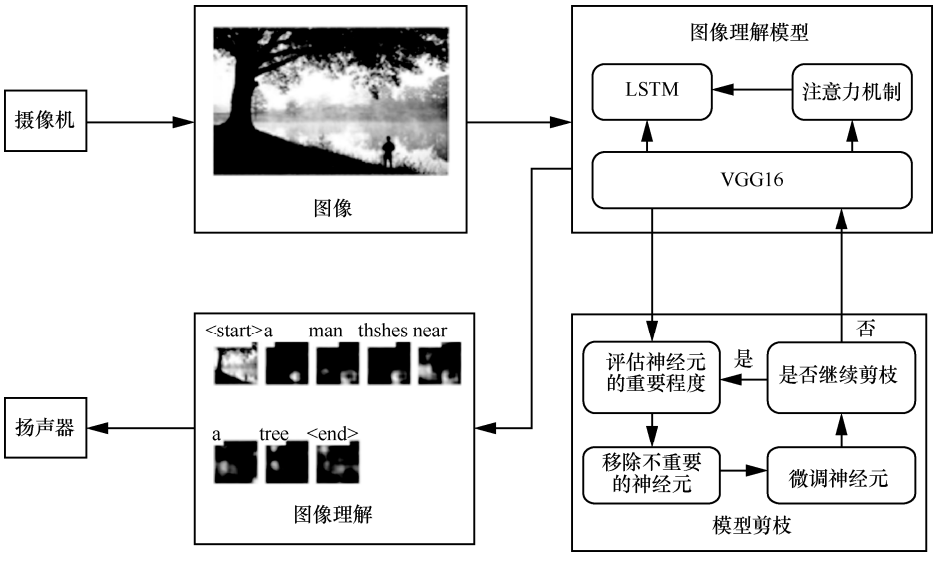


图 3 离线盲人辅助系统模型框架

就是令其等于 0，这时：

$$\left|\Delta C(h_i)\right| = \left|C\left(D \mid W'\right) - C\left(D \mid W\right)\right| = \left|C\left(D, h_i = 0\right) - C\left(D, h_i\right)\right| \tag{6}$$

根据泰勒公式：

$$f(x) = \sum_{p=0}^{P} \frac{f^{(p)}(\mathrm{a})}{p!}(x-a)^p + R_p(x) \tag{7}$$

将式 $C\left(D, h_i = 0\right)$ 对 h_i 展开：

$$C\left(D, h_i = 0\right) = C\left(D, h_i\right) - \frac{\delta C}{\delta h_i} h_i + R_1(h_i = 0) \tag{8}$$

因为拉格朗日余项 $R_1(h_i=0)$的值很小，将其忽略，则判断是否剪枝某一特征图的目标函数变为：

$$\left|\Delta C(h_i)\right| = \left|C\left(D, h_i\right) - \frac{\delta C}{\delta h_i} h - C\left(D, h_i\right)\right| = \left|\frac{\delta C}{\delta h_i} h\right| \tag{9}$$

（2）移除不重要的神经元

神经元的移除可以根据是否满足某个阈值，也可以按照重要程度进行排序。根据第一步的结果，只需设置一个门信号进行移除：

$$g_l = \begin{cases} 0, & \text{剪去某个 特征图} \\ 1, & \text{保留某个 特征图} \end{cases} \tag{10}$$

门信号控制卷积计算输出的结果为：

$$z_l^{(\mathrm{k})} = g_l^{(\mathrm{k})} R\left(\mathrm{z}_{l-1} * w_l^{(\mathrm{k})} + b_l^{(\mathrm{k})}\right) \tag{11}$$

（3）微调神经网络

剪枝类似于一种对完整的网络结构进行有损失调整的操作，势必会对网络模型的精度造成影响。如果剪枝后不进行微调，那么多轮次剪枝后，网络模型的精度将会出现断崖式的下降。因此每次剪枝后需要对模型重新进行训练微调，这在整个流程中至关重要。

