Bachelor's thesis

Information and Communications Technology

2023

Georgi Georgiev

Analyzing the performance of AI summarization with limited resource allocation

TURKU AMK

TURKU UNIVERSITY OF APPLIED SCIENCES

Bachelor’s Thesis | Abstract

Turku University of Applied Sciences

Information and Communications Technology

2023 | 39 pages

Georgi Georgiev

Analyzing the performance of AI summarization with limited resource allocation

The advances in AI text processing in the recent years have been significant. However, the resource demands to achieve this task have also risen considerably. Natural language processing (NLP) is the specific field of machine learning that works with text and AI summarization is a part of NLP. The aim of this thesis was to assess whether the NLP summarization task is possible with limited hardware resources at all and if it was, what its capabilities were.

To achieve that, 11 models in total were prepared for an examination. The selection was made after examining available architectures and the best examples of each were chosen. From those 11 models, 6 were fine-tuned for the summarization task, while the rest were in their base state. While most collected models were kept in their original state, two versions of T5-Small were fine-tuned to examine in more detail the optimizations that can be achieved with it.

The collected models were then evaluated by comparing their hardware utilization and ROUGE scores. Efficiency scores for each model were then calculated based on those values. Finally, the produced summaries were reviewed, possible improvements were proposed, and potential applications were examined.

Keywords:

Artificial Intelligence, Language Model, Machine Learning, Natural Language Processing, Text Summary

Contents

Appendices

Figures

Tables

List of abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CPU	Central Processing Unit
GPU	Graphical Processing Unit
LM	Language Model
ML	Machine Learning
NLP	Natural Language Processing
ROUGE	Recall-Oriented Understudy of Gisting Evaluation
SOTA	State-of-the-art

1 Introduction

There is a vast amount of information stored online and a significant quantity of newly written information that is added to that pre-existing data daily. However, the average person cannot utilize it to its full potential due to our still limited capabilities. Summaries are a great way to extract the most important information out of text. They are present in most research papers but many other information sources, especially long-form ones, lack them. However, even for those that do have summaries, creating one of good quality still requires one to go through the document several times to attain good understanding of it before being able to summarize it.

AI is the option for efficiently automating work and, fortunately, it provides tools that process and analyze text, including summarization. The field that is involved with it is the Natural Language Processing (NLP). The downside of advanced machine learning is that it requires considerable hardware resources to be used to its full potential.

Therefore, the aim of this thesis is to provide a good overview of NLP and test its capabilities while using constrained hardware resources. The testing is focusing on summarization of news articles. Language model options are considered, and the best ones are examined in detail. Python, PyTorch and their respective libraries are the tools used throughout this thesis to train models and produce summaries. The generated summaries are rated based both on the accuracy of the summary and how efficiently were machine resources used during the process. Finally, based on the produced results, the possible applications are considered.

2 Summarization with Natural Language Processing

In this section, an overview of NLP is given and the options for summarization are reviewed. The selected options are explored in more detail.

2.1 NLP Overview

People use language as their day-to-day communication tool, and it is considered as a generally simple concept. However, when one tries to translate it to something a machine can understand, they realize that it is not as simple as it looks. Languages have many abstract concepts that are difficult to put in concrete terms. The natural language processing field attempts to resolve the problems that arise from this translation.

The applications of NLP are numerous and of great value so there has always been active research and development in this field. Machine translation, speech recognition, sentiment analysis and question answering are only among some of the implementations of AI text processing. From the 1950s to now, much progress has been made, moving from hand-coded sets of rules to the current neural networks with unsupervised learning. In the last several years, the NLP field has seen great progress, with the most recent models like ChatGPT and Bing performing close to a human level on many NLP tasks.

In the context of summarization, NLP can be divided into five primary stages of analysis – tokenization, lexical analysis, syntactic analysis, semantic analysis, and pragmatic analysis (de Oliveira et al., 2021). Tokenization, the first step in most types of NLP, is transferring raw text into smaller semantic units that can be properly refined further. It involves removing punctuation, connecting words, entities, and similar parts of a text to avoid possible confusion of algorithms. The lexical analysis involves transferring the multiple variants of one word into its simple form, thus allowing for better discovery of connections between words in a sentence. That aids the following phase, syntactic analysis, which is concerned with that exact issue. The fourth phase tries to clear obscurities caused by structures such as fixed expressions or similar, e.g., ‘hot’ and ‘dog’ have their own meanings, but ‘hot dog’ becomes something different. Finally, the pragmatic analysis compares similarities between sentences to try obtaining a comprehensible summary. Each of those steps can be achieved through many different methods, however, those will not be examined here.

2.2 Summary Types

With the general steps in obtaining a summary from NLP examined, it is time to define what would be a desired output and how that would be achieved. Any good summary should have the following characteristics – it is short and concise, generally a paragraph long; the main ideas of the original text are referred to and no own opinions are added; some copied parts that contain and support the stated ideas. The last characteristic is what separates most NLP summaries. Extractive and abstractive are the two types of summarizations.

In extractive summarizations, the main structure of the corpus is kept the same and the content is just concatenated into the shorter summary, where only the sentences that are deemed most important remain. Some models copy the first several sentences word by word. Overall, this can lead to overly verbose summaries produced by this method. However, it can be argued that the ideas of the original text are better preserved this way.

On the other hand, the abstractive generates a completely new text that summarizes the original in a similar way a human made one would. This method is more advanced as it involves complex sentence generation after already understanding all the semantics. It is usually similar in performance to the extractive method in short texts, but it is usually applied to long form texts like books for example.

Therefore, this thesis prioritizes examining extractive models over abstractive as the latter generally require more training to perform adequately resulting in more resources spent. However, there are still some abstractive models for comparison.

2.3 Evaluation Metrics

Unlike other NLP tasks like translation or question answering, rating summarization can be more challenging. Subjectiveness, relying only on lexical overlap and no evaluation of coherence or fluency are some of the issues with automatic metrics. Nevertheless, they provide a comparable evaluation standard which is the main score for model performance. However, as it is unreliable, a manual comparison is usually made between the different summaries to ascertain quality. This thesis follows the same procedure in the evaluation process.

Recall-Oriented Understudy of Gisting Evaluation (ROUGE) is the main evaluation metric for summarization. It offers a percentage score of overlap between the generated and original summaries. In detail, it combines the values of recall and precision where

$$Recall = \frac{Number\ of\ overlapping\ words}{Total\ number\ of\ words\ in\ reference\ summary}$$

$$Precision = \frac{Number\ of\ overlapping\ words}{Total\ number\ of\ words\ in\ generated\ summary}$$

After the calculation, the values are used to obtain the F1-score which represents the harmonic mean of them and the usual reported final result.

