Globalizing BERT-based Transformer Architectures for Long Document Summarization

Quentin Grail
NAVER LABS Europe,
Meylan, France
Univ. Grenoble Alpes, CNRS, LIG,
Grenoble, France
quentin.grail@naverlabs.com

Julien Perez
NAVER LABS Europe,
Meylan, France
julien.perez@naverlabs.com

Eric Gaussier
Univ. Grenoble Alpes, CNRS, LIG,
Grenoble, France
eric.gaussier@imag.fr

Abstract

Fine-tuning a large language model on downstream tasks has become a commonly adopted process in the Natural Language Processing (NLP) (Wang et al., 2019). However, such a process, when associated with the current transformer-based (Vaswani et al., 2017) architectures, shows several limitations when the target task requires to reason with long documents. In this work, we introduce a novel hierarchical propagation layer that spreads information between multiple transformer windows. We adopt a hierarchical approach where the input is divided in multiple blocks independently processed by the scaled dot-attentions and combined between the successive layers. We validate the effectiveness of our approach on three extractive summarization corpora of long scientific papers and news articles. We compare our approach to standard and pre-trained language-model-based summarizers and report state-of-the-art results for long document summarization and comparable results for smaller document summarization.

1 Introduction

Language model pre-training has become a key component to improve performances on a majority of Natural Language Processing (NLP) tasks (Wang et al., 2019). Most of the recent competitive architectures (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019b; Radford et al., 2018) are based on the efficient transformer layer introduced in Vaswani et al. (2017). BERT (Devlin et al., 2019) is one of these architectures that has been widely adopted for comprehension and generation tasks. It is a multi-layer transformer network, pre-trained with different self-supervised objectives. Numerous variations of transformer architectures have been proposed to improve this approach (Lan et al., 2020; Liu et al., 2019b; Radford et al., 2018). However, this type of process is only evaluated on tasks composed of relatively short input text, GLUE (Wang et al., 2019), SQUAD (Rajpurkar et al., 2016), SWAG (Zellers et al., 2018). Indeed, for the tasks that require reasoning with longer documents, this approach exhibits several limitations. The transformer self-attention memory quadratically increases with the number of input tokens, making it technically impossible to compute on document-scale sequences. In addition, they usually require to define a fixed maximum input length, typically of 512 tokens, at the pre-training stage.

One solution is to pre-train the entire model on longer sequences. However, this will still require a massive computation power and will only push the length limitation further. Other alternatives have been proposed to extend multi-layer transformers architectures to longer sequences without modifying this maximum length limitation. The first one is to limit the input sequence to its first tokens by removing the text beyond the length limit. Obviously, it cannot be a reasonable solution to treat long documents that are consistently longer than this limit. The second alternative is to apply the model on a window that slides all over the document. It has been used in Wolf et al. (2019) to deal with SQUAD documents that are longer than the 512 token limitation and in Joshi et al. (2019) for a co-reference resolution task on long documents. This approach can only work if the tokens need to be contextualized only in their surroundings because there is no interaction between the different windows. It seems to be a solution for co-reference resolution (Joshi et al., 2019) as they usually can be solved with a reasonably sized window. Another

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1792–1810
April 19 - 23, 2021.

approach adopted to deal with long documents or multi-document is to select a sub-sample of the input that is small enough for the transformer model. Most of the state-of-the-art pipelines on the multi-hop question answering dataset HotpotQA (Yang et al., 2018) use a first model to retrieve the relevant pieces of text before feeding them to a transformer-based architecture (Fang et al., 2019a; Tu et al., 2019).

We argue that these solutions are not feasible to deal with tasks that require a global understanding of long documents. An example is extractive summarization, where the decision for each sentence should be based on the information of the complete document. To address these challenges, we propose a simple adaptation of the multi-layer transformer architecture that can scale to long documents and benefit from pre-trained parameters with a relatively small length limitation. The general idea is to independently apply a transformer network on small blocks of a text, instead of a long sequence, and to share information among the blocks between two successive layers. To the best of our knowledge, this is the first attempt to introduce hierarchical components directly between the layers of a pre-trained model and not only on top of it (Fang et al., 2019b; Zhang et al., 2019b; Tu et al., 2020). Between each of the transformer layers, we use a Bidirectional Gated Recurrent Unit (BiGRU) network (Cho et al., 2014) to spread global information across the blocks. Adding these propagation layers between the transformer layers preserves the original structure of the pre-trained model and makes it possible to transfer parameter weights from a large pre-trained language model with only few additional parameters to propagate information between blocks.

The contributions of this paper can be summarized as follows: (i) we propose a novel architecture dedicated to long documents which interweaves recurrent hierarchical modules with transformer layers and which exploits pre-trained language models like BERT, and (ii) we demonstrate that this architecture is able to build informative representations in the context of extractive summarization.

2 Global BERT-based Transformer Architecture

In this part, we briefly recall the transformer layer from Vaswani et al. (2017) and its integration in the BERT model (Devlin et al., 2019). Then we describe our modifications of this architecture that allow the model to read longer documents.

Transformers: The transformer architecture, based on a sequence of transformer layers, has been initially introduced in Vaswani et al. (2017). The key idea of this layer is to produce a contextualized representation of an input sequence of tokens. It is composed of the succession of a multi-head self-attention, a first normalizer, a feed-forward neural network, and a second normalizer. This model, which has originally been introduced for machine translation, has then been adopted for most natural language comprehension tasks. Most of the successful approaches (Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2020) are composed of multiple stacked transformer layers. In the remainder, we denote by T^ℓ the transformation corresponding to the ℓ^{th}, $1 \leq \ell \leq L$, transformer layer (T^ℓ is a function from $\mathbb{R}^{N \times h}$ to $\mathbb{R}^{N \times h}$, where N denotes the length of the sequence and h the hidden dimension).

BERT (Devlin et al., 2019) is a multi-layer transformer encoder pre-trained on large text corpora. Two BERT architectures have been proposed in Devlin et al. (2019): $\text{BERT}_{\text{BASE}}$ composed of 12 stacked transformer layers with hidden dimension of 768 ($L = 12, h = 768$) and $\text{BERT}_{\text{LARGE}}$ composed of 24 layers of hidden dimension 1024 ($L = 24, h = 1024$). For both architectures, the input length is limited to 512 WordPiece tokens and the pre-training includes two self-supervised tasks, namely masked language modeling and next sentence prediction. For masked language modeling, 15% of all the WordPiece tokens of the input sequence are masked or corrupted, and the model is used to predict the original token with a cross-entropy loss. For next sentence prediction, the model is trained as a classifier to predict if two sentences are contiguous or not. The pre-training procedure uses the BooksCorpus (Zhu et al., 2015) and documents from English Wikipedia. It requires 4 days of optimization on 16 TPU chips for $\text{BERT}_{\text{BASE}}$ and 64 TPU chips for $\text{BERT}_{\text{LARGE}}$.