（4）重复上述操作，进入下一轮的剪枝。

根据上述算法，剪枝一次需要微调一次神经网络。如果一次只剪掉一个特征图，那么剪枝过程就需要进行多次微调神经网络操作，这无疑增加了训练时间。而训练时间过长会带来许多不便，例如系统在使用过程中往往会为了提高用户的体验感、升级功能、修复存在的漏洞等方面进行版本更新，训练时间过长则会降低系统迭代更新的速度，不能及时满足用户的需求。相应地，一次剪裁掉多个特征图可以大大缩减整个流程的执行次数进而降低第 3 个步骤的执行次数。具体来说，一次裁剪掉 1 个特征图相较于一次裁剪掉 30 个特征图，就需要多进行 30 次训练。但是每次裁剪掉多个特征图，会导致模型精度下降过快，这使得模型压缩基本失去意义。一方面是因为模型的结构一次性改变过大，使得模型难以恢复；另一方面因为裁剪掉的特征图中存在着不该被剪掉的信息。针对以上问题，本文提出一种改进方案实现一次裁剪多个特征图从而减少微调神经网络所需要的时间同时最小化对剪枝后模型的影响。

按照理论，在不改变任何参数和输入的情况下，每次评估时应该会得到同样的结果，然而实验结果并不是这样。当增加评估次数时，会产生不同的剪枝结果同时存在一些重合的特征图，这些重合的特征图在每轮的评估结果中所处的排序位置也不完全相同，这说明某一特征图在某一评估轮次中最应该被剪掉而在其他轮次中有可能不应该被剪掉。

根据上述现象，本文在每轮剪枝中，为了降低不同的评估实验对结果的影响，首先将“评估神经元的重要程度”这一操作执行 5 次，模型的代价函数将变为：

$$\sum_{i=1}^{N} \min_{\mathrm{M} \cdot \mathrm{W}'} \left|C\left(D \mid W'\right) - C\left(D \mid W\right)\right|_i \ \text{s.t.} \left\|W'\right\|_0 \leqslant B \tag{12}$$

其中，i 为执行评估的次数，$N=5$，M 为每次评估后选取的特征图的个数，其他参数含义详见表 1。

在增加评估次数后，对模型剪枝算法进行改进，具体方案如下。

步骤 1　在增加评估次数的基础上，选取重合的特征图，重合次数越多的特征图就越应该被裁剪掉。

表 1　剪枝算法公式相关符号物理意义说明

参数	描述
$D=\{X=\{x_0,x_1,\cdots,x_N\}, Y=\{y_0,y_1,\cdots,y_N\}\}$	输入集合 X 和输出集合 Y 组成的训练集 x_i，y_i 分别表示第 i_{th} 个输入和输出
$W=\{(w_1^1,b_1^1),(w_1^2,b_1^2),\cdots,(w_L^{C_l},b_L^{C_l})\}$	网络模型参数，表示第 i 层的网络模型参数
W'	剪枝后的网络模型参数集合，$W\in W'$
$C(D\mid W)$	预训练模型损失函数
$C(D\mid W')$	剪枝后的模型损失函数
$C(D,h_l)$	定义 h_l 的模型损失函数
B	非 0 参数的个数
$h=\{z_0^{(1)},z_0^{(2)},\cdots,z_L^{(C_l)}\}$	特征图集合
z_l	第 l 层卷积层的特征图
$w_l^{(k)}$	第 l 层卷积层的第 k 个卷积核
g_l	$g_l\in\{0,1\}$

步骤 2　将第一步中选取的特征图的 Oracle-abs 值按从小到大的规则进行排序，裁剪掉排名靠前的特征图。

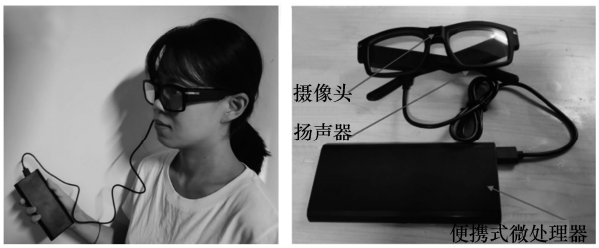
模型此时的代价函数为：

$$\min_{W'}\sum_{i=1}^{N}\min_{M\cdot W'}\left|C(D\mid W')-C(D\mid W)\right|_i \text{ s.t.}\left\|W'\right\|_0\leqslant B \tag{13}$$