To achieve a good representation, several ROUGE metrics are used, most commonly those being ROUGE-1, ROUGE-2, and ROUGE-L. The number stands for the type of N-gram comparison while the ROUGE-L checks the longest matching sequence of words.

3 Language Models

There are three main categories of NLP models – autoregressive, autoencoding and sequence-to-sequence (Seq2Seq). This categorization is based only on the way a model is pre-trained, in detail, by what method the data has been input in the model. Both autoregressive and autoencoding models can be used with the same architecture (HuggingFace, n.d.). It should be also noted that most recent architectures are built upon the Transformer (Vaswani et al., 2017), employing both the encoder-decoder and attention mechanisms, further lessening the separation. Despite all that, there are still notable differences and this section’s goal is to explore in detail the mentioned model categories, select several examples of each category for testing and provide the reasoning behind the choices.

3.1 Autoregressive models

As language has a natural sequential ordering, approaches that extract a result from conditional probabilities are favoured (Radford et al., 2019). Autoregressive models predict a variable by using a linear combination of the past values of that variable, as seen in Figure 1. For NLP implementation, masks are used to hide everything after the next token so the attention heads can only see the previous text. Additionally, as text length can vary significantly, a “stop token” is necessary to have an indication that the sentence is finished so a full stop needs to be its own token. These models can be tuned to a notable degree as they are following the feed-forward sequence. They can be fine-tuned to many NLP tasks, but their most common application is text generation, which includes summaries.

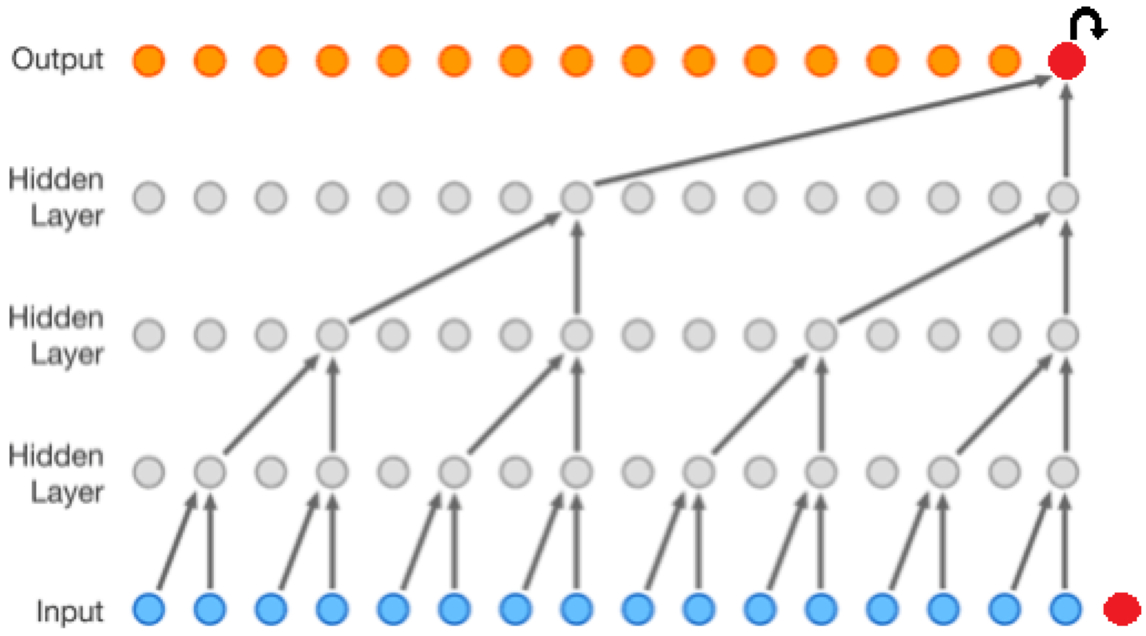

Figure 1. The N_i output generated in a feed-forward fashion. The output is then used in the next generation (Ho, 2019)

GPT-2 (Generative Pretrained Transformer) is one such example. It is an open-source transformer developed by OpenAI in 2019. Thanks to the attention mechanisms it implements, it can detect less occurring but important parts in texts. It produces state of the art results in most NLP tasks and outperforms previous benchmarks of various other architectures in 2019.

As a consequence of this model excelling in text generation, it is best suited to produce abstractive summaries. However, some may argue that it is not as good as other options as it prioritizes predicting the next words and doesn’t try to retain the meaning of the previous. Also, it is a sizeable transformer, with the largest trained model having 1,5 billion parameters and trained on a 40GB of textual data therefore it has higher resource demands. Despite the downsides, it offers unique opportunities as a result of the data – all types of textual data from the web – used in training and the text generation capabilities.

XLNet is another autoregressive model which tries to combine the best of each category in order to produce better results. In detail, it is a generalized autoregressive pretraining method that enables learning bidirectional contexts, combining ideas from both Transformer-XL and BERT (Yang et al., 2019). A significant advantage of autoencoding transformers like BERT over autoregressive ones is the capability to capture meaning bidirectionally, meaning it also understands preceding tokens and not only the following. However, XLNet achieves that by retaining the original positional token encodings which with the proper attention head allow for a permutation of the factorization order.

It is picked for testing because it is a great option for comparison between multiple models as it utilizes their methods and improves upon them to achieve better results.

3.2 Autoencoding models

Autoencoders incorporate representation learning in which a machine is given a large amount of unstructured and unlabelled data, e.g., text, from which patterns and anomalies are extracted. This is achieved by reducing the parts of data that are seen as “noise” by turning high-dimension data into a low-dimension one, as shown in Figure 2, which not only improves accuracy but performance too.

