Machine Learning feasibility – Zappa Roberto

Image captioning model on smartphone

Image captioning is the process that generate a textual description of an image using an AI model made by one Convolutional Neural Network and one Recurrent Neural Network. In the first part the CNN extrapolate the features from the images while the RNN does language modelling.[1] The main difficulty in running the image captioning model locally is how to determine the hardware requirements needed for it to work properly. Given the limitations imposed on access and use of hardware external to what is available in the kits provided to us it is impossible for me to implement and test image captioning models on several recent terminals. For the feasibility study I'm going to examine the hardware used by third party sources to perform image captioning models like what I hypothesized to draw some considerations.

Model

CoCoPIE’s image captioning model.

CoCoPIE stands for Compression-Compilation co-design for Performance, Intelligence, and Efficiency. It is a software framework for enabling real-time AI on mainstream end devices. CoCoPIE holds numerous records on mobile AI: the first framework that supports all main kinds of DNNs, from CNNs to RNNs, transformer, language models, and so on; the fastest DNN pruning and acceleration framework, up to 180X faster compared with current DNN pruning on other frameworks such as TensorFlow-Lite; making many representative AI applications able to run in real-time on off-the-shelf mobile devices that have been previously regarded possible only with special hardware support.[2]

Reference Video:

* [On-mobile real-time image captioning 1 - YouTube](https://www.youtube.com/watch?v=Pm9M9MXN__g)
* [On-mobile real-time image captioning 2 - YouTube](https://www.youtube.com/watch?v=Tm3xjcW1_jE)
* [On-mobile real-time image captioning 3 - YouTube](https://www.youtube.com/watch?v=m87L6Zysyu8)

Hardware In the reference videos the smartphone that is used is the Samsung Galaxy S10.

**Samsung Galaxy S10**

**Chipset:** Qualcomm Snapdragon 855

**CPU:** Qualcomm Kryo 485 Octa-core

**GPU:** Qualcomm Adreno 640

**AI Engine:** Hexagon 690

Qualcomm Snapdragon 855 is the hight end SoC of late 2018 made by Qualcomm using TMSC 7nm process.

In Figure 1 is possible to see some of most famous flagship SoC made since 2018. As you can see every hight-end Soc younger than the Snapdragon 855 is powerful enough to do the inference of the model in real time.

Figure 1 – Partial list of hight end Soc since 2018 [3]

Even in the Mid-range market nowadays there are SoC more powerful than SD855 as shown in Figure 2.

Figure 2 – Partial list of Mid-range Soc since 2018 [3]

Since in reference video there is no waiting time from taking the picture and receiving the description, we can assume that even SoC that are not as powerful as SD855 could do the inference on the device using of course more time. Otherwise, if the smartphone does not have the minimum requirements, it could be possible to use a third-party cloud service like the Amazon Web Services (AWS) to do the inference. In this scenario using an internet connection the smartphone will send to the servers the photo and receive the description of it.

References Cited

1. Image Captioning in Deep Learning

URL: https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

1. CoCoPIE

URL: <https://arxiv.org/pdf/2003.06700.pdf>

1. AI-Benchmark

URL: https://ai-benchmark.com/ranking\_processors.html