在后续的实验中，分别使用只有步骤（1）的改进方案（以下称为改进方案 1）和包含步骤（1）、步骤（2）的完整改进方案（以下称为改进方案 2）对模型压缩的精确度进行验证。

2.3　盲人视觉辅助设备

通过前文描述得到剪枝后的图像描述模型后，本文将其在携式处理器上进行了部署，并最终搭建了完整的盲人离线视觉辅助系统，具体包含：用于拍摄周围场景的广角摄像机、将图像描述用语音转述的扬声器及功能模块、用于图像处理的便携式微处理器设备。其中，广角摄像机和扬声器借助智能眼镜的形式实现，便携式微处理器选取了搭载 Inter 4 核 Z8350 CPU、4 GB 内存、电池容量为 5 000 mAh（约为普通智能手机的电池容量）的便携式计算机。盲人辅助设备使用说明如图 4 所示。当视觉障碍人士处于日常生活环境并需要了解周围环境情况时，可以通过摄像机拍摄周围环境信息，图像描述模型能够对拍摄的图片处理并生成图像描述结果，语音播报功能模组对得到的图像描述信息进行播报。通过实验，视觉障碍人士在拍摄照片后 2 s 左右即可收听到周围环境信息的语音描述，符合实际生活的需求。更为具体的实验结果将在第 3 节给出。



(a) 设备使用佩戴示意图　　(b) 盲人辅助设备细节

图 4　盲人辅助设备使用说明

3　实验分析

3.1　实验条件

（1）实验设置

本文利用阿里云服务器对所采用的图像描述模型进行了剪枝和训练，所采用的数据集为经典的 Flickr8k 数据集[26]。上述数据集中每张图像带有 5 句关于该图像的描述，每一句描述语句都有一个 0～1 的得分，得分越大则语句描述越准确。Flickr8k 数据集示例如图 5 所示。

获得剪枝后的图像描述模型后，本文将其部署在了第 3.3 节所描述的离线盲人系统中，并分别用数据集和实际场景进行了定性和定量的分析。软件环境为 Ubuntu18.04 系统、Python 3.6 和 Pytorch 1.0.0。

（2）评价指标

本文采用双语替换评测（bilingual evaluation understudy，BLEU）[27]和基于召回率的评估指标（recall-oriented understudy for gisting evaluation，ROUGE）[28]对图像描述模型的精度进行评价。

图 5　Flickr8k 数据集示例

BLEU 采用一种 *n*-gram 的匹配原则，即对生成的一句话进行 *n* 个连续单词的截断。根据 *n* 的取值，BLEU 可以划分成多种评价指标，常见的有 BLEU-1、BLEU-2、BLEU-3、BLEU-4。具体来说，BLEU-1 衡量的是单词级别的准确性，而更高阶的 BLEU 可以衡量句子的流畅性，具体的计算方法可以参看文献[27]，本文不赘述。

ROUGE 评估指标是一组能够评估自动文摘以及机器翻译的指标，通过将预测语句和参考语句进行比较得出召回率，以衡量自动生成的语句与参考语句之间的相似度。其中，有 3 个评价标准，分别是 ROUGE-N、ROUGE-L 和 ROUGE-S。ROUGE 和 BLEU 几乎一模一样，区别是 BLEU 只计算准确率，而 ROUGE 只计算召回率，具体的计算方法可以参看文献[28]，本文不赘述。

（3）实验参数

在本文的方案中，执行评估的次数 N=5，每次评估后选取的特征图的个数 M=50，选择一次需要裁剪掉的特征图个数为 50。在改进方案 1 中，当重合的特征图个数不满 50 时，需要扩大执行评估的次数 N。在对模型进行参数微调时，批尺寸为 64、所有训练样本的训练次数为 20、编码学习率为 5×10^{-5}、解码学习率为 5×10^{-6}。