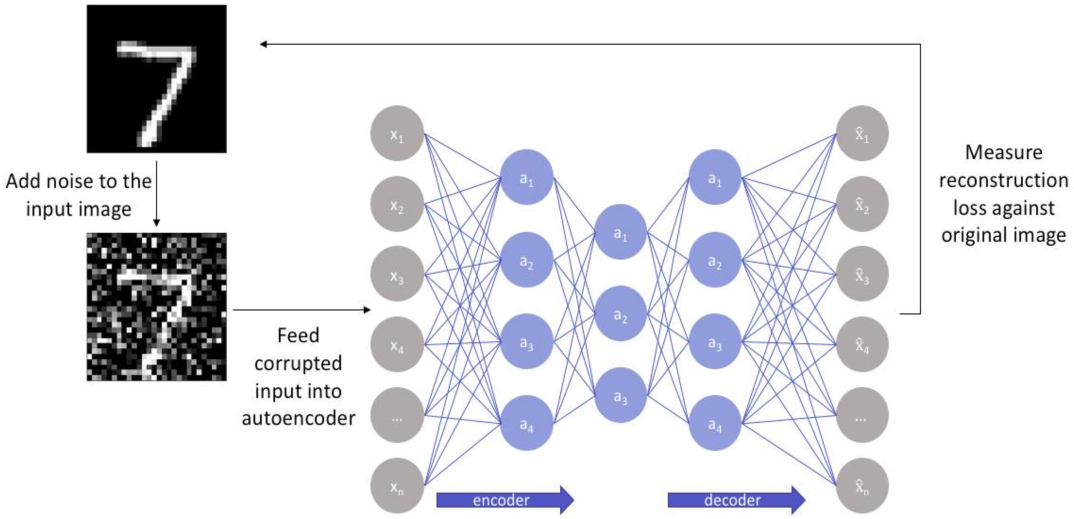

Figure 2. Example of an autoencoder (Bandyopadhyay, 2023)

NLP autoencoders corrupt some of the input tokens and their goal is to reconstruct the original sentence as close as possible, retaining the meaning.

The Bidirectional Encoder Representations from Transformers (BERT) introduced a new practice in language model training that is now being applied in almost all current models. Instead of training a model with a complex architecture for a specific task, a LM is first pre-trained with unlabelled data to obtain a general language understanding. Afterwards, the top level of the model is fine-tuned for a specific NLP task. In the case of BERT, the exact same architecture is used in fine-tuning as it can be seen from Figure 3.

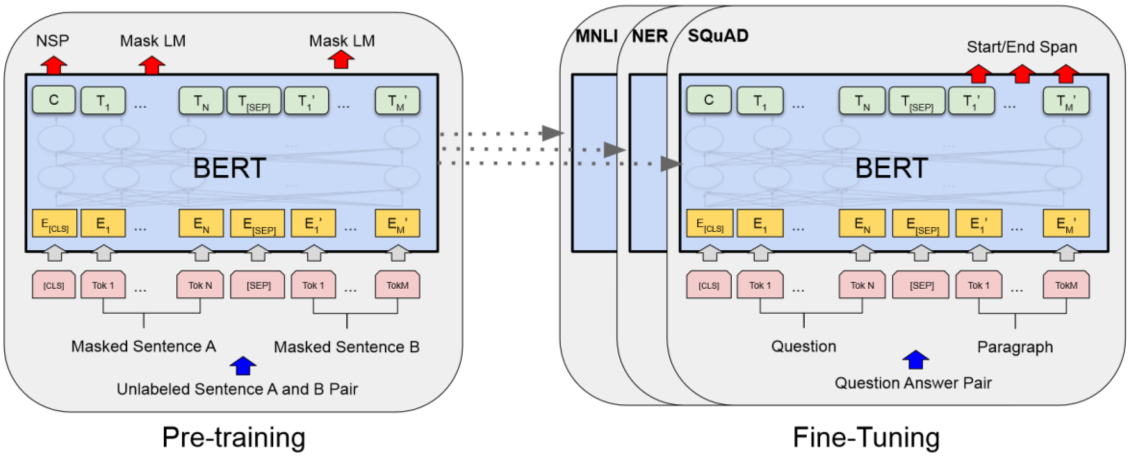

Figure 3. Overall pre-training and fine-tuning procedures for BERT. (Devlin et al., 2019)

To achieve the bidirectional representation, part of the input tokens is masked so the model can’t indirectly see text that it is supposed to predict. The [CLS] and [SEP] a BERT specific tokens serving to separate and identify parts of sentence pairs. This is also used in the fine-tuning process, and it is especially useful in tasks like question answering or summarization where there are “question” and “answer” pairs. Despite being called “sentences”, the pairs can be of arbitrary sequence length and that is why summarization can also fall under this category.

There are numerous implementations of the BERT architecture but as the focus of the thesis will be to test performance, only two will be tested. The first utilizes BERT to extract text embeddings and then identify sentences closest to the centroid for summary selection (Miller, 2019). The second model (Tran, 2020) combines PreSumm (Liu and Lapata, 2019), and mobileBERT (Sun et al., 2020). It performs outstandingly fast and produces satisfactory results.

3.3 Sequence-to-sequence models

Compared to the previous two categories, Seq2Seq models do not alter text semantics in any way. They are fed a text sequence, encode it, and then decode it to produce another text sequence as a result. This allows for better and easier handling of more complex NLP tasks. Most models that produce SOTA results on the summarization task fall under this category.

Text-to-Text Transfer Transformer – T5 – was introduced by Google in 2020. It combines an encoder like BERT and a decoder like GPT-2. Additionally, it was trained on the 700GB C4 dataset, compared to the 40GB one GPT-2 used. Despite it having double the parameters thanks to the encoder-decoder method, it has a similar computational cost (Raffel et al., 2020). Figure 4 displays the method of unifying tasks so the same model and hyperparameters can be used in each.

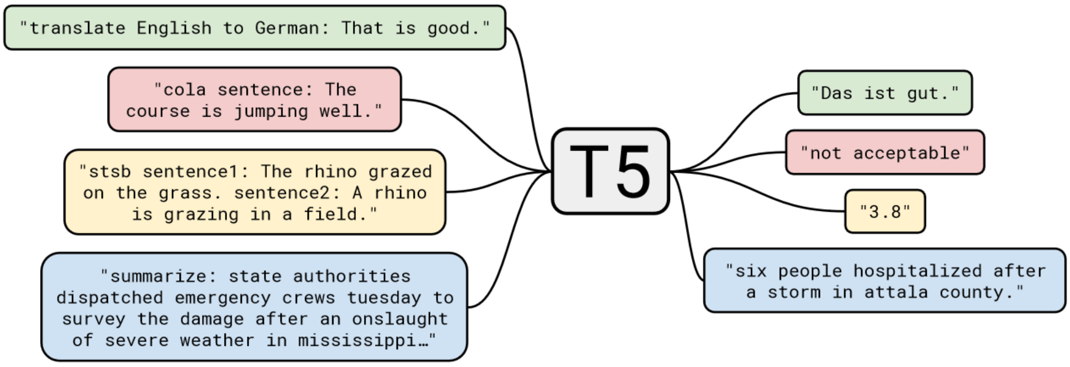
"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 4. Diagram of the T5 framework. (Raffel et al., 2020)

Similar to the goals of this thesis, T5 was trained with the ideas of exploring different existing methodologies in search of the best ones. Furthermore, it has summarization as one of its pre-trained tasks so inference testing can be done easily. As a result, T5 is selected as one of the test examples. A version of T5, T5-small, is fine-tuned on summarization and compared against other T5 models to provide a more in-depth analysis.

