

Multimodal Machine Learning for Extraction of Theorems and Proofs in the Scientific Literature

ABSTRACT

Scholarly articles in mathematical fields feature mathematical statements such as theorems, propositions, etc., as well as their proofs. Extracting them from the PDF representation of the articles requires understanding of scientific text along with visual and font-based indicators. We pose this problem as a multimodal classification problem using text, font features, and bitmap image rendering of the PDF as different modalities. In this paper we propose a multimodal machine learning approach for extraction of theorem-like environments and proofs, based on late fusion of features extracted by individual unimodal classifiers, taking into account the sequential succession of blocks in the document. For the text modality, we pretrain a new language model on a 11 GB scientific corpus; experiments shows similar performance for our task than a model (RoBERTa) pretrained on 160 GB, with faster convergence while requiring much less fine-tuning data. Font-based information relies on training a 128-cell LSTM on the sequence of font names and sizes within each block. Bitmap renderings are dealt with using an EfficientNetv2 deep network tuned to classify each image block. Finally, a simple CRF-based approach uses the features of the multimodal model along with information on block sequences. Experimental results show the benefits of using a multimodal approach vs any single modality, as well as major performance improvements using the CRF modeling of block sequences.

1 INTRODUCTION

Context and problem definition. Scholarly articles in mathematical fields typically include theorems (and other theorem-like environments) along with their proofs. We consider an application where the objective is to transform this scientific literature from a collection of PDF articles to an open knowledge base of theorems, where mathematical results are the focus, which can be explored in new ways, such as searching for all theorems depending on a given result, or for all proofs including a specific feature (e.g., induction proofs on trees). We also envision applications to provide better ways of navigating dependencies of results contained in papers and references within.

As a first step in this direction, it is necessary to develop information extraction methods that automatically identify theorem-like environments and proofs in PDF scientific articles. A human would typically be able to perform this task by relying on the formatting of the text, on specific keywords identifying the environments, and on other visual clues: examples include (see Figure 1) the presence of keywords such as “Theorem” or “Proof” in bold or italics, the fact that an entire block of text might be in italics, the comparatively high proportion of mathematical characters, the presence of a QED symbol at the end of a proof, etc. However, precise formatting depends on document formats; a classifier that would only use such kinds of hard-coded rules does not generalize well for arbitrary formats, or for proofs that span multiple paragraphs.

Theorem 1.8 ([HRVW09]). Let $f : \{0,1\}^n \rightarrow \{0,1\}^n$ be a one-way function, let X be uniformly distributed in $\{0,1\}^n$, and let (Y_1, \dots, Y_m) be a partition of $Y = f(X)$ into blocks of length $O(\log n)$. Then (Y_1, \dots, Y_m, X) has next-block accessible entropy at most $n - \omega(\log n)$.

Proof. Since f is (t, ε) -one-way, the distributional search problem $(\Pi^f, f(X))$ where $\Pi^f = \{(f(x), x) : x \in \{0,1\}^n\}$ is (t, ε) -hard. Clearly, $(f(X), X)$ is supported on Π^f , so by applying Theorem 3.8, we have that $(\Pi^f, f(X), X)$ has witness hardness $(\Omega(t), \log(1/\varepsilon))$ in relative entropy and $(\Omega(t), \log(1/\varepsilon) - \log(2/\delta))$ in $\delta/2$ -min relative entropy. Thus, by Theorem 4.7 we have that (Y_1, \dots, Y_m, X) has next-block inaccessible relative entropy $(\Omega(t \cdot \Delta \cdot \ell^2 / (n^2 \cdot 2^t)), \log(1/\varepsilon) - \Delta)$ and next-block inaccessible δ -min relative entropy $(\Omega(t \cdot \delta \cdot \ell^2 / (n^2 \cdot 2^t)), \log(1/\varepsilon) - \log(2/\delta) - \Delta)$, and we conclude by Theorem 4.9. \square