2.1 Stacked Propagation Layers

We propose a hierarchical structure that uses pre-trained transformers to encode local text blocks that will be used to compute document level representations. The novel contribution of this work, depicted in Figure 1, is to incorporate recurrent hierarchical modules between the different transformer layers and not only on top of the model, as proposed in

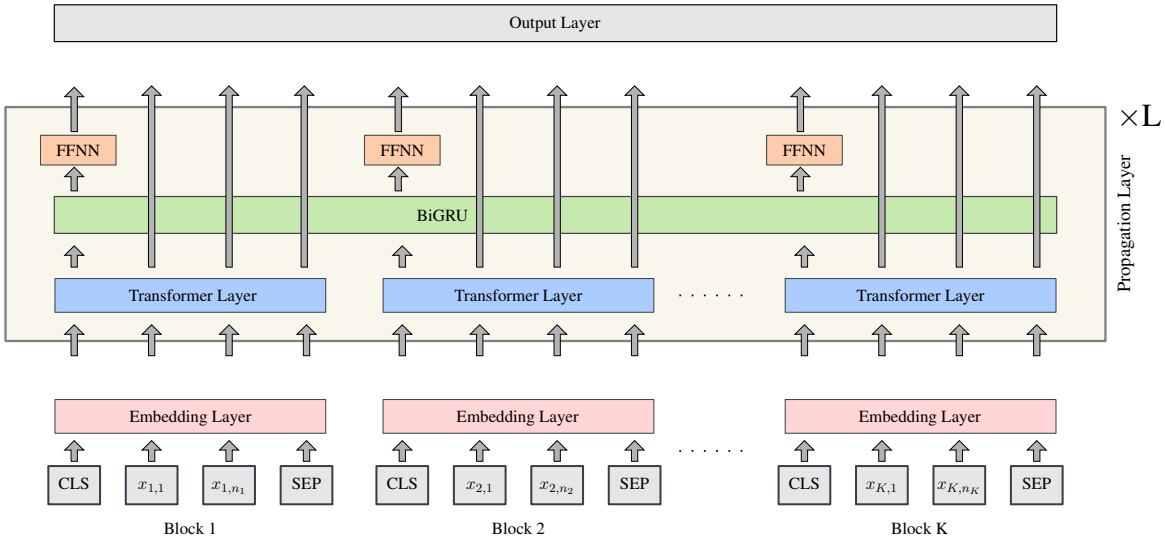

Figure 1: Our proposed modification of a multi-layer transformer architecture. The input sequence is composed of K blocks of tokens. Each transformer layer is applied within the blocks, and a bidirectional GRU network propagates information in the whole document by updating the [CLS] representation of each block.

several recent works (Fang et al., 2019b; Zhang et al., 2019b; Tu et al., 2020). Because we construct and propagate document level information between the layers, global and local information are fused at every level of the architecture. The text blocks can be sentences, paragraphs, or sections. We experiment using sentences as blocks because it generally does not exceed the maximum length allowed by pre-trained models and because BERT has demonstrated to be well adapted to represent such sequences.

We start by splitting the original sequence into multiple blocks. Let D be a document composed of K blocks, $D = \{B_1; B_2; \cdots; B_K\}$ where a block B_k, $1 \leq k \leq K$, is composed of n_k tokens. To follow the convention of BERT, special tokens [CLS] and [SEP] are respectively added at the beginning and end of each block of the document, so that: $B_k = \{[\text{CLS}]; x_{k,1}; x_{k,2}; \cdots x_{k,n_k}; [\text{SEP}]\}$ where $x_{k,i}$ is the index of the WordPiece token i of block k. In the remainder, the index 0 (resp. $n_k + 1$) will be used to refer to the representation of the [CLS] (resp. [SEP]) token in each block.

Embedding Layer Because our goal is to reuse the available pre-trained BERT parameters, token representations are kept the same as in the original BERT and are composed of a token embedding, a segment embedding, and a positional encoding that represents the position of the token in its block. We will denote by E_k ($E_k \in \mathbb{R}^{(n_k+2)\times h}$, $1 \leq k \leq K$) the embedding representation of block k.

Propagation Layers Our model is composed of L stacked identical hierarchical layers, called *propagation layers*, that comprise a transformer layer, a BiGRU to propagate information across blocks and, finally, a feed-forward network. For any layer ℓ, $1 \leq \ell \leq L$, let $U_k^\ell \in \mathbb{R}^{(n_k+2)\times h}$ be the representation of block k after the $(\ell-1)^{th}$ layer, the representation for the first layer being initialized with the output of the embedding layer: $U_k^1 = E_k, \quad \forall k \in \{1, \cdots, K\}$. We first apply the pre-trained transformer function T^ℓ individually on each block of the document to compute local, token-aware representations $V_k^\ell \in \mathbb{R}^{(n_k+2)\times h}$:

$$V_k^\ell = T^\ell(U_k^\ell), \quad \forall k \in \{1, \cdots, K\}.$$

The next step is to propagate information across all the blocks of the document in order to compute a global block-aware representation for the document at layer ℓ, denoted by $W^\ell \in \mathbb{R}^{K\times h}$, $1 \leq k \leq K$. To do so, we use a BiGRU network, fed with the representation vectors of the different blocks, and apply a feed-forward neural network to preserve the hidden dimension of the transformer. Each block k is represented by its [CLS] vector, *i.e.*, the vector (represented by $V_{k,0}^\ell \in \mathbb{R}^h$) at the first position in the local representation of the block. These representations are then concatenated to form the input to the BiGRU. The global, block-aware representation is then computed by applying the feed-forward neural network (FFNN) to all K outputs of the BiGRU:

$$W_k^\ell = \text{FFNN}(\text{BiGRU}_k([V_{1,0}^\ell; \cdots; V_{K,0}^\ell])),$$

Datasets	avg. doc length		avg. summary length	
	sentences	words	sentences	words
arXiv	204	5038	5.6	165
PubMed	88	3235	6.8	205
CNN/DM	32	757	4.1	57

Table 1: Statistics on arXiv, PubMed and CNN/DailyMail validation datasets in terms of documents and summary lengths.

where BiGRU_k denotes the k^{th} output of the BiGRU and $[;]$ is the concatenation operation.

At this stage, we have computed, for a given document, local block representations V_k^ℓ $(1 \leq k \leq K)$ and a global representation W^ℓ. We combine them to build the output representation of the layer:

$$U_k^{\ell+1} = [W_k^\ell; V_{k,1}^\ell; \cdots; V_{k,n_k+1}^\ell], \quad 1 \leq k \leq K.$$

As one can note, $U_k^{\ell+1} \in \mathbb{R}^{(n_k+2)\times h}$ is a representation of block k in which the [CLS] vector representation has been enriched with document level information propagated from other blocks. $U_k^{\ell+1}$ is then used as input for the next propagation layer.

2.2 Output Layer

In this work, we validate our approach on the task of extractive summarization described in Section 3. This task can be considered as a binary classification problem where each block has to be labeled as selected or not. We use a feed-forward neural network followed by a Softmax function on the top of the block level representations after the last layer L to compute $Y \in \mathbb{R}^{K\times 2}$.

$$Y_k = \text{Softmax}(\text{FFNN}(W_k^{L+1})).$$

Using a recurrent architecture to propagate information between blocks has two interesting properties. First, it allows our model to scale to long sequences of blocks without using an attention mechanism that would not scale. Second, it does not require to implement any positional encoding on block representations.

3 Experiments

We evaluate our approach, which we refer to as GBT-EXTSUM (for 'Global BERT-based Transformer for Extractive Summarization'), in the context of extractive summarization, the goal of which being to identify and extract from a document the pieces of text that are the most important (Kupiec et al., 1995). We view this task as a sentence-level classification problem where each sentence has to be labeled according to its belonging to the summary or not. To validate the effectiveness of our approach, we propose to test it on three summarization datasets, namely ArXiv, PubMed and CNN/DailyMail:

- The **ArXiv** and **Pubmed** datasets have been introduced in Cohan et al. (2018). They contain long scientific documents from `arXiv.org` and `PubMed.com` and use their abstracts as the ground-truth summaries. We use the original splits that respectively contain 203,037/6,436/6,440 samples in the training, validation, and test sets for arXiv, and 119,924/6,633/6,658 for PubMed.

- The **CNN/DailyMail** dataset contains news articles associated with short summaries. We use the splits of Hermann et al. (2015), where entities have not been anonymized. This dataset contains 287,226 training samples, 13,368 validation samples, and 11,490 test samples.

Table 1 presents some statistics on these three datasets. As one can note, for the scientific articles, the average number of tokens in the documents to summarize is way beyond the capabilities of a standard transformer pre-trained with BERT.

3.1 Evaluation Metrics

We evaluate the quality of the extracted summaries using the ROUGE metric (Lin, 2004), and more particularly ROUGE-1 (overlap of unigrams), ROUGE-2 (overlap of bigrams), ROUGE-3 (overlap of trigrams) and ROUGE-L (longest common subsequence between the produced summary and the gold-standard one).