BRIO has currently the highest scores in summarization. It modifies two other LMs, BRIO and PEGASUS, which are also among the best performers in summarization. It works similarly to PreSumm (Liu and Lapata, 2019) but instead of picking candidate sentences, candidate full summaries are chosen.

This is the final model selected for testing. BART LMs are also used for comparison as the BRIO model in use implements BART.

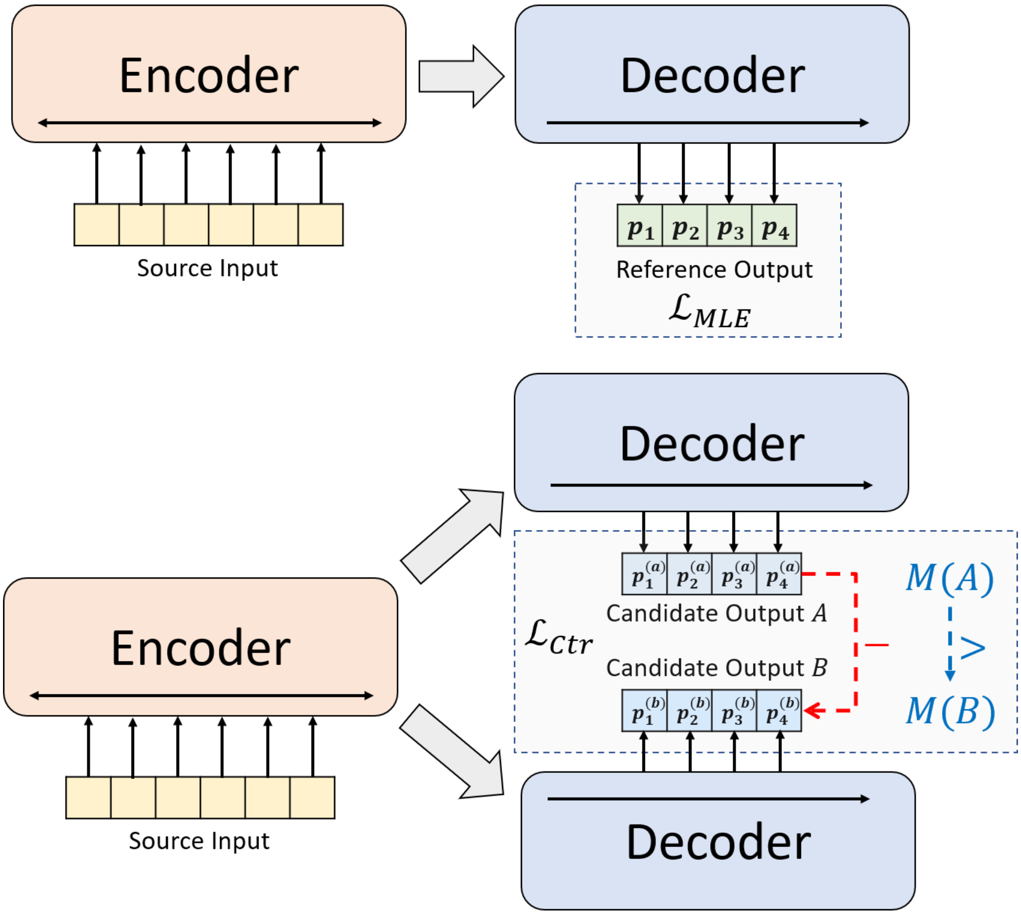

Figure 5. The comparative method BRIO uses to select summaries. (Liu et al., 2022)

4 Model fine-tuning

Fine-tuning is a powerful technique used in machine learning to train a model for a specific task. A pre-trained model is leveraged and only its weights are adjusted. This allows for improvement of models with significantly less resources compared to complete training. This section will demonstrate the fine-tuning process for the T5-small language model.

4.1 HuggingFace

HuggingFace is an American company, founded in 2015, that provides vast number of open-source resources, especially in NLP, that include models, datasets, and courses. Most of the testing and training in this thesis is done thanks to the provided materials by them. Additionally, the HuggingFace Hub makes thousands of community-made language models available freely. This short section is just for clarification as the HuggingFace name will be used often in the following text and also to give proper credit, as this thesis wouldn't be possible without all of the readily available HuggingFace tools.

4.2 Dataset

The *CNN/Daily Mail* dataset was released by See, Liu and Manning (2017). It contains 311 971 unique articles written by journalists at CNN and Daily Mail from 06.2010 to 04.2015. Each article has highlights which can be concatenated to be used for summarization training. It is one of the staple datasets that are used for assessing a model's capabilities for summarization. As a result of its frequent usage and availability on HuggingFace, it will be used for the training so comparison between models is easier.

The dataset is separated into three parts – training, validation, and test datasets. However, the whole set is too large for the available GPU for training. Therefore, only the validation and test datasets will be used. The validation and test dataset splits contain 13,368 and 11,490 articles respectively. The mean token amount is 781 for the articles and 56 for the highlights.

4.3 Training process

This is the relevant hardware of the machine used for fine tuning:

- Intel i7-6700k CPU
- NVIDIA RTX 2080Ti GPU with 11GB memory
- 32GB DDR4 3600MHz RAM

As I do not possess a CUDA-capable GPU, the training process was "outsourced" to the above-mentioned machine. Consequently, only a limited amount of testing and training could be done, and the results displayed below are far from optimal. Despite that, they still serve as a good example and comparison point.

From the data shown in Table 1, it can be seen that the resource demand of larger models might not be always warranted.

Table 1. Reported results on summarization of different T5 implementations (Raffel et al., 2020)

Model	Parameters (millions)	Size	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L
T5-Small	60	240MB	41.12	19.56	38.35
T5-Base	220	870MB	42.05	20.34	39.40
T5-Large	770	2,7GB	42.50	20.68	39.75
T5-3B	3000	10,6GB	42.72	21.02	39.94
T5-11B	11 000	45,2GB	43.52	21.55	40.69

The ROUGE scores of T5-Small and T5-11B only differ by approximately 2 points, while the memory requirements are ~188 times higher. Additionally, with model dimensions varying, loading any additional data for training would make the difference grow exponentially. For example, HuggingFace (n.d.) provides materials where a BERT-Large model (345mil parameters) is used as an example. Training on a test dataset with 512 sequences of length 512 tokens and batch size of just 4 required ~15GB of GPU memory. On the other hand, I trained a T5-Small model with 24 858 sequences also of length 512 with 6,4GB used.