3.2　图像描述模型实验结果和分析

（1）剪枝算法改进前后精度和效果对比

为了验证剪枝前后模型的精度未发生较大变化，本文首先以 BLEU-4 为评估指标分别记录两种改进剪枝算法迭代过程图像描述模型的精度。图 6 表示改进方案 1 和改进方案 2 在剪枝迭代过程中 BLEU-4 的变化情况。其中，原剪枝方案表示经过原始剪枝方法进行模型压缩的图像描述模型。改进方案 1 表示在原剪枝方案的基础上增加评估次数，裁剪重合次数较多的特征图。改进方案 2 表示在第一步增加评估次数的基础上，计算特征图的 Oracle-abs 值并按从小到大的

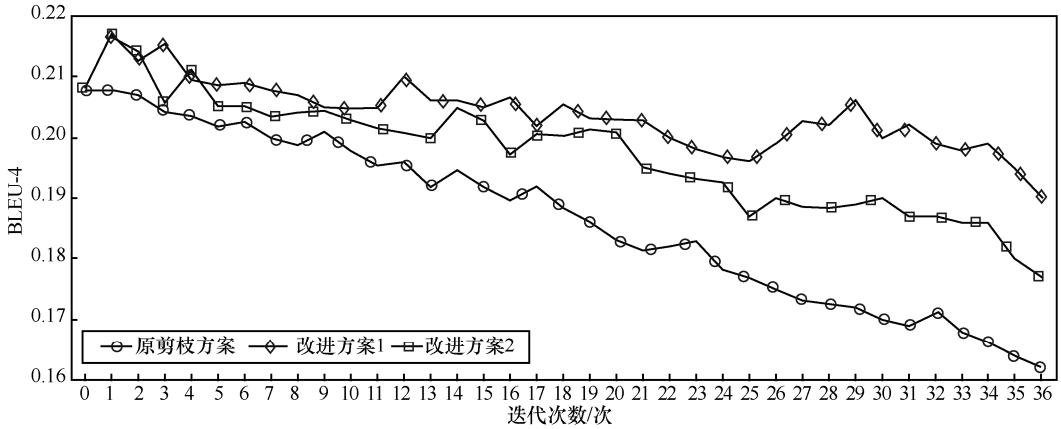


图 6　图像描述模型在原剪枝算法和改进方案上的精度对比

规则进行排序，裁剪掉那些排名靠前的特征图。

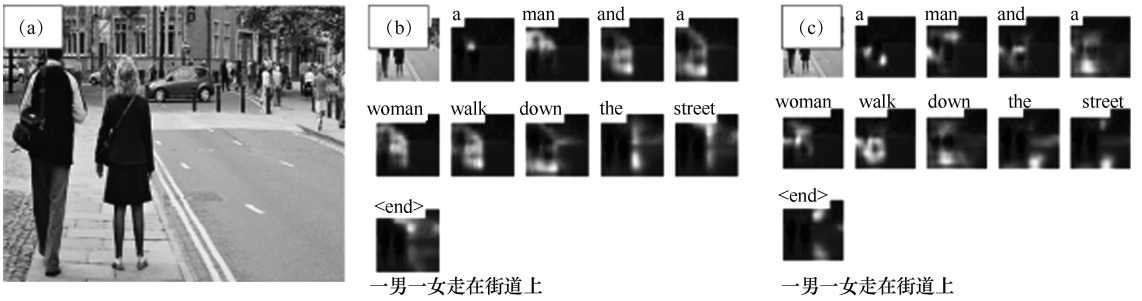
可以看出，随着剪枝迭代次数的增加，原方法剪枝后的模型精度下降越来越快，改进后的两种方案均优于原方法，尤其是改进方案 2。改进方案 1 的整体效果不如改进方案 2，方案 1 在后期（迭代次数为 21 时）呈现速率下降快的趋势，因为重合的特征图在几次评估中排名是随机的，而剪枝后期重合的特征图越来越多且没有对重合的特征图做进一步的处理，单以重合的特征图作为剪枝的依据不再具有代表性。需要注意的是，在迭代次数为 25 时，将裁剪掉的特征图个数改为 30，之后剪枝的效果有明显的提升。同时可以看出在迭代次数大于 34 时，随迭代次数的增加模型精度下降较快，说明最终迭代次数应设置为 34。通过这些实验数据表明，本文提出的改进方案能够有效地改善模型下降速率过快的问题，同时因为将单次剪裁特征图的数量设置为 30，图像描述模型剪枝的训练时间缩减为原算法的 $\frac{1}{30}$ 左右。