3.2 Label Generation

In order to train extractive summarizers, one needs annotations in the form of sentence-level binary labels. To compute such annotations, we follow the work of Kedzie et al. (2018) and label all sentences by greedily optimizing the ROUGE-1 score of the extracted summary against the gold-standard summary associated with each article. These labels are only used at training time, the evaluation of the extracted summaries being done against the gold-standard summaries provided in the datasets.

		PubMed				arXiv			
	Summarizer	RG-1	RG-2	RG-3	RG-L	RG-1	RG-2	RG-3	RG-L
	Oracle	58.15	34.16	24.11	52.99	57.78	30.43	18.41	51.24
	Lead	37.77	13.35	7.64	34.31	35.54	9.50	3.33	31.19
Abstractive or Mix	Attn-Seq2Seq (Nallapati et al., 2016)	31.55	8.52	7.05	27.38	29.30	6.00	1.77	25.56
Abstractive or Mix	Pntr-Gen-Seq2Seq (See et al.)	35.86	10.22	7.60	29.69	32.06	9.04	2.15	25.16
Abstractive or Mix	Discourse summarizer (Cohan et al., 2018)	38.93	15.37	9.97	35.21	35.80	11.05	3.62	31.80
Abstractive or Mix	TLM-I+E (G,M) (Subramanian et al., 2019)	42.13	16.27	8.82	39.21	41.62	14.69	6.16	38.03
Abstractive or Mix	DANCER PEGASUS (Gidiotis and Tsoumakas, 2020)	46.34	19.97	-	42.42	45.01	17.60	-	40.56
Abstractive or Mix	PEGASUS (Zhang et al., 2019a)	45.97	20.15	-	28.25	44.21	16.95	-	25.67
Abstractive or Mix	BIGBIRD-Pegasus (Zaheer et al., 2020)	46.32	20.65	-	42.33	46.63	19.02	-	41.77
Extractive	SumBasic (Vanderwende et al., 2007)	37.15	11.36	5.42	33.43	29.47	6.95	2.36	26.30
Extractive	LexRank (Erkan and Radev, 2004)	39.19	13.89	7.27	34.59	33.85	10.73	4.54	28.99
Extractive	LSA (Steinberger and Jezek, 2004)	33.89	9.93	5.04	29.70	29.91	7.42	3.12	25.67
Extractive	Sent-CLF (Subramanian et al., 2019)	45.01	19.91	**12.13**	41.16	34.01	8.71	2.99	30.41
Extractive	Sent-PTR (Subramanian et al., 2019)	43.30	17.92	10.67	39.47	42.32	15.63	7.49	38.06
Extractive	Bert Ranker (Nogueira and Cho, 2019)	43.67	18.00	10.74	39.22	41.65	13.88	5.92	36.40
Extractive	BERTSUMEXT (Liu and Lapata, 2019b)	41.09	15.51	8.64	36.85	41.24	13.01	5.26	36.10
Extractive	BERTSUMEXT (SW) (Liu and Lapata, 2019b)	45.01	20.00	12.05	40.43	42.93	15.08	6.01	37.22
Extractive	Longformer-Ext (Beltagy et al., 2020)	43.75	17.37	10.18	39.71	45.24	16.88	8.06	40.03
Extractive	Reformer-Ext (Kitaev et al., 2020)	42.32	15.91	9.02	38.26	43.26	14.86	6.66	38.10
Extractive	GBT-EXTSUM (Ours)	**46.87**	**20.19**	12.11	**42.68**	**48.08**	**19.21**	**9.58**	**42.68**

Table 2: Summarization results on PubMed and arXiv. Except for BERT-based approaches, for Reformer-Ext and for Longformer-Ext, which we have reimplemented, the results of the baselines are taken from their associated paper as well as from Cohan et al. (2018). Bold results correspond to the best scores of extractive summarizers.

3.3 Baseline Models

We compare our approach to several well known published methods described below. These methods include SumBasic (Vanderwende et al., 2007), LexRank (Erkan and Radev, 2004), LSA (Steinberger and Jezek, 2004), Attn-Seq2Seq (Nallapati et al., 2016), Pntr-Gen-Seq2Seq (See et al.) and Discourse-aware summarizer (Cohan et al., 2018). The results for these models are the ones reported in the paper (Cohan et al., 2018). We also report the results of Sent-CLF and Sent-PTR, which are hierarchical sentence pointer and classifier, TLM-I+E (G,M) a mixed extractive/generative transformer language model from Subramanian et al. (2019), BIGBIRD (Zaheer et al., 2020), PEGASUS (Zhang et al., 2019a) and DANCER (Gidiotis and Tsoumakas, 2020) which are three abstractive methods. Lastly, we developed several baseline models based on BERT, Longformer (Beltagy et al., 2020) and Reformer (Kitaev et al., 2020):

BERT Ranker: We used a BERT ranker, similar to Nogueira and Cho (2019) in which each sentence of the document is processed individually. We apply BERT on each sentence[1] and use a Sigmoid layer, the input of which consists of the [CLS] representation of the sentence, to model the probability of the sentence to be selected.

BERTSUMEXT has been introduced in Liu and Lapata (2019b). This model is an adaptation of BERT for extractive summarization. Because this model takes as input the concatenation of all the tokens of the document, it cannot scale to the arXiv and PubMed datasets. We propose two variants: the first one is to take as input only the first 800 tokens of the document, as suggested in the original paper. This solution is displayed as BERTSUMEXT in Table 2. The second is to apply BERTSUMEXT per sliding windows on the original document and to use, as a token representation, its representation in the window that maximizes its surrounding context. We name this sliding window implementation BERTSUMEXT (SW) in Table 2.

Longformer-Ext and Reformer-Ext: The Longformer and Reformer models were respectively introduced by Beltagy et al. (2020) and Kitaev et al. (2020). They both propose an adaptation of the Transformer self-attention that scale to long sequences. We add the same classification head as the one used in our model on top of the contextualized representation of the first token of each sentence to label them as selected or not in the summary.

[1]This is possible as no sentence exceeds BERT token limitation.

We also present the Oracle extractive results as an upper bound as well as the Lead baseline (which respectively select the first 3, 6, 7 sentences for CNN/DailyMail, arXiv and PubMed datasets). Several models are reported only on CNN/DailyMail dataset and not on arXiv/Pubmed as they do not scale to long documents.

3.4 Implementation details

We run all our experiments using the Pytorch library (Paszke et al., 2019). We built our model using the "bert-base-uncase"[2] version of BERT and its implementation in the HuggingFace library (Wolf et al., 2019). Our architecture is composed of $L = 12$ propagation layers with a transformer hidden dimension of $h = 768$. The hidden dimension of the BiGRU is set to 384 and we share its parameters among all the propagation layers. The FFNN inside the propagation layers maps the output of the BiGRU of dimension 2×384 to a vector of dimension 768. The FFNN of the output layer is a binary classifier that projects the sentence representations of dimension 768 to an output of dimension 2. We fine-tuned our model on the cross-entropy loss, for 5 epochs on 4 GPUs V100 and use Adam optimizer (Kingma and Ba, 2015) with the initial learning rate set to 3×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, no learning rate warmup and a linear decay of the learning rate. We describe implementation details of BERTSUMEXT, Longformer-Ext and Reformer-Ext baselines in the Supplementary Material, Appendix A.

We used *Trigram Blocking* to avoid the repetition of trigrams in the extracted summaries as suggested in Paulus et al. (2018). Given the extracted summary so far, we only added candidate sentences that had no overlapping trigram with the current summary. We limited the summary to 3 sentences for the CNN/DailyMail dataset, 6 sentences for arXiv, and 7 for PubMed.

3.5 Results

Our main results are shown in Tables 2 and 3. On the arXiv and PubMed datasets, our model outperforms the baseline models on almost all of the reported metrics. Our approach manages to summarize long documents while preserving informativeness (evaluated by ROUGE-1) and fluency (evaluated by ROUGE-L) of the summaries. In addition to the previously published methods, our approach also improves over the BERT-based, Longformer-Ext and Reformer-Ext baselines we have developed. Among them, BERTSUMEXT, which focuses on a truncated version of the document, is the less effective. As documents are significantly longer than the 800 tokens limitation of this model, this result is not surprising. The sliding window adaptation of this model, that allows it to scale to long documents, is the one that achieves results that are the most comparable to ours. Our approach still outperforms this adaptation, demonstrating that summaries require to propagate information beyond a single BERT window.