After some testing, only T5-Small was determined to be usable for any reasonable fine-tuning without having “out-of-memory” issues. Two different models were trained with different sized datasets to compare performance and memory usage. As it can be seen for Table 2, the main difference between the two models is the input tokens. Despite T5-Small being only able to process sequences of 512 tokens, raising the max token length to 1024 is a crude method to test memory usage. T5-S-512 used 6,4GB and T5-S-1024 used 9,2GB of memory, without the weights, the used memory will be 6,1GB and 8,9GB respectively. However, there is also doubling of the data size, changing length from 1024 to 512, and increasing batch size from 4 to 8. With so many variables and only two examples, it is impossible to determine the exact relation between all the parameters. To achieve it, an examination of the model architecture and its interaction with PyTorch is necessary, as those vary significantly between the LMs. However, it can be still observed that data length has an enormous influence on memory requirements.

Table 2. Training parameters for the two fine-tuned T5-Small models. Articles were obtained from the “validation” and “test” split of the CNN/DM dataset.

Trained models	T5-S-512	T5-S-1024
Training dataset size	13 368	9 192
Validation dataset size	11 490	2 298
Batch size	8	4
Max input tokens	512	1 024
FP16 precision	Yes	Yes
Training epochs	4	4
Memory used	6,4GB	9,2GB
Size of input dataset	106MB	48MB

The batch size is kept higher for better precision and also divisible by 4 to improve the FP16 performance. The FP16 option allows for a mixed precision, meaning some computations are be performed at half the precision(16-bit) without losing accuracy, resulting in faster training. This, however, requires weights to be stored both in 16-bit and 32-bit values, which increases the GPU requirements of the model by 50% (HuggingFace n.d.). Therefore, this option should be turned off if the only constraint is memory. Additionally, there are further possibilities for optimization, like gradient accumulation and checkpointing, or using Adafactor optimizer instead of Adam. These options can reduce memory requirements by several fold. Nevertheless, they are only of significance when applied on large model training that also involve larger datasets and training times. For smaller sized ones, it will lead to longer training times and possibly worse results, which wouldn’t justify the minimal memory savings (~200MB in this case).

The results in Table 3 are better as expected, with T5-S-512 having a larger dataset that was also properly pre-processed. That is why the training time is faster, despite the larger dataset. In hindsight, the validation set for T5-S-512 was unnecessarily large and more additional could have been added to the training set instead. However, as mentioned, the available training time was too limited to produce an optimally trained model.

Table 3. Training results for the two models

Model	Training time	Loss	ROUGE-1	ROUGE-2	ROUGE-L
T5-S-512	25,3min	1,86	35,59	14,80	24,98
T5-S-1024	33,2min	1,95	34,97	13,92	24,76

5 Evaluation of the models

This section presents and compares the hardware resource utilization, ROUGE scores and actual produced summaries of all the LMs mentioned in the previous sections.

5.1 Setup

All ROUGE scores presented here differ by a large margin from the stated values for the different tested language models. However, that number is consistent across all models and so it isn't because of erroneous testing but most likely due to a different implementation of the ROUGE scoring method, different formatting of the tested text or different size of testing set. Nevertheless, the stated scores in this thesis should only be compared against other values from the document and not with values from other sources to achieve an accurate representation.

This is the relevant hardware of the machine used for the inference, only the CPU was used, and models were loaded into the RAM instead of the GPU memory:

- AMD Ryzen 5 5600X CPU
- 16GB DDR4 3600MHz RAM

To assess my fine-tuned models, 400 articles from the CNN/DM “training” split, which isn't used in the training process, are scored. For the rest, 400 articles from the “test” split are used instead. The “evaluate” module from HuggingFace is the ROUGE implementation that is used. RAM usage is not recorded as it is nothing more than the model being loaded in memory, so the size of the model should be referred to instead. CPU utilization percentage and inference times were measured with “psutil” and “timeit” respectively.

Multiple variations of a LM are tested and also LMs trained on CNN/DM and base ones are compared in an effort to produce the most comparable results.

BERT, GPT2 and XLNet summaries are produced with “bert-extractive-summarizer” (Miller, 2019). It identifies a cluster sentence that is closest to its centroid and combines the sentences to produce a summary.

MobileBERT-PreSumm summaries are made with the code available at https://github.com/chriskhanhtran/bert-extractive-summarization/

T5, BART and BRIO summaries are done solely with the HuggingFace tools.

LEAD3 is used a common baseline for comparisons. It is a summary that just collects the first three sentences of an article.

The CPU and time values include the ROUGE scoring process as there was no simple way to properly separate it and the inference time to evaluate them. Therefore, there are ~0,5s added to the inference times and ~5% added to the CPU average utilization from ROUGE. The results can be seen in Table 4.

‘

Table 4. Hardware usage and ROUGE scores of all tested models

Model name	Size	Average inference time	CPU % utilization (AVG/Peak)	ROUGE-1	ROUGE-2	ROUGE-L
GPT2-Base	1,52GB	5s	28,1 / 63,4	28,49	9,72	18,63
GPT2-Large	3.25GB	9,16s	28.7 / 66,2	28,01	9,54	18,29
XLNet-Base	440MB	2,61s	21.2 / 52,8	29,21	10,28	19,13
BERT-Large	1.35GB	4,62s	27.8 / 60	21,06	10,70	19,42
MobileBERT-PreSumm*	131MB	2,4s	14 / 35,3	32,18	12,19	20,56
T5-Small	242MB	4,34s	34,9 / 58,1	33,28	12,36	22,96
T5-S-512*	236MB	4,38s	33,5 / 59,1	35,59	14,80	24,98
T5-S-1024*	236MB	4,2s	30,5 / 56,9	34,97	13,92	24,76
BART-Base*	558MB	5,11s	28,1 / 57,7	37,93	14,97	25,69
BART-Large*	1,63GB	8,39s	34,3 / 62,3	38,71	16,45	26,38
BRIO*	1,63GB	9,26s	32,4 / 65,6	43,81	19,39	29,38
LEAD3	-	-	-	34,49	14,27	22,85

*Fine-tuned on the CNN/Daily Mail dataset

5.2 Performance values

There is too much varying data in Table 4 to make an accurate comparison. Raffel et al. (2020) observe that ROUGE metrics are highly corelated, which can be also confirmed from Table 4, so they use only the ROUGE-2 values. Therefore, I am going to use it as well to calculate the efficiency of the results.