Figure 1: Example of text, styling, and visual indicators of theorems and proofs

To clarify, in the whole of this paper we use *theorem* in the same sense as it is used in \LaTeX (say, by the `\newtheorem` command): a theorem-like environment is a structured statement, possibly numbered, formatted in a specific way and used to represent a formal (usually mathematical) statement: it can be a theorem, a lemma, a proposition, etc., but also a definition, a formal remark or example, etc. By *theorem* in the following we mean any statement of this kind. By *proof* we mean what would typically be rendered in \LaTeX in a proof environment: a proof or proof sketch of a result.

We propose to approach the theorem-proof identification problem by designing an approach based on multimodal machine learning that classifies each paragraph of an article into *basic*, *theorem*, and *proof* labels, based on the scientific language, on typographical information, and on visual rendering of PDF documents. Additionally, we also take into account information about the *sequence* of paragraph blocks, as well as spacing between paragraphs and page breaks, to exploit the fact that the label of a paragraph heavily relies on that of the previous (and possibly following) ones.

Related work. This extraction problem has received little interest in past research, though we now discuss two highly relevant works [6, 29].

Ginev et al. [6] proposed the task of identifying proofs and theorem-like environments from arXiv¹ articles using their HTML rendering via \LaTeX XML. Their approach involves detecting mathematical statements (along with other regions such as abstract and acknowledgements), introduced as a 50-class classification problem. They show that there is some link in the contextual information among paragraphs, which is then exploited by the textual modality over a BiLSTM-based encoder/decoder approach. This approach has two major limitations, however, which make it unsuitable for our needs:

(1) Their approach does not operate on raw PDFs but on HTML renderings, which makes it only possible when \LaTeX source code is available.²

¹<https://arxiv.org/>

²Note that in such settings, extracting theorems and proofs from the source, as we do in Section 4 to build a labeled dataset, seems a better alternative.

(2) Their approach is only evaluated on the first logical paragraph within a marked-up environment belonging to the label set (e.g., only the first paragraph of every proof), which makes the task much simpler, since the first word is highly indicative of the label in most cases. In contrast, we aim at differentiating such environments from regular text, and we aim at classifying all paragraphs within an environment, not just the first one. This is a significantly more challenging problem.

In addition, note that accessing the dataset of [6] requires signing an NDA³ whose terms prevent free use for research.

Another early work [29] towards theorem-proof identification was to build a proof of concept evaluating various unimodal approaches based on different evaluation metrics using NLP, computer vision, and a mix of heuristics (detection of specific keywords) and font-based information to identify mathematical regions of interest. The problem was posed as a 3-class classification problem operated on text *lines* extracted from raw PDFs obtained using pdfalto⁴. That work had some important limitations. First, text lines do not usually contain entire sentences and offer little context, which means they are hard to classify using automated NLP approaches. Second, the computer-vision approach was posed as an object detection problem, which is ill-adapted to detecting text blocks; this also makes it easier to integrate the computer-vision approach with other modalities, which operate at the text line level. Finally, the third modality (heuristics and font features) relied on mostly hand-crafted features. In addition, [29] does not offer any uniform way of comparing all three modalities used, since each used a different section of information for determining labels. In this paper, we aim at providing a uniform way of comparing the performance of different modalities, as well as combining them together.

Methodology and contributions. We use as a basic component for our task the Grobid⁵ tool [27], which is state of the art for information extraction from scholarly documents. Grobid parses a PDF document and use machine learning techniques to interpret it into a succession of paragraph blocks, presented in a structured XML representation. Because PDF is such a low-level document format, however, reconstructing paragraph blocks is not a trivial task, and Grobid sometimes fails at identifying the proper limits of paragraphs; we take this into account for our task. Though our target class labels for our problem are *basic*, *theorem*, and *proof*, we add a fourth class label *overlap* which aims at classifying blocks that were extracted by Grobid overlapping several paragraphs of different labels. The goal is to improve the performance of classifiers by allowing them to use this *overlap* label, but we only report performance on the main three classes.