On the CNN/DailyMail dataset, one can see that our model outperforms all the models that do not use pre-trained parameters. This includes several transformer-based and hierarchical models. However, while having comparable results, we do not achieve stronger performance than the current extractive state of the art from Zhong et al. (2020). This is not surprising as the majority of the CNN/DailyMail examples contains their oracle summary sentences in the first positions of the articles, as shown in the Supplementary Material, Appendix B.

Lastly, we evaluate the impact of several elements of our proposed model in Table 4. We first study the influence of the underlying language model by considering both RoBERTa (Liu et al., 2019b) and PEGASUS (Zhang et al., 2019a) pre-trained models, respectively referred to as GBT-EXTSUM-RoBERTa and GBT-EXTSUM-

Model	R-1	R-2	R-L
Oracle	56.22	33.74	52.19
Lead-3	40.11	17.54	36.32
LATENT (Zhang et al., 2018)	41.05	18.77	37.54
NEUSUM (Zhou et al., 2018)	41.59	19.01	37.98
SUMO (Liu et al., 2019a)	41.00	18.40	37.20
TransformerExt (Liu and Lapata, 2019b)	40.90	18.02	37.17
MASK-LM^{global} (Chang et al., 2019)	41.2	19.1	37.6
PNBERT (Zhong et al., 2019)	42.69	19.60	38.85
BERT-ext + RL (Bae et al., 2019)	42.76	19.87	39.11
HIBERT$_M$ (Zhang et al., 2019b)	42.37	19.95	38.83
BERTSUMEXT (Liu and Lapata, 2019b)	43.25	20.24	39.63
BERTSUMEXT w/o interval embedding	43.20	20.22	39.59
BERTSUMEXT (large)	43.85	20.34	39.90
MatchSum (RoBERTa) (Zhong et al., 2020)	**44.41**	**20.86**	**40.55**
Reformer-Ext (Kitaev et al., 2020)	38.85	16.46	35.16
Longformer-Ext (Beltagy et al., 2020)	43.00	20.20	39.30
GBT-EXTSUM (Ours)	42.93	19.81	39.20

Table 3: Comparison of ROUGE scores on CNN/DailyMail wrt extractive models. All results are taken from original papers but Reformer-Ext and Longformer-Ext which we have reimplemented.

[2] https://github.com/google-research/bert

Model	PubMed			arXiv		
	R-1	R-2	R-L	R-1	R-2	R-L
GBT-EXTSUM	46.87	20.19	42.68	48.08	19.21	42.68
GBT-EXTSUM-RoBERTa	46.02	19.29	41.84	47.42	18.62	42.03
GBT-EXTSUM-PEGASUS	44.11	17.34	40.03	43.50	15.35	38.41
GBT-EXTSUM-NoShare	46.84	20.19	42.63	48.11	19.30	42.75
GBT-EXTSUM-AveragePool	45.24	18.13	40.94	45.71	17.36	40.43
GBT-EXTSUM-Transformer	46.46	19.62	42.17	47.64	18.82	42.22

Table 4: Analysis of the influence of different key components of our proposed architecture.

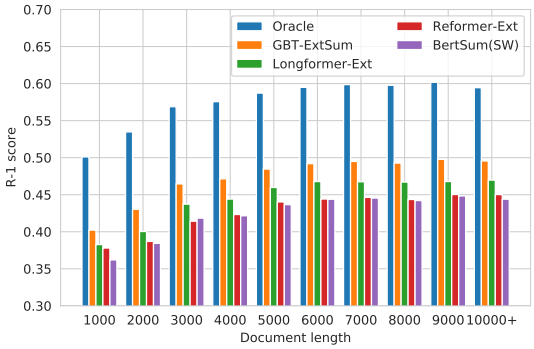

Figure 2: Average R-1 scores of extracted summaries according to the number of words in the input documents from arXiv test dataset.

PEGASUS. As one can see, the results show that BERT-base architecture performs best in terms of ROUGE scores on both arXiv and PubMed. One major difference between PEGASUS and BERT/RoBERTA pre-trained models is that BERT/RoBERTA are only encoders while PEGASUS is a pre-trained encoder/decoder architecture. This could explain why BERT/RoBERTA outperform PEGASUS on extractive summarization tasks. We then compare an alternative of our implementation of GBT-EXTSUM in which the parameters of the BiGRU are not shared among all the propagation layers (GBT-EXTSUM-NoShare) and found no clear difference with the version in which the parameters are shared. Lastly, we compare three architectures of propagation layers, including an average pooling of the [CLS] representations of the sentences, a Transformer layer between the [CLS] tokens (associated to a block position embedding), and a BiGRU layer. Among these three layers, the average pooling layer, which introduces no additional trainable parameters, performs the worst. Furthermore, the BiGRU layer slightly outperforms the Transformer layer in terms of ROUGE scores.

Analysis. In Figure 2, we compare the R-1 score of several models regarding the number of words in the source documents. One can see that GBT-EXTSUM consistently outperforms BERTSUMEXT (SW), Reformer-Ext and Longformer-Ext regardless of the number of words in the source documents.

We present in Table 5 two example summaries of a document from the PubMed test set (Kamio et al., 2009), respectively obtained by GBT-EXTSUM and BERTSUMEXT (SW). The numbers in the margin indicate the position of the sentences in the original document, which is composed of a total of 78 sentences. As one can observe, GBT-EXTSUM extracts sentences from various parts of the document whereas BERTSUMEXT (SW) mostly focuses on the beginning of the document. Among the sentences selected by the two models, the most meaningful one, in terms of ROUGE, is the last one selected by GBT-EXTSUM. This sentence appears at position 66, in the last section (*Discussion*) of the original paper. In contrast, BERTSUMEXT (SW) proposes sentences that are less relevant for summarization purposes. Additional summaries of the PubMed and arXiv articles are provided in the Supplementary Material, Appendices C and D.

To analyse the influence of the positions of the sentences in the input document, we present in Figure 3 the histograms of the positions of the sentences of the Oracle summary as well as that of the predicted positions of different models, on the PubMed test set. One can see that if most relevant sentences appear at the beginning of a document, other Oracle sentences are still relevant further down the document. GBT-EXTSUM is the model that behaves the most closely to the Oracle, followed by BERTSUMEXT (SW), Reformer-Ext and Longformer-Ext. These last two models tend to over-select sentences from the beginning while focusing less on the ones appearing later in the document. Our model remains influenced by the sentence position but is still able to select sentences from all over the document and is closer to the Oracle distribution.