To achieve that, a simple unification is done. For the CPU values, inference times are multiplied by average CPU utilization % and then ROUGE-2 is divided by the result to obtain CPU efficiency (CE). For memory efficiency (ME), model sizes in GB are divided by ROUGE-2 value. Finally, for an overall efficiency (OE) value, a mean is calculated from the CPU and memory efficiency values. The first entry in Table 4, GPT2-Base, can be used as an example.

$$\frac{9{,}72}{5 * \frac{28{,}1}{100}} = 6{,}9\ CE$$

$$\frac{9{,}72}{1{,}52} = 6{,}4\ ME$$

$$\frac{6{,}9 + 6{,}4}{2} = 6{,}7\ OE$$

The results are arbitrary and serve no other purpose except to generalize the data in Table 4. The idea is that faster models with lower memory usage should have higher efficiency values. All efficiency results are collected in Table 5.

Table 5. Efficiency values of models, ordered by overall efficiency.

Model name	CPU Efficiency	Memory Efficiency	Overall Efficiency	ROUGE-2
MobileBERT-PreSumm	36,3	93	64,7	12,19
T5-S-512	10	62,7	36,4	14,80
T5-S-1024	10,7	58,9	34,8	13,92
T5-Small	8,2	51	29,6	12,36
XLNet-Base	18,6	23,4	21	10,28
BART-Base	10,8	26,8	18,8	14,97
BRIO	6,7	11,9	9,3	19,39
BERT-Large	8,3	7,9	8,1	10,70
BART-Large	5,7	10,1	7,9	16,45
GPT2-Base	6,9	6,4	6,7	9,72
GPT2-Large	3,5	2,9	3.3	9,54

From Table 4 it can be seen that something like LEAD3 can outperform half of the tested models, proving ROUGE isn’t a perfect metric and further evaluation is necessary to determine the best models. Therefore, actual summaries have to be manually reviewed for proper evaluation. The produced summaries are of four models – mobileBERT (best efficiency), T5-S-512 (fine-tuning test), BRIO (best ROUGE score) and GPT2-Large (worst efficiency and ROUGE score). The summaries are the direct result, with no formatting or corrections. There are two articles to check performance with different length. All inferences were constrained to produce summaries of length no more than 4 sentences.

Table 6. Summaries of a CNN/DM article

Model	Summary
Reference Summary	Afghan President Hamid Karzai says his country will help review of war on terrorism. He says U.S. forces shouldn't leave Afghanistan soon since task is not over. Karzai says after civilian causalities, home searches have soured public support. Karzai suggested forces should focus on poppy fields that fuel a drug trade
MobileBert	Obama and Secretary of State Hillary Clinton tapped Holbrooke as special representative for the two countries, a signal of how the new administration considers Afghanistan and Pakistan intertwined in any solution to the war in Afghanistan and the terrorist threat along their shared border. Holbrooke is visiting Afghanistan after a trip to neighboring Pakistan. At the news conference in Kabul, Holbrooke said Sunday that he conveyed the administration 's support of the upcoming elections on August 20 , a date recently set by Afghanistan 's electoral commission .
T5-S-512	Speaking at a joint news conference with visiting U.S. diplomat Richard Holbrooke, Karzai said he is very , very thankful '' that President Barack Obama accepted his proposal to join the review . Holbrooker is visiting Afghanistan after a trip to neighboring Pakistan. Obama and Secretary of State Hillary Clinton tapped Holbrooken as special representative for the two countries.
BRIO	Afghan President Karzai says his country can't afford for U.S. troops to leave soon. Karzai meets with U.S. diplomat Richard Holbrooke in Kabul. Holbroke says he supports Karzai's support of August 20 elections. Karza says he's thankful Obama accepted his request to join review of Afghan policy.
GPT2-Large	Speaking at a joint news conference with visiting U.S. diplomat Richard Holbrooke , Karzai said he is very , very thankful '' that President Barack Obama accepted his proposal to join the review . Holbrooke 's visit comes as Obama plans to send another 30,000 troops to Afghanistan to fight what he 's called the central front in our enduring struggle against terrorism and extremism . '' If we set ourselves the objective of creating some sort of central Asian Valhalla over there , we will lose, '' Gates said during a recent Senate hearing .

Table 7. Summaries of an NPR article

Model	Summary
Reference Summary	Australia's Defense Department will remove surveillance cameras made by Chinese Communist Party-linked companies from its buildings, the government said Thursday after the U.S. and Britain made similar moves.
MobileBert	Hikvision and Dahua are partly owned by China's Communist Party-ruled government. China's Embassy to Australia did not immediately respond to a request for comment. China's general response to such moves is to defend their high tech companies as good corporate citizens who follow all local laws and play no part in government or party intelligence gathering. Security cameras made by Hikvision were also banned from British government buildings in November.
T5-S-512	At least 913 cameras, intercoms, electronic entry systems and video recorders developed and manufactured by Chinese companies are in Australian government and agency offices. The U.S. government said in November it was banning telecommunications and video surveillance equipment from several prominent Chinese brands including Hikvision and Dahua.
BRIO	hikvision and Dahua are partly owned by China's Communist Party-ruled government. At least 913 cameras, intercoms and video recorders are in Australian government offices, a newspaper reports. The U.S. government banned cameras made by the companies in November.
GPT2-Large	The Australian newspaper reported Thursday that at least 913 cameras, intercoms, electronic entry systems and video recorders developed and manufactured by Chinese companies Hikvision and Dahua are in Australian government and agency offices, including the Defense Department and the Department of Foreign Affairs and Trade. The U.S. government said in November it was banning telecommunications and video surveillance equipment from several prominent Chinese brands including Hikvision and Dahua in an effort to protect the nation's communications network. "Where those particular cameras are found, they're going to be removed," Marles told Australian Broadcasting Corp.

6 Discussion

6.1 Resource utilization

As it can be seen from Table 4 and 5, there are drastic differences in resource usage of the different LMs. Part of it is thanks to some models being fine-tuned while others are not. For example, BERT-Large (325M parameters) when compared to MobileBERT (25M parameters) produces 30% worse scores while being 13 times larger and 2 times slower. Even my two fine-tuned models on a miniature dataset, have a 2-point advantage on the ROUGE-2 score over the base T5-Small model. Additionally, it can be seen that all fine-tuned models in Table 7 perform worse when summarizing a non-CNN/DM article, despite them being trained specifically for news summarization and the article itself being shorter. That is because the LMs are highly dependent on the quality of input text and even different formatting can confuse the model. That is why fine-tuned models for their specific task will always perform considerably better than an all-round one.

Another case to be made is that quality matters more than quantity when training new LMs. For example, BRIO is technically a BART model but optimized, which increases its ROUGE scores by more than 3 which is substantial. As it can be seen from the T5 model examples in Table 1, a difference of 3 between ROUGE scores is what separates a 250MB model from a 45GB model, showing that more training data shows diminishing returns compared to adjusting architecture and implementation.