为了更加直观地验证剪枝后的图像描述模型仍有较高的精度，本文使用以改进方案二剪枝的图像描述模型分别对数据集中室外和室内场景以及生活中的实际场景进行了实验，如图 7、图 8、图 9 所示。可以明显看出，剪枝后图像描述模型的输出与剪枝前的图像描述模型的输出结果并无区别，这同样印证了图 3 的结论。

（2）剪枝算法改进先后图像描述结果相似度对比

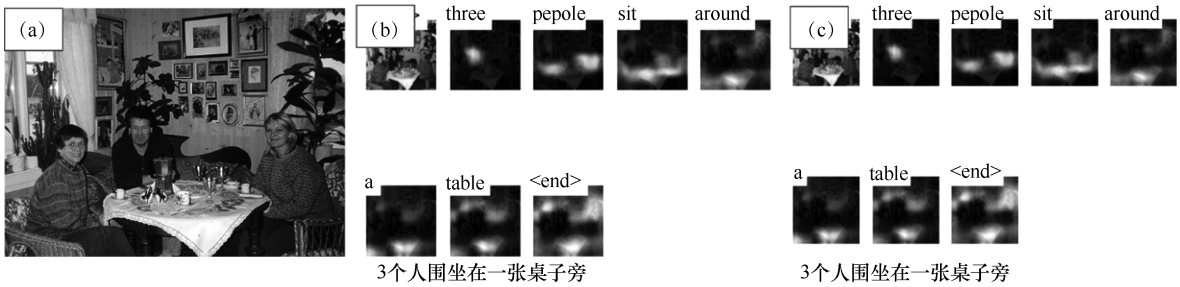
为了验证图像描述模型在本文所提出的剪枝模型的压缩后，模型的准确度没有明显下降，能够提供盲人所需要的周围环境信息，本文随机选取了 100 张图像进行图像描述实验。将未剪枝的图像描述模型的输出作为参考描述，将剪枝后的图像描述输出作为预测描述，采用 ROUGE-1、ROUGE-2、ROUGE-L 评估方式计算剪枝前后的图像描述的召回率，这在一定程度上能够表示剪枝后图像描述模型与原图像描述模型输出的相似度。具体实验结果见表 2，其中，本文将 ROUGE-1、ROUGE-2、ROUGE-L 的结果得分分为 4 个区间：0.91～1、0.71～0.9、0.51～0.7、0～0.5。记录剪枝前后图像描述模型输出结果的召回率得分的比率情况，得分越高表示其输出结果与参考结果越接近。剪枝前后图像描述模型输出结果的召回率情况如图 10 所示。

表 2 表示使用不同评估指标对剪枝前后图像描述模型输出结果评估时不同召回率的占比情况。由图 10 可以看出，剪枝后的图像描述模型与剪枝前图像描述模型的输出中有近 60%的结果召回率大于 0.9，在 ROUGE-1 和 ROUGE-L