4 Related Work

Hierarchical neural architectures have been competitive on a collection of NLP tasks that require to reason over long or multiple documents such as aspect-based sentiment analysis (Paulus et al., 2018), document summarization (Cheng and Lapata, 2016), document segmentation (Koshorek

GOLD		purpose : to investigate whether the glc3a locus harboring the cyp1b1 gene is associated with normal tension glaucoma (ntg) in japanese patients.materials and methods : one hundred forty two japanese patients with ntg and 101 japanese healthy controls were recruited . patients exhibiting a comparatively early onset were selected as this suggests that genetic factors may show stronger involvement . genotyping and assessment of allelic diversity was performed on 13 highly polymorphic microsatellite markers in and around the glc3a locus.results:there were decreased frequencies of the 444 allele of d2s0416i and the 258 allele of d2s0425i in cases compared to controls (p = 0.022 and p = 0.034 , respectively) . however , this statistical significance disappeared when corrected (pc $>$ 0.05) . we did not find any significant association between the remaining 11 microsatellite markers , including d2s177 , which may be associated with cyp1b1 , and ntg (p $>$ 0.05). conclusions : our study showed no association between the glca3 locus and ntg , suggesting that the cyp1b1 gene , which is reportedly involved in a range of glaucoma phenotypes , may not be an associated factor in the pathogenesis of ntg .
GBT-EXTSUM	1-	primary open angle glaucoma (poag) is the most common type of glaucoma .
	15-	we excluded individuals who were diagnosed under 20 or over 60 years of age and who had 8.0 d or higher myopic refractive error of spherical equivalence .
	17-	the cases exhibiting a comparatively early onset were selected as they suggest that genetic factors may show stronger involvement . during diagnosis ,
	30-	the probability of association was corrected by the bonferroni inequality method , ie , by multiplying the obtained p values with the number of alleles compared .
	63-	only two adjacent markers , d2s0416i and d2s0425i , were significantly positive , as shown in table 2 , and the frequency of the 444 allele of d2s0416i and the 258 allele of d2s0425i were decreased in cases compared to controls (p = 0.022 , or = 0.59 and p = 0.034 , or = 0.42 , respectively) .
	66-	the purpose of this study was to investigate whether the glc3a locus is associated with ntg in japanese subjects , based on results from recent studies reporting that the cyp1b1 gene , located at the glc3a locus on chromosome 2p21 , could be a causative gene in poag as well as pcg . to this end , we genotyped 13 microsatellite markers in and around the glc3a locus . here
BERTSUMEXT (SW)	1-	primary open angle glaucoma (poag) is the most common type of glaucoma .
	2-	normal tension glaucoma (ntg) is an important subset of poag ; while many poag patients have high iop,1 patients with ntg have statistically normal iop.24 the prevalence of ntg is higher among the japanese population than among caucasians , and recent studies reported that 92% of poag patients in japan had ntg.58 the diagnosis of glaucoma is based on a combination of factors including optic nerve damage and specific field defects for which iop is the only treatable risk factor .
	7-	of these subjects , 142 were diagnosed with ntg , and 101 were control subjects .
	20-	genomic dna was extracted using the qiaamp dna blood mini kit (qiagen , hilden , germany) or the guanidine method . in this association study , we selected 13 highly polymorphic microsatellite markers that are located in and around the glc3a locus as shown in figure 1 .
	28-	the number of microsatellite repeats was estimated automatically using the genescan 672 software (applied biosystems) by the local southern method with a size marker of gs500 tamra (applied biosystems) .
	22-	polymerase chain reaction (pcr) was performed in a reaction mixture with a total volume of 12.5 l containing pcr buffer , genomic dna , 0.2 mm dinucleotide triphosphates (dntps) , 0.5 m primers , and 0.35 u taq polymerase .

Table 5: An example of summary produced by our method compared to the gold summary and one produced by BERTSUMEXT (SW). With a red scale, we highlight the sentences with the highest ROUGE score when evaluated against the abstract. We show in the margin the position of the extracted sentence in the document. This document (Kamio et al., 2009) is 78 sentences long.

et al., 2018) and text classification (Yang et al., 2016). The hierarchical structure enables the model to learn local contextualized token representations in its lower hierarchy level, while higher-level representations can capture long-distance dependencies within the document. Liu and Lapata (2019a) have proposed a hierarchical modification of the transformer layer-based attention modules to model relations between documents for abstractive summarization but do not investigate parameter transfer form pre-trained language models. Chang et al. (2019) and Zhang et al. (2019b) suggested pre-training processes for hierarchical models, without however testing their approaches on long document summarization nor releasing their pre-trained models. We have not included these models in our comparison for this reason. Transformer-XH (Zhao et al., 2020) introduced an eXtra Hop attention to model dependencies between different transformer windows but requires a graph of related documents.

Long-Document Transformers: Multiple studies have investigated different self-attention mechanisms to extend transformers to long documents. Transformer-XL (Dai et al., 2019) introduced a recurrence between successive transformer windows which run from left to right through the document, preventing global information to bidirectionally flow through the document. Other approaches design the self-attention as a sparse layer, as sparse transformers (Child et al., 2019) or the recently proposed Longformer and BIGBIRD models (Beltagy et al., 2020; Zaheer et al., 2020). One major difference with our work is that these models compute the attention only between a limited set of randomly or *a priori* chosen tokens. Reformer (Kitaev et al., 2020) also tackles the problem of language modeling for long sequences, but it does so by computing the self-attention only between *similar*

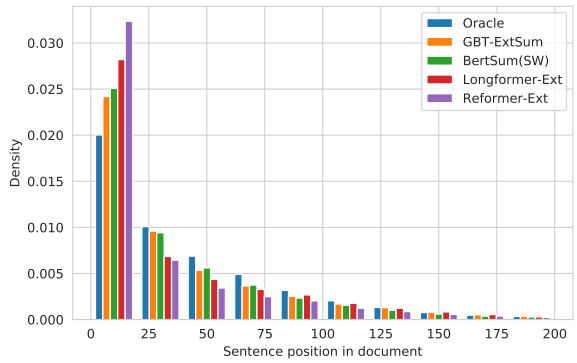

Figure 3: Proportion of the extracted sentences according to their position in the input document from PubMed test dataset.

tokens, based on locality-sensitive hashing.

5 Conclusion

In this paper, we have introduced a novel transformer-based model for long document summarization based on propagation layers that spread information between multiple transformer windows. This model preserves the architecture of commonly used pre-trained language models, thus allowing the transfer of parameters. An evaluation, conducted on top of the BERT model in the context of an extractive summarization task, further revealed its effectiveness in dealing with long documents compared to other adaptations of BERT and previously proposed models. In the future, we plan to adapt our model to other tasks that require understanding long documents, as question-answering and document-scale machine translation.

Acknowledgments

This work was partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003).

References

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sang-goo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. *CoRR*, abs/1909.08752.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Ming-Wei Chang, Kristina Toutanova, Kenton Lee, and Jacob Devlin. 2019. Language model pre-training for hierarchical document representations. *CoRR*, abs/1901.09128.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2019a. Hierarchical graph network for multi-hop question answering. *CoRR*, abs/1911.03631.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2019b. Hierarchical graph network for multi-hop question answering. *CoRR*, abs/1911.03631.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pages 1693–1701. Curran Associates, Inc.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5802–5807. Association for Computational Linguistics.

M Kamio, Akira Meguro, Masao Ota, N Nomura, Kenji Kashiwagi, F Mabuchi, Hiroyuki Iijima, K Kawase, T Yamamoto, M Nakamura, Akira Negi, T Sagara, Teruo Nishida, M Inatani, Hidenobu Tanihara, M Aihara, M Araie, Takeo Fukuchi, H Abe, and Nakamura Mizuki. 2009. Investigation of the association between the glc3a locus and normal tension glaucoma in japanese patients by microsatellite analysis. *Clinical ophthalmology (Auckland, N.Z.)*, 3:183–8.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, page 68–73, New York, NY, USA. Association for Computing Machinery.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5070–5081. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019a. Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8024–8035.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation.

Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher J. Pal. 2019. On extractive and abstractive neural document summarization with transformer language models. *CoRR*, abs/1909.03186.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2019. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. *CoRR*, abs/1911.00484.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. *national conference on artificial intelligence*.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manag.*, 43(6):1606–1618.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1049–1058. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

A Baselines: Implementation Details

BERTSUMEXT: For all experiments with BERTSUMEXT, we started with the original implementation[3] and adapted the code to build the sliding windows version. This implementation leverage *bert-base-uncased* pre-trained model and its associated hyperparameters. We use windows of width 800 with an overlap of 300 tokens between two following windows. If a sentence is in multiple windows, we select its [CLS] representation in the window that maximizes the number of surrounding tokens. We finetune the model for 5 epochs using Adam optimizer with an initial learning rate of 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Longformer-Ext: We built the Longformer-Ext baseline from the Longformer implementation released by HuggingFace[4]. We use the official *longformer-base-4096* pre-trained model trained by AllenAI[5]. This model is based on *RoBERTa-base* and its associated hyperparameters. To increase the maximal position embedding, we drop the pre-trained positional embedding parameters and train a novel token embedding layer to scale Longformer-Ext input up to 12294 tokens. This model computes a sliding self-attention with a window size of 512 tokens on all its 12 Transformer layers. We finetune the model for 5 epochs with only local attention because of memory constraints, using Adam optimizer with an initial learning rate of 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, no learning rate warmup and a linear decay of the learning rate.