However, the opposite is true for small scale fine-tuning like the one in this thesis. The concepts of underfitting and overfitting are what encapsulates this problem. Not enough training data will result in high loss rate (underfitting) while longer training periods can lead to the model learning irrelevant information from the training data, like noise, making it unable to properly analyse new data (overfitting). I could not go in depth on training optimizations due to limited resources, however, it can be assumed from the observed results that a single

GPU can be enough to optimally fine-tune a small model and produce acceptable results. That is because the capabilities of the GPU used allowed for twice the amount of training data from what was used. Additionally, the training-validation split, and the starting learning rate can be adjusted considerably as the ones used are sub-optimal. Therefore, the already acceptable results produced by the fine-tuned models in this thesis can be further improved.

6.2 Quality of the summaries

There are two main factors that attribute to the observed varying qualities of summaries. First, with article length increasing, 512-length models like T5-S-512 will perform worse as they can only “see” up to that length and the rest of the content is omitted. Fully extractive methods, like the one utilizing GPT2-Large, seem to also deteriorate from long form texts as the extracted sentences are usually from the beginning. The second factor is the type of summaries – abstractive or extractive. The more extractive the summary is, the stronger the baseline quality of it will be. However, there is also small room for improvement as the summaries it gives are basically copying the original text sentences word by word, meaning the results will always be too lengthy and failing to synthesize information adequately. On the other hand, with highly abstractive summarization like BRIO, the results can reach human levels of summarization. The downside is that it lacks the baseline of extractive summaries, and it can often create results that hold entirely wrong information or without any coherence. Abstractive summarization, even when done by humans, can omit important details and despite being technically correct, it can lead to misinterpretation of important information. Therefore, the type of text that is processed should be considered carefully before selecting the summarization type. Shorter and more technical texts would be more suited for extractive summarization to avoid erroneous information while longer form texts, like stories, could be better summarized with high amounts of abstraction.

6.3 Possible applications

The purpose of this examining the performance of summarization models was to determine if it possible for them to be implemented as part of a full-scale application which is ran by a machine without specific hardware. The answer to that is yes, but also very situational. Again, it depends entirely on the task. If it is general news summarization for example, it is probably unnecessary to fine-tune a model for the task considering the vast amount of available trained models online on this. However, if the text sources are expected to only come in specific format, the summaries to be of an exact length or the data is only of one type, e.g., economic news, then a fine-tuned model on a properly prepared dataset to fit those requirements will outperform most if not all general summarization models.

However, despite the possibility to run inferences on a CPU and obtaining quality results, the performance will be severely limited. Even with the most efficient model, MobileBERT, 2second inference and ~40% peak CPU utilizations are considerable constraints, especially if the application is expected to have other CPU heavy features that run parallelly. It should be still noted that this performance is on an only 6-core CPU. Nevertheless, that computation should be transferred to the GPU if possible. Teyssier (2021) observes that there is 25 times difference between inference speeds on BERT when done on a single GPU compared to an 8-core CPU. With a 32-core CPU the difference is 15 times. The cost of GPU inference, however, is 10times higher (Teyssier, 2021). As with every other project, the main bottleneck, be it time, cost, or something else, should be identified and addressed appropriately.

7 Conclusion

The purpose of this thesis was to determine if summaries of good quality could be produced by a machine with limited hardware resources. To achieve that, multiple language models with varying size, architecture, and amount of training were tested. From the generated summaries by the models and their efficiency scores, it can be concluded that the summarization task can be achieved with minimal hardware resources. However, that can only be true if one does not require a solution to a specific task but instead, a more general one, like news summarization. Then the large libraries of pre-trained models on such tasks will be sufficient. However, for summarization on specific tasks, dedicated hardware would be necessary unless training time is not an issue. For reference, the fine-tuning process done for this thesis took 30 minutes on the GPU and it would have taken approximately 600 hours for the same dataset if trained with the available CPU instead. Nevertheless, from the fine-tuning process in this thesis, it can be determined that a single mid- to high-end GPU with CUDA available would be sufficient to produce acceptable results. Adding more resources would possibly only give diminishing returns on the summaries' quality. There are multiple cheaper options to consider for optimization before resorting to hardware upgrades.

Despite the large improvement of summarization models over the last few years that could be even seen in this thesis' comparison of the models from several years ago and the current ones, the results seemed to be still far from perfect. The current language models can produce some impressive summaries, however, there are also many incoherent ones. The inconsistency of the results is still too high to warrant the costly resources of the best performing models, especially for small–scale applications. This is supported by the fact that significantly less hardware expensive models do not perform worse by several folds, which is the difference in the resources they use. With this considered, building a large commercial scale level application might not be a worthwhile endeavour. However, if the purpose is for personal or small team/business use, with minimal investment and correct integration, a possibly good workflow improvement can be achieved.

For anyone with average experience in machine learning and NLP, this document could be seen as too simplistic and often stating obvious facts. Despite that, this thesis still provides a good overview of the AI summarization field and could be used as a solid reference point for anyone who wishes to obtain better performance in this NLP task.

References

Bandyopadhyay, H. (2023) *Autoencoders in Deep Learning: Tutorial & Use Cases*. January 03, 2023 [Blog], accessed January 19, 2023.

de Oliveira, N.R. et al. (2021) Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. *Information*, 12(1):38 [online]. doi: https://doi.org/10.3390/info12010038

Devlin, J. et al. (2019) BERT: Pre-training of deep bidirectional transformers for language under-standing. arXiv preprint https://doi.org/10.18653/v1/N19-1423

Ho, G. Autoregressive Models in Deep Learning – A Brief Survey. March 09, 2019 [Blog], accessed January 19, 2023.

Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain. Association for Computational Linguistics. Available at https://aclanthology.org/W04-1013/

Liu, Y. (2022) BRIO: Bringing order to abstractive summarization. arXiv preprint https://doi.org/10.48550/arXiv.2203.16804

Liu, Y., Lapata, M. (2019) Text Summarization with Pretrained Encoders. arXiv preprint https://doi.org/10.48550/arXiv.1908.08345

Miller, D. (2019) Leveraging BERT for Extractive Text Summarization on Lectures. arXiv preprint https://doi.org/10.48550/arXiv.1906.04165

Radford, A. et al. (2019) *Language models are unsupervised multitask learners* [PDF]. Available at https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf , accessed January 22, 2023.

Raffel, C. et al. (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv preprint https://doi.org/10.48550/arXiv.1910.10683

See, A., Liu, P., Manning, C., (2017) Get To The Point: Summarization with Pointer-Generator Networks. arXiv preprint https://doi.org/10.48550/arXiv.1704.04368

Sun, Z. et al. (2020) MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. arXiv preprint https://doi.org/10.48550/arXiv.2004.02984

Teyssier, V. BERT inference cost/performance analysis CPU vs GPU. April 19, 2021 [Blog] Accessed February 10, 2023.