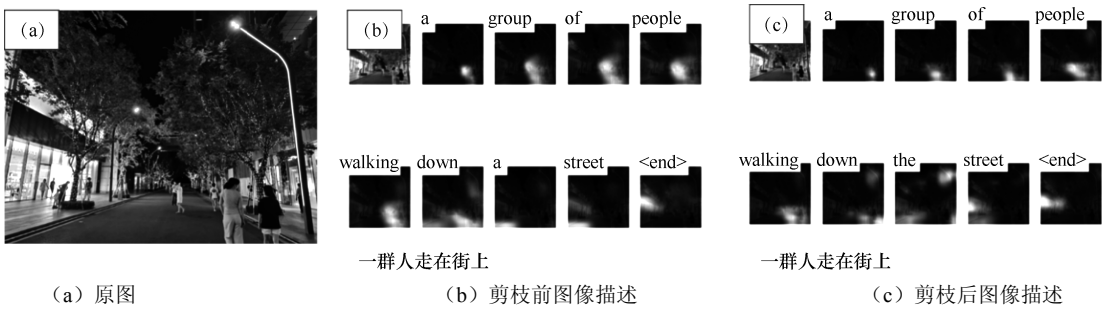
（a）原图　　（b）剪枝前图像描述　　（c）剪枝后图像描述

图 7　模型剪枝前后室外场景图像描述对比



（a）原图　　（b）剪枝前图像描述　　（c）剪枝后图像描述

图 8　模型剪枝前后室内场景图像描述对比



（a）原图　　（b）剪枝前图像描述　　（c）剪枝后图像描述

图 9　模型剪枝前后实际场景图像描述对比

评估指标中有近 90%的结果召回率大于 0.7，在 ROUGE-2 评估指标中也有超过 70%的结果召回率大于 0.7。这说明剪枝后的图像描述模型相比于剪枝前的图像描述模型的精度有降低，但与剪枝前的图像描述相比能够达到 70%以上的相似度，模型精度下降不大。

表 2　剪枝前后图像描述模型输出结果的召回率情况

得分	0.91～1	0.71～0.9	0.51～0.7	0～0.5
ROUGE-1	0.62	0.28	0.08	0.02
ROUGE-2	0.59	0.12	0.22	0.07
ROUGE-L	0.62	0.28	0.08	0.02

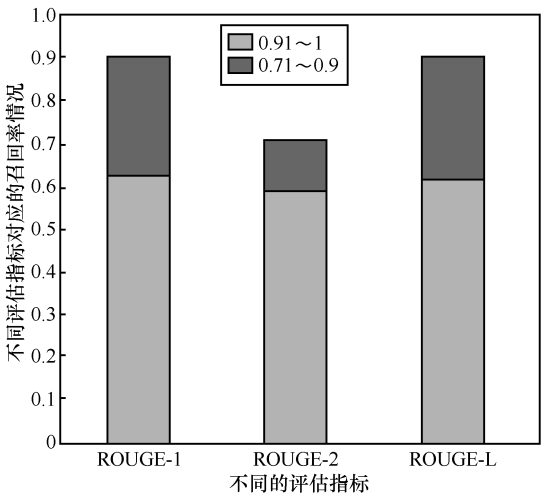
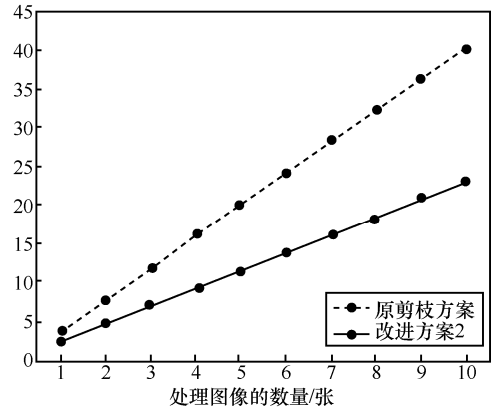


图 10　剪枝前后图像描述模型输出结果的召回率情况

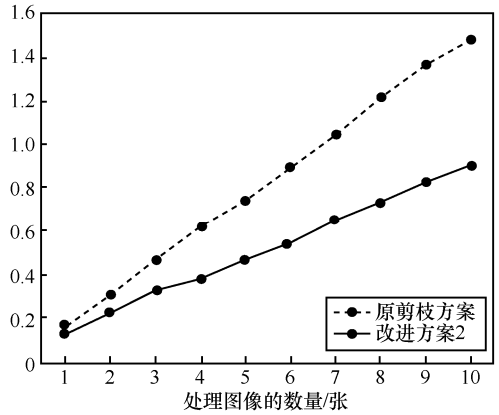
（3）剪枝算法改进先后消耗时间、电源容量对比

为了验证本文所提出的离线盲人视觉辅助系统的高可用性，本文分别测试了剪枝前以及使用改进方案 2 剪枝后的图像描述模型在本文所用硬件上所消耗的时间和电源容量的变化情况。本文在相同的 10 张图片上进行了实验，并累计对应所消耗的时间和功率。