Reformer-Ext: We started from the HuggingFace implementation of Reformer to build Reformer-Ext baseline. We use a Reformer configuration composed of six layers of attention. We use Locality-Sensitive Hashing Attention with 128 buckets on the input sequence and Local Self-attention on chunks of 64 tokens. We use hidden sates of dimension 256, a feed-forward layer of dimension 512, and 12 attention heads in Transformer encoders. We train this model for 5 epochs using Adam optimizer with an initial learning rate of 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, no learning rate warmup and a linear decay of the learning rate.

[3] https://github.com/nlpyang/PreSumm

[4] https://github.com/huggingface/transformers

[5] https://github.com/allenai/longformer

B Datasets Statistics

Figure 4 presents the distribution of the document lengths in arXiv, PubMed and CNN/DailyMail, after tokenization with pretrained BERT-base tokenizer. It also provides the histograms of the position of the [CLS] tokens of the Oracle sentences in input documents. One can see that the three datasets contain an important number of documents longer than 512 tokens, the standard length limitation of pre-trained language models. However, one can also notice that CNN/DailyMail contains a large part of its Oracle sentences within this first window of 512 tokens. As a consequence, a model that is not able to "read" beyond this limitation is not penalized. It is also a reason why Lead baseline is quite strong on this dataset. On the contrary, on arXiv and PubMed, one can see that a large part of Oracle sentences occur beyond this 512 windows. This explains why models capable of reading long sequences are required to achieve good results on these datasets.

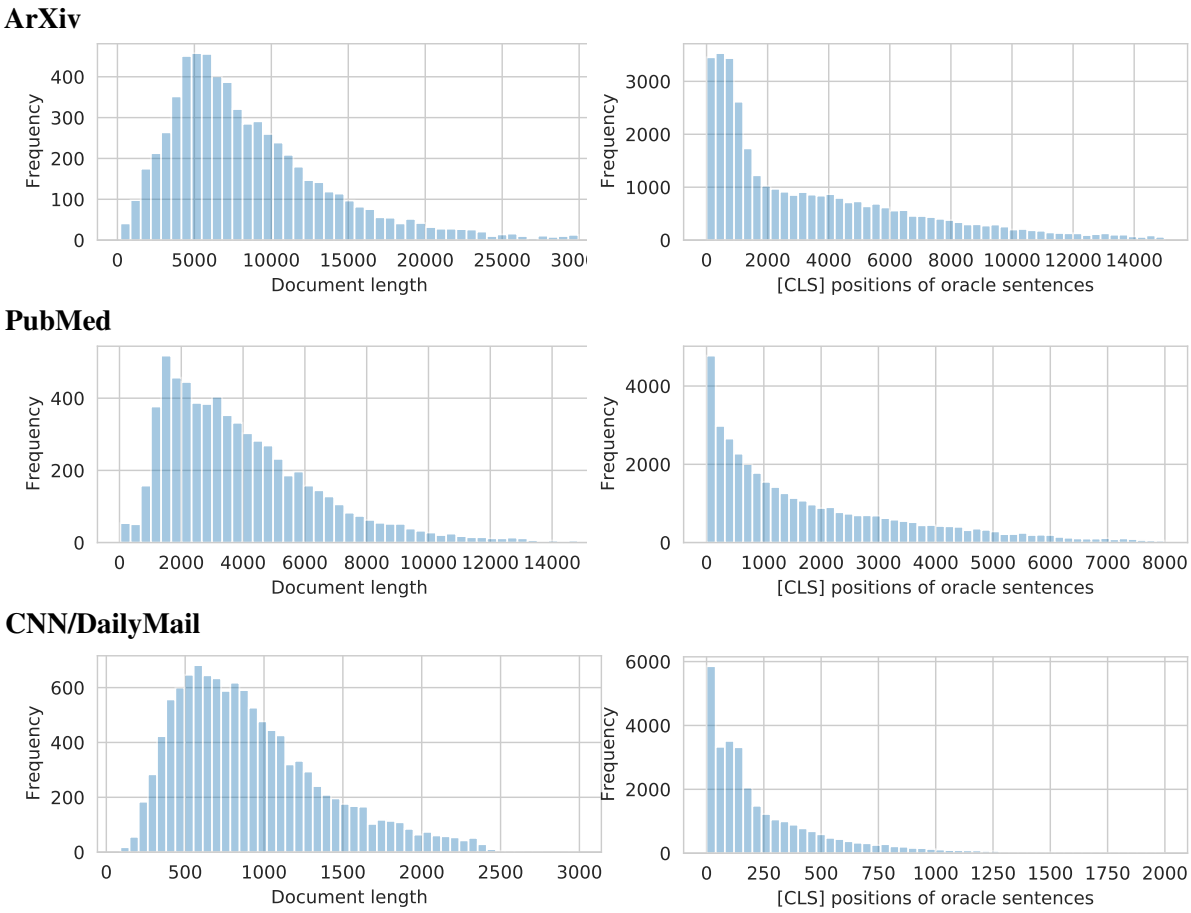

Figure 4: Document lengths after tokenization with pretrained BERT-base tokenizer and position of the [CLS] tokens of Oracle sentences in the input documents.

C PubMed Summaries

GOLD		aim . to investigate incidental adrenal enlargement clinical characteristics and functional status and analyze functional lesion risk factors . materials and methods . this retrospective study included 578 patients with adrenal imaging features showing enlargement . incidental adrenal enlargement cases (78) were considered eligible . demographics , functional diagnosis , adrenal imaging features , and concomitant diseases were analyzed . results . the number of adrenal enlargements and proportion of incidental adrenal enlargement increased each year . mean patient age was 50.32 years . thirty - nine cases had unilateral enlargement on the left side and 3 on the right side ; 36 had bilateral enlargement . routine medical checkup was found to have the greatest chance (43.59%) of revealing clinical onsets leading to discovery . biochemical and functional evaluation revealed 54 (69.23%) cases of nonfunctional lesions , 12 (15.38%) of subclinical cushing syndrome , 6 (7.69%) of primary hyperaldosteronism , 1 (1.28%) of metastasis , and 5 (6.41%) of unknown functional status . nodular adrenal enlargement (or , 7.306 ; 95% ci , 1.72728.667 ; p = 0.006) was a risk factor for functional lesions . age and lesion location were not significant factors . conclusion . incidental adrenal enlargement is a frequent radiographic finding and is accompanied by diverse clinical factors that require proper evaluation and management . nodular adrenal enlargement was a risk factor .
GBT-EXTSUM	8-	data retrieved included patient demographics , final functional diagnosis , adrenal imaging features , and concomitant diseases .
	14-	smooth enlargement was defined as enlargement of the gland with a smooth contour and no measureable or diffuse nodules . after obtaining patient history and physical examination , all patients underwent biochemical evaluation to assess their functional status .
	16-	patients with an aldosterone - rennin ratio (arr) $>$ 20 underwent any 1 of 3 confirmatory tests (saline infusion , captopril challenge , or postural stimulation) to confirm or exclude definitively primary hyperaldosteronism (pa) .
	25-	as shown in table 1 , routine medical checkup was found to have the greatest chance (43.59%) of revealing clinical onsets leading to the discovery of adrenal enlargement .
	29-	nodular adrenal enlargement (or 7.306 ; 95% ci , 1.72728.667 ; p = 0.006) was the risk factor for functional lesions .
	31-	our study shows that the proportion of incidental adrenal enlargement has gradually increased by year .
	46-	acth - independent macronodular hyperplasia (aimah) and primary pigmented nodular adrenal hyperplasia often manifest as adrenal hyperplasia . the clinical features of aimah tended to be atypical .
BERTSUMEXT (SW)	4-	it is a common term for a variety of adrenal disorders , but its cause must be properly assessed so that patients needing treatment , such as those with hormone hypersecretion or malignant disease , can receive appropriate care . however , there is a lack of literature on functional status and its follow - up to provide comprehensive insight to these findings .
	5-	patients with incidental adrenal enlargement were evaluated in a tertiary referral hospital with endocrinological departments in china .
	7-	this retrospective study included 578 patients with adrenal imaging features showing adrenal enlargement who were hospitalized at the department of endocrinology in pla general hospital (beijing , china) between january 1993 and july 2013 .
	29-	nodular adrenal enlargement (or 7.306 ; 95% ci , 1.72728.667 ; p = 0.006) was the risk factor for functional lesions .
	36-	in addition , smooth enlargement was more common , in 53 (83%) cases , and together these statistics reflect the likelihood that adrenal enlargement will be bilateral , smooth , and found in men .
	37-	however , our study did not show this tendency , likely because the research goals and thus , study populations , differed between the 2 studies .
	38-	's study aimed to explore prevalence , while the present study aimed to evaluate functional status .