Yang, Z. et al. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint https://doi.org/10.48550/arXiv.1906.08237

CNN / Daily Mail Dataset Article. See, A., Liu, P., Manning, C., (2017).

KABUL, Afghanistan -LRB- CNN -RRB- -- Afghan President Hamid Karzai said Sunday that his country would join the strategic review of the U.S.-led war on terrorism.

Afghanistan President Hamid Karzai, right, meets with Richard Holbrooke in Kabul on February 15, 2009 .

Speaking at a joint news conference with visiting U.S. diplomat Richard Holbrooke, Karzai said he is `` very, very thankful '' that President Barack Obama accepted his proposal to join the review .

Holbrooke is visiting Afghanistan after a trip to neighboring Pakistan. Obama and Secretary of State Hillary Clinton tapped Holbrooke as special representative for the two countries, a signal of how the new administration considers Afghanistan and Pakistan intertwined in any solution to the war in Afghanistan and the terrorist threat along their shared border .

At the news conference in Kabul, Holbrooke said Sunday that he conveyed the administration 's support of the upcoming elections on August 20 , a date recently set by Afghanistan 's electoral commission .

`` President Obama and Secretary Clinton and the United States government were very gratified to hear President Karzai reaffirm his support of the August 20 decision , '' Holbrooke said .

Holbrooke 's visit comes as Obama plans to send another 30,000 troops to Afghanistan to fight what he’s called the `` central front in our enduring struggle against terrorism and extremism. ''

In an interview on CNN 's `` Fareed Zakaria GPS, '' which aired Sunday , Karzai said that , with a resurgent Taliban , a still-flourishing drug trade and a border with Pakistan believed to be home base for al Qaeda , his country ca n't afford for U.S. troops to leave any time soon .

`` U.S. forces will not be able to leave soon in Afghanistan because the task is not over , '' Karzai said . `` We have to defeat terrorism. We 'll have to enable Afghanistan to stand on its own feet. We 'll have to enable Afghanistan to be able to defend itself and protect for its security ...

`` Then, the United States can leave and, at that time , the Afghan people will give them plenty of flowers and gratitude and send them safely back home . ''

At the same time, Karzai said the actions of troops currently in Afghanistan have turned some of the public against them.

`` It’s the question of civilian causalities. It’s a question of risk of Afghans. It’s the question of home searches, '' he said. `` These activities are seriously undermining the confidence of the Afghan people in the joint struggle we have against terrorism and undermining their hopeful future.

``We 'll continue to be a friend. We 'll continue to be an ally. But Afghanistan deserves respect and a better treatment.''

While he said he welcomes additional U.S. troops, Karzai suggested they need to work along the Afghan-Pakistan border and in the poppy fields that fuel a drug trade that threatens to turn the nation into a narco-state -- not in the villages where most Afghans live.

`` We have traveled many years on. What should have happened early on did n't unfortunately happen, '' Karzai said. `` Now, the country is not in the same mood as it was in 2002. And so any addition of troops must have a purposeful objective that the Afghan people would agree with. ''

The Obama administration is conducting several reviews of U.S. policy in Afghanistan, including a review by Gen. David Petraeus, the commander in the region. Defense Secretary Robert Gates has said the original mission in

Afghanistan was `` too broad '' and needs to be more `` realistic and focused '' for the United States to succeed.

`` If we set ourselves the objective of creating some sort of central Asian Valhalla over there, we will lose, because nobody in the world has that kind of time, patience and money, '' Gates said during a recent Senate hearing.

He called for concrete goals that can be reached in three to five years.

Speaking via satellite from Kabul, Karzai called former President George Bush `` a great person, '' but said he can work with Obama -- despite the president 's comments as a candidate that Karzai had `` not gotten out of the bunker '' to improve security and infrastructure in Afghanistan.

`` President Obama is a great inspiration to the world, '' he said. `` The people of America have proven that they can really be the light holders for change and the will of the people in the world.

`` And his coming to power by the vote of the American people is a manifestation of that great power of the American people. ''

Karzai also acknowledged corruption in the Afghan government but defended the work he's done to combat it.

`` Sure, corruption in the Afghan government is as much there as in any other third world country, '' he said.

`` Suddenly this country got so much money coming from the West, suddenly so many Afghans came from all over the world to participate . Suddenly there were projects -- suddenly there were this poverty that turned into some sort of form of prosperity for this country, '' he said.

He said a government department has been created to deal with corruption and that corrupt judges, administrators and other officials are dismissed `` daily '' over corruption charges.

NPR Article. Link to article (Accessed 13.02.23)

CANBERRA, Australia — Australia's Defense Department will remove surveillance cameras made by Chinese Communist Party-linked companies from its buildings, the government said Thursday after the U.S. and Britain made similar moves.

The Australian newspaper reported Thursday that at least 913 cameras, intercoms, electronic entry systems and video recorders developed and manufactured by Chinese companies Hikvision and Dahua are in Australian government and agency offices, including the Defense Department and the Department of Foreign Affairs and Trade.

Hikvision and Dahua are partly owned by China's Communist Party-ruled government. China's Embassy to Australia did not immediately respond to a request for comment. China's general response to such moves is to defend their high tech companies as good corporate citizens who follow all local laws and play no part in government or party intelligence gathering.

The U.S. government said in November it was banning telecommunications and video surveillance equipment from several prominent Chinese brands including Hikvision and Dahua in an effort to protect the nation's communications network.

Security cameras made by Hikvision were also banned from British government buildings in November.

Defense Minister Richard Marles said his department was assessing all its surveillance technology.

"Where those particular cameras are found, they're going to be removed," Marles told Australian Broadcasting Corp.

"There is an issue here and we're going to deal with it," Marles added.

An audit found that Hikvision and Dahua cameras and security equipment were found in almost every department except the Agriculture Department and the Department of Prime Minister and Cabinet.

The Australian War Memorial and National Disability Insurance Agency have said they would remove the Chinese cameras found at their sites, the ABC reported.

Opposition cybersecurity spokesman James Paterson said he had prompted the audit by asking questions over six months of each federal agency, after the Home Affairs Department was unable to say how many of the cameras, access control systems and intercoms were installed in government buildings.

"We urgently need a plan from the ... government to rip every one of these devices out of Australian government departments and agencies," Paterson said.

Both companies were subject to China's National Intelligence Law which requires them to cooperate with Chinese intelligence agencies, he said.

"We would have no way of knowing if the sensitive information, images and audio collected by these devices are secretly being sent back to China against the interests of Australian