同一组10张图片在剪枝前后的两个模型所累积消耗的时间如图 11（a）所示。无论是单张图片的理解速度还是累计所需要的时间，剪枝后的模型处理速度比剪枝前的速度快，同时随着处理图片数量的增加，其差距也越来越大。根据表 2，剪枝前从图像输入到输出结果单张图像平均用时为 4.049 s；而剪枝后的模型对相同的 10 张图像进行图像描述时，单张图像的平均用时仅为2.337 s，



(a) 剪枝前后图像描述所消耗的时间对比



(b) 剪枝前后图像描述所消耗的电源容量对比

图 11　剪枝前后图像描述对比

缩短了 42%。图像描述用时的缩短能够为视觉障碍人士及时地提供附近环境信息，特别是在危险、紧急的情况下为视觉障碍人群获得宝贵的反应时间。

同一组 10 张图片在剪枝前后的两个模型所累积消耗的电源容量如图 11（b）所示。从图 11 中可以看出，对一张图片进行图像描述时功耗消耗相差不大，但随着处理图像数量的增多，剪枝后的模型相比于剪枝前的模型对电源容量累计消耗增长缓慢，即处理单张图片剪枝后的模型所消耗的电源容量更低，即使处理了 10 张图片后，剪枝后的模型所消耗的功率也只近似于剪枝之前的模型消耗的一半。这是由于剪枝后图像描述所需要处理的数据减少，内存占用率随之减少。根据表 3，剪枝前从图像输入到输出结果每张图片所消耗的处理平均电源容量为 0.269 mAh；而在剪枝后的模型对相同的 10 张图像进行图像描述时，每张图像所消耗的平均电源容量仅为 0.164 mAh。

表 3　剪枝前后图像描述模型所消耗的平均处理时间和平均电源容量功率

模型	处理一张图片平均消耗的时间/s	处理一张图片平均消耗的电源容量/mAh
剪枝前的图像描述模型	4.049	0.164
剪枝后的图像描述模型	2.337	0.269

假定模型剪枝前处理一张图片消耗的电池容量约为 0.16 mAh，剪枝后处理一张图片消耗的电池容量约为 0.27 mAh，将本实验设备用于日常生活可以处理约 30 000 张图片，而在同等条件下剪枝前的图像描述模型只能处理约 18 000 张图片。剪枝前后图像描述所消耗的功率减小对于将此盲人视觉辅助系统装载于低成本便携小巧的移动设备提供了极大的便利，延长了视障人士使用该系统时的时间。

4　结束语

本文提出了基于图像描述模型算法的离线盲

人视觉辅助系统，为了使得图像描述模型能够在便携式低性能移动式设备上离线使用，本文对模型进行了剪枝处理。视觉障碍人士可以利用本文的盲人视觉辅助系统对周围场景拍照作为输入，之后扬声器将图像描述后的信息以语音的形式播报，从而能够感知周围环境的信息。结果表明，剪枝后的模型在图像描述的精度上与剪枝前的模型差别不大，但在处理时间和能耗上分别有较大的降低，这让视觉障碍人士能够长时间稳定及时地感知周围地环境，在一定程度上提升其生活幸福感。后续将进一步开展对现有模型的优化，力求探索出计算机视觉相关模型在实际生活应用的最佳实践模式。

参考文献：

[1] 康帅, 章坚武, 朱尊杰, 等. 改进 YOLOv4 算法的复杂视觉场景行人检测方法[J]. 电信科学, 2021, 37(8): 46-56.
KANG S, ZHANG J W, ZHU Z J, et al. An improved YOLOv4 algorithm for pedestrian detection in complex visual scenes[J]. Telecommunications Science, 2021, 37(8): 46-56.

[2] MAO J H, XU W, YANG Y, et al. Explain images with multimodal recurrent neural networks[EB]. 2014.

[3] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015.

[4] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6077-6086.

[5] LUO Y P, JI J Y, SUN X S, et al. Dual-level collaborative transformer for image captioning[EB]. 2021.

[6] YANG X, TANG K H, ZHANG H W, et al. Auto-encoding scene graphs for image captioning[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 10685-10694.