To design a multimodal approach to the theorem-proof identification problem, we take inspiration from how a human being would solve the task, i.e., with the help of:

(1) Understanding of the scientific vocabulary and how mathematical writing is organized: it might be possible to recognize a proof or a theorem by the presence of such phrases as “We conclude by” or “Assume ... Then ...”.

(2) Visual features such as symbols and the use of bold or italic fonts: some document classes, for instance, format the content of a theorem all in italics and end all proofs with a QED symbol.

(3) Use of different font types and sizes in order within paragraphs, such as starting a paragraph with a word in bold or in italics.

(4) Sequential organization of blocks within a document. For example, if we know the label of both previous and next paragraphs are *proof*, it is likely that of the current paragraph is also *proof* (or possibly *theorem*; *basic* is unlikely); this is all the truer if vertical spacing between these blocks is small.

This suggests, respectively, the use of a language model to capture text-level information; the use of a computer-vision approach to capture visual features; the use of styling information to capture font-based information; and the use of a sequential model to capture information from block sequences. In addition, we want to be able to combine all these features in a unified multimodal approach.

We thus provide the following contributions in this paper:

(1) Three unimodal (text, vision, font information) models for the theorem-proof identification problem relying on modern machine learning techniques (transformers, CNNs, LSTMs) with a focus on reasonably efficient models as opposed to very large ones. Note that the text modality approach relies on pretraining a language model specific to our corpus, which may have applications beyond our task.

(2) A multimodal late fusion model that combines the features of all three modalities.

(3) A block sequential approach, based on a simple conditional random field model, that can be used to improve the performance of any unimodal and multimodal model by capturing dependencies between blocks.

(4) An experimental evaluation on a dataset of roughly 200k English-language papers from arXiv, with a separate validation dataset of 3.5k papers (amounting to 529k blocks).

Outline. We first present in Section 2 the three unimodal models. We then discuss in Section 3 how to combine them into a multimodal model, and how to add support for information about block sequences. We then provide a description of our dataset and how it has been constructed in Section 4. Before concluding, we provide detailed experimental results on all unimodal and multimodal models in Section 5. Note that related work relevant to a specific modality is described at the appropriate place in the paper. Because of space constraints, some information is relegated to the supplementary material.

2 UNIMODAL MODELS

We now present in turn the methodology of our three unimodal models: a pretrained transformer-based (RoBERTa-based) language model for text; an EfficientNetv2 [35] CNN for vision; and an LSTM model trained on font information sequences within each paragraph for fonts.

2.1 Text Modality

Pretraining language models. Recent advancements in natural language processing have significantly improved the capabilities of language models. One well-established benchmark for assessing

³<https://sigmathling.kwarc.info/resources/arxmliv-statements-082018/>

⁴<https://github.com/kermitt2/pdfalto>

⁵<https://github.com/kermitt2/grobid>

these models is the SuperGLUE benchmark [38], which provides an average score over eight different downstream language tasks. These tasks serve as a crucial metric for highlighting the overall understanding of trained models. To achieve high scores on downstream tasks, it is often critical to pretrain the language models to develop a generic understanding of the domain before applying task-specific fine-tuning. In this paper, we evaluate language models of comparable sizes to investigate the impact of pretraining on our task of theorem-proof identification.

In our task, as in many, the vocabulary and semantic relationships between words can be very different from what a model pretrained on general English would learn. To capture language understanding, models are often trained in a self-supervised fashion using masked language modeling (MLM) loss. Therefore, the performance of the model depends not only on the architecture but also on the pretraining data provided. Although it is possible to fine-tune English-based models on a task like text classification, they often fail to capture domain-specific terminologies and their context when adapted for any other task.