GOLD		background and objective . antimicrobial resistance is now a major challenge to clinicians for treating patients . hence , this short term study was undertaken to detect the incidence of multidrug - resistant (mdr) , extensively drug - resistant (xdr) , and pandrug - resistant (pdr) bacterial isolates in a tertiary care hospital . material and methods . the clinical samples were cultured and bacterial strains were identified in the department of microbiology . the antibiotic susceptibility profile of different bacterial isolates was studied to detect mdr , xdr , and pdr bacteria . results . the antibiotic susceptibility profile of 1060 bacterial strains was studied . 393 (37.1%) bacterial strains were mdr , 146 (13.8%) strains were xdr , and no pdr was isolated . all (100%) gram negative bacterial strains were sensitive to colistin whereas all (100%) gram positive bacterial strains were sensitive to vancomycin . conclusion . close monitoring of mdr , xdr , or even pdr must be done by all clinical microbiology laboratories to implement effective measures to reduce the menace of antimicrobial resistance .
GBT-EXTSUM	5-	multidrug resistant (mdr) was defined as acquired nonsusceptibility to at least one agent in three or more antimicrobial categories . extensively drug
	36-	no mdr or xdr strain was isolated from streptococcus sp . all (100%) gram positive cocci were sensitive to vancomycin and linezolid .
	38-	e. coli was the commonest isolate 261 (35%) , followed by pseudomonas aeruginosa 212 (28.4%) .
	40-	out of 200 klebsiella pneumoniae strains isolated , 75 (37.5%) and 25 (12.5%) were detected as mdr and xdr , respectively . out of 42 acinetobacter and other nonfermenter species isolated , 19 (45.2%) and 8 (19%) were mdr and xdr strains , respectively . amongst 250 gnb - mdr strains isolated ,
	62-	, it has been reported that most frequent mdr pathogens were pseudomonas aeruginosa followed by e. coli .
	67-	unless and until multidrug resistant organisms are detected and their incidence is known , the strategies for their control can not be adopted properly in healthcare setup . hence , detection , prevention of transmission of mdros by following infection control practices , antimicrobial surveillance , and stewardship are need of the hour .
	69-	we hereby conclude that early detection and close monitoring of mdr , xdr , or even pdr bacterial strains must be started by all clinical microbiology laboratories to reduce the menace of antimicrobial resistance which is now a global problem .
BERTSUMEXT (SW)	9-	this short term cross - sectional study was conducted in the department of microbiology from 15th of april to 15th of july , 2014 .
	10-	the bacterial strains were isolated from different clinical samples and were identified by conventional methods .
	17-	methicillin resistant staphylococcus aureus (mrsa) strains were detected by meca - mediated oxacillin resistance using cefoxitin disk (30 g) on mueller hinton (mh) agar plate inoculated with test strains as per standard disk diffusion recommendations and incubated at 3335c for 1618 hours .
	20-	an increase in diameter of 5 mm with ceftazidime plus clavulanic acid as compared to ceftazidime disk alone was considered positive for esbl detection .
	36-	no mdr or xdr strain was isolated from streptococcus sp . all (100%) gram positive cocci were sensitive to vancomycin and linezolid .
	38-	e. coli was the commonest isolate 261 (35%) , followed by pseudomonas aeruginosa 212 (28.4%) .
	65-	the limitation of this study is that this is a single center study for only three - month period in a tertiary care hospital in central india . to reflect the trend of infections caused by mdr and xdr strains of bacteria in the region , a multicenter study involving all types of healthcare setups for a minimum period of one year

GOLD		background suicide is a grave public health issue that is responsible for a high mortality rate among individuals aged 1544 years . attitudes toward suicide among medical staff members have been associated with appropriate therapeutic responses to suicidal individuals . the aim of this study was to examine the effects of parental rearing on attitudes toward suicide among japanese medical college students.methodswe examined the association between parental bonding and attitudes toward suicide in 160 medical college students in japan . the parental bonding instrument was used to assess the attitudes and behaviors of parents . the attitudes toward suicide were evaluated using the japanese version of the attitudes toward suicide questionnaire.resultsthe mean age of the subjects was 25.24.0 years old . the majority of the participants in our study agreed that anyone could commit suicide (88.8%) and that suicide is preventable (86.3%) . after adjusting for age and sex , multivariate regression analysis revealed that maternal care approached a statistically significant association with the right to suicide attitude . under the same conditions , maternal care was shown to be significantly associated with the common occurrence attitude . no other significant relationships were observed between parental bonding and attitudes toward suicide.conclusionthis study suggests that a higher level of maternal care ensures that children think that suicide occurs less commonly . the promotion of best practices for suicide prevention among medical students is needed . child rearing support might be associated with suicide prevention .
GBT-EXTSUM	3-	previous studies have shown that difficulties with parental bonding during childhood could be a predisposing factor for the onset of many psychiatric conditions , such as anxiety , depressive states , and maladjusted behaviors.68 parental bonding and premorbid personality traits play an important role in shaping the developmental trajectory of an individual , including his / her ability to adjust to stressful events .
	5-	the objective of this study was to investigate whether parental bonding is associated with attitudes toward suicide among medical college students in japan .
	8-	the demographic data (age and sex) were obtained from self - questionnaires and interviews .
	14-	higher scores on the care and protection dimensions reveal that participants perceive their parents to be more caring and/or protective .
	39-	right to suicide was significantly associated with common occurrence , unjustified behavior , and preventability / readiness to help .
	43-	the majority of the participants in our study agreed that anyone could commit suicide (88.8%) and that suicide is preventable (86.3%) .
	44-	in addition , the multiple regression analysis revealed that participants who reported a higher level of maternal care thought that suicide was a common occurrence and tended to think that people do not have the right to commit suicide .
BERTSUMEXT (SW)	6-	students in their fifth year of medical school at hirosaki university , hirosaki , japan , participated in the study .
	7-	the surveys were distributed to 226 medical students . of the distributed 226 surveys , 160 questionnaires (116 males and 44 females)
	13-	the overprotection dimension of the pbi reflects parental overprotection and control in contrast to the encouragement of autonomy .
	14-	higher scores on the care and protection dimensions reveal that participants perceive their parents to be more caring and/or protective .
	15-	we employed the japanese version of the attitudes toward suicide questionnaire (atts) to assess the attitudes toward suicide held by the study participants.12 we employed a six factor model that was previously developed in studies of japanese attitudes , including
	16-	common occurrence , suicidal expression as mere threat , unjustified behavior ,
	17-	impulsiveness.12,13 each item , with the exception of items 10 and 28 , was scored on a five point scale from 1 (strongly agree) to 5 (strongly disagree) .