[7] CHEN S Z, JIN Q, WANG P, et al. Say as you wish: fine-grained control of image caption generation with abstract scene graphs[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9962-9971.

[8] WANG Z Y, FENG B, NARASIMHAN K, et al. Towards unique and informative captioning of images[M]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, [S.l.:s.n.], 2020: 629-644.

[9] XU G H, NIU S C, TAN M K, et al. Towards accurate text-based image captioning with content diversity exploration[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 12637-12646.

[10] DENTON E, ZAREMBA W, BRUNA, et al. Exploiting linear structure within convolutional networks for efficient evaluation[C]//Advances in neural information processing systems. Cambridge:MIT Press, 2014: 1269-1277.

[11] ZHUANG Z W, TAN M K, ZHUANG B H, et al. Discrimination-aware channel pruning for deep neural networks[EB]. 2018.

[12] RASTEGARI M, ORDONEZ V, REDMON J, et al. Xnor-net: imagenet classification using binary convolutional neural networks[C]//European conference on computer vision. Berlin: Springer, 2016: 525-542.

[13] WANG K, LIU Z J, LIN Y J, et al. HAQ: hardware-aware automated quantization with mixed precision[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 8612-8620.

[14] CHEN H T, WANG Y H, XU C, et al. Data-free learning of student networks[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 3514-3522.

[15] LUO L C, SANDLER M, LIN Z, et al. Large-scale generative data-free distillation[EB]. 2020.

[16] YU X Y, LIU T L, WANG X C, et al. On compressing deep models by low rank and sparse decomposition[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 7370-7379.

[17] YANG Z, WANG Y, LIU C, et al. Legonet: efficient convolutional neural networks with lego filters[C]//International Conference on Machine Learning. New York: ACM Press, 2019: 7005-7014.

[18] CHEN H T, WANG Y H, XU C J, et al. AdderNet: do we really need multiplications in deep learning?[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 1468-1477.

[19] XU Y, XU C, CHEN X, et al. Kernel based progressive distillation for adder neural networks[EB]. 2020.

[20] SONG D H, WANG Y H, CHEN H T, et al. AdderSR: towards energy efficient image super-resolution[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 15648-15657.

[21] PARK Y, YUN I D. Fast adaptive RNN Encoder-Decoder for anomaly detection in SMD assembly machine[J]. Sensors (Basel, Switzerland), 2018, 18(10): 3573.

[22] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[EB]. 2015.

[23] XINGJIAN S H I, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Advances in neural information processing systems. Cambridge:MIT Press, 2015: 802-810.

[24] MOLCHANOV P, TYREE S, KARRAS T, et al. Pruning convolutional neural networks for resource efficient inference[EB]. 2016.

[25] 王从徐. 基于泰勒级数展开及其应用探讨[J]. 红河学院学报, 2021, 19(02): 154-156.
WANG C X. Discussion on Taylor series expansion and its application[J]. Journal of Honghe University, 2021, 19(02): 154-156.

[26] HODOSH M, YOUNG P, HOCKENMAIER J. Framing image description as a ranking task: data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2013, 47: 853-899.

[27] 蔡鑫. 基于 Bert 模型的互联网不良信息检测[J]. 电信科学, 2020, 36(11): 121-126.
CAI X. Internet bad information detection based on Bert model[J]. Telecommunications Science, 2020, 36(11): 121-126.

[28] LIN C Y. Rouge: a package for automatic evaluation of summaries[C]//Text summarization branches out. Barcelona: ACL, 2004: 74-81.

[作者简介]

陈悦（1998− ），女，北京科技大学计算机与通信工程学院硕士生，主要研究方向为计算机视觉与人工智能。

郭宇（1992− ），男，博士，北京科技大学计算机与通信工程学院讲师，主要研究方向为无线传感器网络、云计算、多机器人系统。

谢圆琰（1996− ），女，北京科技大学计算机与通信工程学院博士生，主要研究方向为云机器人、服务科学与云计算。

米振强（1983− ），男，博士，北京科技大学计算机与通信工程学院副教授，主要研究方向为服务计算、多机器人系统、移动环境中的点云计算。