The complex and technical understanding of scientific terms poses a challenge for natural language processing models. To address this issue, a domain-adapted pretraining step is necessary to enhance the model’s ability to encode and comprehend scientific language. Recent research by Suchin et al [9] indicates that the performance of a language model can be further increased by pretraining on top of an already trained model such as RoBERTa; they show incremental improvements by further pretraining on a corpus of Biomed, News, Reviews, and CS datasets [3]. To better compare the overall performance, we do not employ this strategy. Instead, we pretrain the model from scratch only on a corpus of mathematical articles and then compare it with the language model trained on English.

Note that final performance of the language model at our task is not our only target: we are also interested in models that, after pretraining, require fewer samples to fine-tune for our task (and possibly other tasks).

Related work. Several existing works have built a domain-specific language model for scientific papers, such as SciBERT [2], BioBERT [23], and MathBERT [31].

MathBERT, a language model designed to capture mathematical relationships [31], involves pretraining on text within papers with mathematics, which is jointly trained with mathematical formulas and their corresponding contexts as they appear in scientific articles. MathBERT performs comparably well on mathematical information retrieval, formula topic classification, and formula headline generation as compared to the standard BERT model. The authors mention that, for the preprocessing task, they apply a \LaTeX tokenizer which requires access to \LaTeX sources of papers at inference time, in contrast to our setting where only the PDF version is available.

BioBERT is trained on scientific papers in the medical science domain, which is not quite aligned with our application domain. On the other hand, SciBERT is trained on 1.14 million papers from Semantic Scholar⁶, of which 18% papers (around 200k, roughly the same size that our pretraining dataset, see Section 4) come from the computer science domain, while the rest is from a broad

⁶<https://www.semanticscholar.org/>

biomedical domain. We keep SciBERT as a comparison point in our experiments.

Many studies in the past, such as CamemBERT [28], which was trained on 4 GB of web-crawled data, have shown better performance than models trained on 130+ GB of data, and have demonstrated performance gains on several domain-specific tasks involving the French language. While BERT-like models can still make good guesses for text classification tasks without benefiting from scientific vocabulary, there are several advantages to applying a domain-specific pretraining step:

(1) By changing the model’s head, the same model can be used for other downstream tasks where performance may differ drastically, such as tasks that requires a more specific understanding of vocabulary, such as masked language modeling or entity recognition for formulas or scientific QA.

(2) A pretrained model can give better or on-par performance compared to models that do not benefit from domain knowledge, while training on a fraction of pretraining and fine-tuning data [28].

(3) Pretrained models also tend to converge faster, which can be extremely useful for few-shot learning scenarios for tasks where the model has very few samples to generalize upon.

Methodology. Based on the previous discussion, we pretrain a language model from scratch on a 50k vocabulary size (with byte-pair encoding), similar to the configuration of RoBERTa base (124M) [25]. While masking 15% of the tokens we kept the configuration similar to the original RoBERTa ($L = 12$, $H = 768$, $A = 12$), but on a different vocabulary. The model was provided with dynamic masking and trained on MLM loss.

In our pretraining, we used 11 GB of pretraining text data (196 846 scientific articles, see Section 4), trained over 11 epochs. We used the LAMB optimizer [42] and produced the results using a total batch size of 256 across 4 NVIDIA A100 GPUs with a distributed mirrored strategy and an initial learning rate of 2×10^{-5} . The total pretraining time was 176 hours.

2.2 Vision Modality

Vision classifiers. CNNs have been successfully applied to many image classification tasks and used as feature backbones in complex visual-language tasks. However, most widely accepted benchmarks used to evaluate these architectures are based on the ImageNet and CIFAR datasets, which evaluate their top-1% accuracy. Many times, weights trained on the ImageNet dataset can be used for transfer learning-based tasks for other real-world image classification problems, where there is a lack of data, offering significant gains over training the network from scratch.