D ArXiv Summaries

GOLD		in vivo calcium imaging through microscopes has enabled deep brain imaging of previously inaccessible neuronal populations within the brains of freely moving subjects . however , microendoscopic data suffer from high levels of background fluorescence as well as an increased potential for overlapping neuronal signals . previous methods fail in identifying neurons and demixing their temporal activity because the cellular signals are often submerged in the large fluctuating background . here we develop an efficient method to extract cellular signals with minimal influence from the background . we model the background with two realistic components : (1) one models the constant baseline and slow trends of each pixel , and (2) the other models the fast fluctuations from out - of - focus signals and is therefore constrained to have low spatial - frequency structure . this decomposition avoids cellular signals being absorbed into the background term . after subtracting the background approximated with this model , we use constrained nonnegative matrix factorization (cnmf , @xcite) to better demix neural signals and get their denoised and deconvolved temporal activity . we validate our method on simulated and experimental data , where it shows fast , reliable , and high quality signal extraction under a wide variety of imaging parameters .
GBT-EXTSUM	1-	. continued advances in optical imaging technology are greatly expanding the number and depth of neuronal populations that can be visualized .
	2-	specifically , in vivo calcium imaging through microendoscopic lenses and the development of miniaturized microscopes have enabled deep brain imaging of previously inaccessible neuronal populations of freely moving mice (@xcite) . while these techniques have been widely used by neuroscientists ,
	20-	like the proposed cnmf in @xcite , our extended cnmf for microendoscopic data (cnmf - e) also has the capability of identifying neurons with low signal - to - noise ratio (snr) and simultaneously denoising , deconvolving and demixing large - scale microendoscopic data . to accomplish this : (1) we replace the rank-1 nmf approximation of the background with a more sophisticated approximation , which can better account the complex background and avoid absorbing cellular signals , and (2) we develop an efficient initialization procedure to extract neural activities with minimal influence from the background .
	71-	@xmath56 is a template matching filter to detect spatial structures with similar shapes and sizes . for flat structures in the small regions , like background , filtering them with @xmath56
	134-	in this paper , we proposed an efficient method for extracting cellular signals from microendoscopic data ; such methods are in very high demand in the neuroscience community .
	136-	our method shows credible performances in recovering the real neuronal signals and outperforms the previous standard pca - ica method .
BERTSUMEXT (SW)	0-	monitoring the activity of large - scale neuronal ensembles during complex behavioral states is fundamental to neuroscience research
	11-	our work is based on a matrix factorization approach , which can simultaneously segment cells and estimate changes in fluorescence in the temporal domain .
	26-	the video data we have are observations from the optical field for a total number of @xmath2 frames .
	64-	we estimate the temporal component of one neuron @xmath15 from spatially filtered data and then use it to extract the corresponding spatial footprint @xmath14 from the raw data . in the step of estimating @xmath14 , we re - order all frames to make nearby frames share the similar local background levels and then take the temporal differencing to remove the background signals temporally .
	105-	we also display @xmath98 tightly clustered neurons in the simulated data (figure [fig : sim]e) to demonstrate that our cnmf - e approach can accurately detect and demix their activity (figure [fig : sim]g) .
	107-	in contrast , pca - ica based detection can only detect two neurons and the calcium traces have high level of noise .

GOLD		statistical learning theory chiefly studies restricted hypothesis classes , particularly those with finite vapnik - chervonenkis (vc) dimension . the fundamental quantity of interest is the sample complexity : the number of samples required to learn to a specified level of accuracy . here we consider learning over the set of all computable labeling functions . since the vc - dimension is infinite and a priori (uniform) bounds on the number of samples are impossible , we let the learning algorithm decide when it has seen sufficient samples to have learned . we first show that learning in this setting is indeed possible , and develop a learning algorithm . we then show , however , that bounding sample complexity independently of the distribution is impossible . notably , this impossibility is entirely due to the requirement that the learning algorithm be computable , and not due to the statistical nature of the problem .
GBT-EXTSUM	6-	an alternative approach , and one we follow in this paper , is simply to consider a single learning model that includes all possible classification methods .
	8-	since the vc - dimension is clearly infinite , there are no uniform bounds (independent of the distribution and the target concept) on the number of samples needed to learn accurately @xcite .
	10-	, it is natural to allow the learning algorithm to decide when it has seen sufficiently many labeled samples based on the training samples seen up to now and their labels . since the above learning model includes any practical classification scheme , we term it universal (pac-) learning .
	11-	we first show that there is a computable learning algorithm in our universal setting .
	19-	our results imply that computable learning algorithms in the universal setting must waste samples " in the sense of requiring more samples than is necessary for statistical reasons alone .
	81-	then we will contrast this to the case of an uncomputable learning algorithm .
BERTSUMEXT (SW)	50-	(semantic requirements) for any @xmath27 , for any concept @xmath8 , and distribution @xmath9 over @xmath2 , if the oracle returns pairs @xmath28 for @xmath29 drawn iid from @xmath9 , then @xmath0 always halts , and with probability at least @xmath12 outputs a hypothesis @xmath13 such that @xmath30 ¡ { varepsilon}$]
	64-	suppose @xmath36 is an infinite sequence of iid samples drawn from @xmath9 .
	75-	the learning algorithm queries the oracle as necessary for new learning samples and their labeling .
	78-	note that it seems necessary to expand the hypothesis space to include all partial recursive functions because the concept space of total recursive functions does not have a recursive enumeration (it is uncomputable whether a given program is total recursive or not) .
	79-	we will see in theorem [thm : nobound] that there is no bound @xmath55 on the number of samples queried by any computable learning algorithm in our setting .
	80-	let us obtain some intuition for why that is true for the above learning algorithm .

GOLD		in this paper , we propose majority voting neural networks for sparse signal recovery in binary compressed sensing . the majority voting neural network is composed of several independently trained feedforward neural networks employing the sigmoid function as an activation function . our empirical study shows that a choice of a loss function used in training processes for the network is of prime importance . we found a loss function suitable for sparse signal recovery , which includes a cross entropy - like term and an @xmath0 regularized term . from the experimental results , we observed that the majority voting neural network achieves excellent recovery performance , which is approaching the optimal performance as the number of component nets grows . the simple architecture of the majority voting neural networks would be beneficial for both software and hardware implementations .
GBT-EXTSUM	40-	requires only several matrix - vector products to obtain an output signal , which is an estimate signal of the sparse vector @xmath12 .
	48-	the signal propagates from left to right and the output signal @xmath17 eventually comes out from the output layer . the network should be trained so that the output signal @xmath17 is an accurate estimation of the original sparse signal @xmath12 .
	168-	in this paper , we proposed sparse signal recovery schemes based on neural networks for binary compressed sensing .
	169-	our empirical study shows a choice of the loss function used for training neural networks is of prime importance to achieve excellent reconstruction performance .
	170-	we found a loss function suitable for this purpose , which includes a cross entropy like term and an @xmath0 regularized term .
	173-	the simple architecture of the majority voting neural network would be beneficial for both software and hardware implementation .
BERTSUMEXT (SW)	19-	the paper @xcite presents binary iterative hard thresholding (biht) algorithm by reforming iterative hard thresholding (iht) algorithm @xcite .
	20-	although the known sparse recovery algorithms exhibit reasonable sparse recovery performance , it may not be suitable for applications in high speed wireless communications .
	48-	the signal propagates from left to right and the output signal @xmath17 eventually comes out from the output layer . the network should be trained so that the output signal @xmath17 is an accurate estimation of the original sparse signal @xmath12 .
	137-	the outputs from these neural network are combined by soft majority voting nodes and the final estimation vector is obtained by rounding the output from the soft majority voting nodes . combining a several neural networks to obtain improved performance is not a novel idea , e.g. , @xcite , but it will be shown that the idea is very effective for our purpose . from statistics of reconstruction errors occurred in our computer experiments , we observed that many reconstruction error events (i.e. , @xmath97) occur due to only one symbol mismatch .
	149-	note that implementation of neural networks with fpga is recently becoming a hot research topic @xcite .
	151-	the length of the sparse signal is set to @xmath59 and the sparseness parameter is set to @xmath110 .) , width=317] from fig.[fig : rr_and_m_k6] , we can observe significant improvement in recovery performance compared with the performance of the single neural network . a single feedforward neural network discussed in the previous section