In our case, we aim to capture certain mathematical symbols or fonts, as well as the overall appearance of a paragraph block, in order to identify proofs and theorem-like environments. This is a task different enough from that of ImageNet to decide in favor of training a model from scratch. One simple visual indicator that a vision model could look for is the token “Proof.” in italic or bold font, which can be a strong indicator of the label of the paragraph. Another strong indicator is the QED symbol at the end of the proof. Note that such clues are not fully available to the text modality.

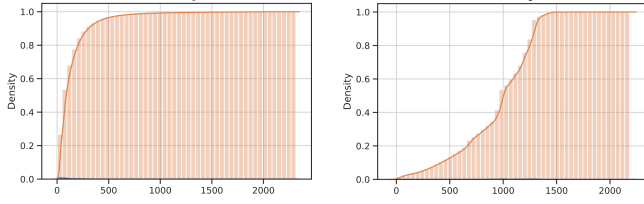


Figure 2: Cumulative distribution of heights (left) and widths (right) of paragraph blocks in our dataset

Preprocessing. To apply vision classifiers to our problem setting, we have two preprocessing challenges:

(1) **Text images with different aspect ratios:** Images need a standard size (fixed aspect ratio) for which these vision architectures are specifically crafted. In our setting, our paragraphs come in numerous aspect ratios, due both to the document class (e.g., one-column vs two-column paragraphs), or the particular style of specific authors. To standardize input dimensions to the CNN model, one very common method usually adopted for natural scenic objects is to apply an interpolation method to resize the images to a standard size. However, we argue that such a technique may completely alter the visual geometry of words, symbols, and underlying fonts, and hence is not suited for our task. Instead, we fix a standard aspect ratio, chosen based on actual widths and heights of blocks in our corpus. We show the cumulative distribution function of the heights and widths of blocks in pixels after rasterization on Figure 2. From the CDF plots, we can say that at least 80% of the paragraphs have their heights less than or equal to 400 and width less than or equal to 1400 pixels. Thus, we freeze the input size to be (400×1400). For any size above this resolution, we crop the area that does not fit within, whereas for the patches that are under the resolution, we vertically and horizontally pad them with white to a fixed standard 400×1400 resolution.

(2) **Background color as white in scientific articles:** Many studies [14] have indicated that training neural networks on negative images adversely impacts the performance of trained CNN models. Moreover, some standard CNN architectures apply max-pooling, an operation to detect a strong intensity feature present in a neighborhood denoted by the kernel size. This can often pick up the background if it is denoted by a higher pixel value. Since scientific papers originally have a white (maximum intensity) background with black text indicating the characters, to avoid this issue, we employ a bitwise not operator to invert the color of images, making it very similar to MNIST samples (white text on a black background).

Related work. Most vision architectures are now determined based on their performance on benchmark datasets such as ImageNet, Flowers, and CIFAR100. However, these real-world objects are very different from text-filled images in scientific papers. Moreover, the network parameters are tuned for the specific image resolution, such as square images of 224^2 or 331^2 resolution, specific to ImageNet. Therefore, finding the right architecture for a custom dataset like ours can be challenging.

The design of the architecture almost always plays a crucial role in determining the performance of the model. From one of the

very first CNN architectures, Lenet-5 [22], several advancements have been made in the deep learning community that have overall provided better CNN designs:

(1) Introduction of skip connections [11] in residual networks to avoid the gradient vanishing problem for deep layered networks.

(2) Use of batch normalisation [18] to reduce internal covariance shift.

(3) Dropouts in fully connected layers [34] to act as a regularizer.

(4) Better activation functions switching from Sigmoid to Relu, Leaky Relu, Swish, Gelu (for later architectures) [12] for transformer models to avoid neuron saturation problem.

Many papers have adopted these changes within the architecture, for example AlexNet [20] uses Relu activation within the architecture, similarly ResNet [11] adopts batch normalization and skip connections. Thus the problem simply boils down to finding the right proposed architecture which should include one or several of these advancements without having to figure out more specific design choices. In this paper we chose our vision backbone based on networks with skip connections, to avoid the vanishing gradient problem that happens when training plain neural networks without a skip connection such as in AlexNet [20] and VGGNet [33]; an interesting study [24] tries to visualize the loss surface of plain networks vs networks with skip connections which points towards faster convergence with low number of local minima surfaces.

ResNet [11] in 2015 was one of the first architecture to adopt the idea of skip connections. Later, this idea was extended in DenseNet where every n -th layer was connected to all preceding layers within the network [16]. NASNet [43] introduced neural architecture search (NAS), an approach based on reinforcement learning to select the best network architecture.

More recent architectures such as EfficientNet [35] use NAS to find the right set of hyperparameters of a baseline network (comprising depthwise convolutions) and then principally scaling the depth, width, and the resolution of input images simultaneously with the help of an empirically determined ratio; in their paper this approach was referred to as *compound scaling*. This does not only work for EfficientNet but also when scaling previous architectures such as MobileNets [15] and ResNet [11] on several image classification tasks including the ImageNet dataset. The resultant architecture was 5 times more efficient in terms of number of floating-point operations (FLOPs) at the inference time as compared to previously proposed NASNet, while also giving better results and requiring less memory. This makes it very practical for a backbone in a multimodal setting such as ours, where several backbones for different modalities will need to be combined, and therefore minimizing computations on each modality is a goal. We thus focus on EfficientNet architectures.

After the introduction of Vision transformers [5] for image classification where a pure transformer model can outperform CNNs (EfficientNet in this case) when provided with large amounts of data (JFT-300M), there has been a constant debate in the community between the use of transformers over CNNs. It would be interesting to compare the performance differences of both architectures in our setting but due to computation resources we have decided to restrict ourselves with CNN-based approaches only. Furthermore, several studies [26, 36, 39] have indicated that more recent CNNs perform

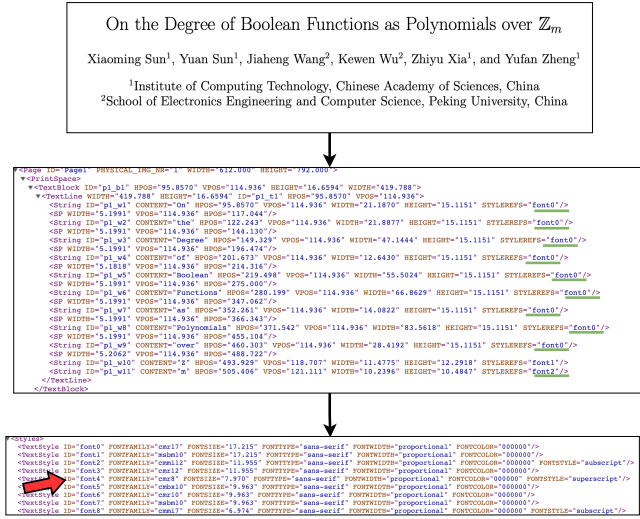


Figure 3: Font information as presented in the output of the pdftalto tool. Note that, for instance, the \mathbb{Z} or m characters of the title are written in different fonts from the other tokens.

comparably to vision transformers while requiring far less computational resources (here measured in FLOPs) and training time. One very prominent CNN architecture that showcases comparable performance to a transformer architecture is EfficientNetv2 [36].

EfficientNetv2 [36] introduced a new family of model, with 11 times faster training speed and up to 6.8 times better parameter efficiency on several datasets. They were able to achieve this performance by switching the depth-wise convolutions (MBConv) to FusedMBConv in the earlier part (stage 1–3) of the network. The baseline architecture was found using training-aware NAS which jointly optimizes accuracy, parameter efficiency, and training efficiency. EfficientNetv2 focuses on a lower kernel size of 3×3 but adds more depth to compensate for the loss.

Choice of model. EfficientNet comes with several variants (B0–B7) where B7 has the largest receptive field due to compound scaling. We select in our experiments a base network (B0), a medium-sized network (B4) and the largest network (B7). EfficientNetV2 also comes with different sizes. We focused on the small (EfficientNetV2s) and medium-sized (EfficientNetV2m) models.

2.3 Font Modality

The last modality we consider is styling information present in the PDF in terms of the sequence of fonts (font family and font size) used in a specific paragraph. This information can be obtained using the pdftalto tool, which produces a list of fonts used in a given document, and associates each text token to a particular font, as illustrated in Figure 3. Fonts are usually standard \LaTeX fonts, such as cmr10 for Computer Modern Roman in 10 point.

From the training data, we build a font vocabulary of 4 031 unique fonts including their sizes, and represent every paragraph block as a sequence of font identifiers. To match input dimensions among training samples, we apply left padding with a maximum length of 1000. We then feed the entire sequence to a simple 128-cell

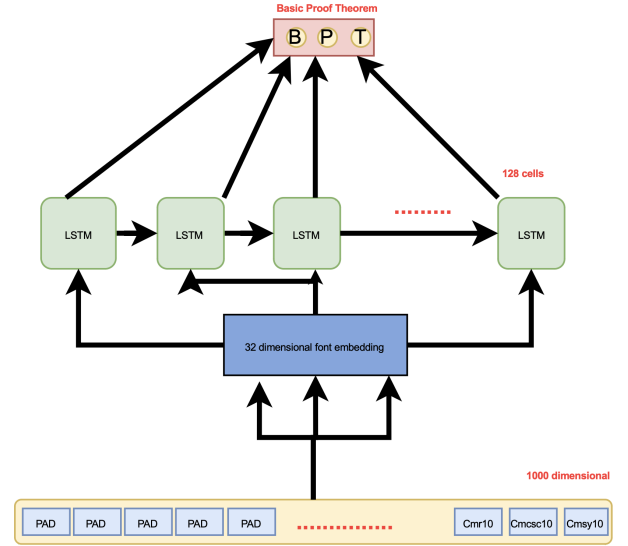


Figure 4: LSTM model for the font modality

LSTM [13] cell network to monitor the loss, represented in Figure 4. The choice of the LSTM model is purely to capture sequential information within fonts that can be used to identify the label of the paragraphs.

3 MULTIMODAL AND SEQUENTIAL MODELS

We now go beyond unimodal models by showing how all three modalities can be combined into a single late-fusion multimodal model, and how block sequence information can be captured.

3.1 Multimodal Approach

Related work. Multimodal machine learning for document AI has recently gained popularity. However models from the literature are typically poorly adapted to scientific articles (not taking into account specific encoding of font features, scientific terminology, or documents potentially consisting of dozens of pages, each containing dozens of blocks). Indeed, the main target application in the literature is towards visual document understanding benchmarks such as FUNSD [8], CORD [30] or RVL-CDIP [10] (invoices, receipts, forms, memos, etc.). One of the representative approach is LayoutLM [40], which takes into account 2D positional embedding via masked visual-language model loss and multi-label document classification loss. Training such models for scientific texts would come with an extra cost of adding another objective on top of language pretraining which already captures some domain-specific corpus. LayoutLMv2 [41] adds new text–image alignment and text–image matching tasks, while LayoutLMv3 [17] adds a word–patch alignment objective. Although the performance of these models are compared with unimodal approaches they are not compared to other multimodal baselines, such as as simple fusion baseline.

In this paper we focus on late fusion of modalities instead of fusing them early on, so as not to require any extra pretraining cost. Two popular works made in this direction are based on *gated*

multimodal units (GMU) [1] and bilinear gated units [19]. Though this is usually done in the context of only two modalities (typically, text and image), [1] extends the idea to multiple modalities.

Methodology. We re-implement the architecture of [1] in our setting, incorporating features from all three unimodal classifiers. Hence, our GMU operates on features extracted from a RoBERTa-like model pretrained from scratch (see Section 2.1), an EfficientNetV2m model chosen as the best performing vision approach (see Sections 2.2 and 5), and the font sequence LSTM of Section 2.3. We also set the weights of these backbones to be non-trainable. The detailed architecture is in the supplementary material.

3.2 Sequential Approach

In addition to the modalities, considering the sequencing of the blocks, i.e., the order in which they appear in the document, allows us to determine with greater confidence the class of each block. For example, if one seeks to predict the class of a block that is itself framed by two blocks that have been classified as *proof*, then there is a good chance that this block is itself a *proof*. We consider how to integrate unimodal and multimodal models into a sequential prediction model.

Conditional Random Fields (CRFs) [21] are a statistical, sequential, learning model based on an undirected probabilistic graphical model. We use the simplest variant of CRFs here, linear-chain order-one CRFs, which amounts to making a first-order Markov assumption for the dependency of block sequences: the label of a given block is modeled to only depend (in addition to the features) on the label of the previous block. Note that since CRF optimization searches for a global optimal over the entire sequence, in practice the label of other blocks will also have an impact.

We use the following features, depending on the modalities:

- The last hidden state of the CLS token for the unimodal text model. It corresponds to a set of 768 features.
- The last dense representation (just after the flattening operation) for the unimodal vision model. It corresponds to a set of 1 280 features.
- The last LSTM embedding representation of font-based unimodality. It corresponds to a set of 128 features.
- The last embedding layer of the multimodal approach. It corresponds to a set of 768 features.
- We also add the following four features describing the geometrical positions of blocks: the normalized page number, i.e., the ratio between the current page of the considered block over the total number of pages of the sequence; the normalized horizontal distance with the previous block, with respect to the horizontal length of the current page; the normalized vertical distance with the previous block, still with respect to the horizontal height of the current page; and a binary feature describing whether the previous block is on the same page as the current block.

4 DATASET AND SETUP

We now discuss the dataset used. The different steps of dataset preparation are summarized in Figure 5.

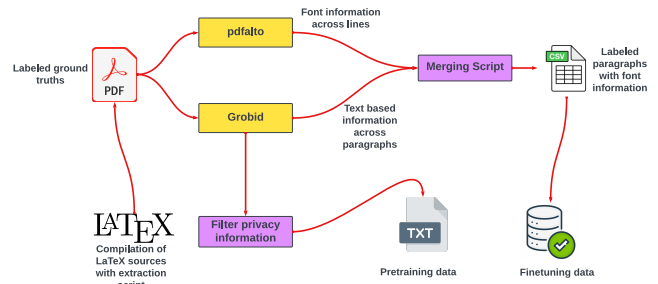


Figure 5: Overview of dataset preparation pipeline

We collected all arXiv papers (PDF and sources) until May 2020, using arXiv’s bulk data access on Amazon S3⁷. We implemented a ground-truth annotation script to determine the positions of theorem-like environments and proofs in PDF documents based on their \LaTeX sources; note this is a non-trivial step, as \TeX is a Turing-complete language – we designed a \LaTeX package that instruments commands such as `\newtheorem` in order to add some annotation to the resulting compiled PDF.

We filtered articles from the dataset to only keep those written in English, for which \LaTeX source is available (which is, according to arXiv’s policy, all those that have been produced using \LaTeX), that were compilable on a modern \LaTeX distribution, that contained at least a theorem or a proof environment, and for which none of the tools necessary (our ground-truth annotation package, Grobid for extraction of blocks, pdfalto for line-by-line font sequences, image rasterization for production of inputs to the CNN) failed to produce a valid output.

The output of Grobid and that of pdfalto is merged in order to obtain the final dataset sequence of paragraphs within the PDF document, labeled with their text, their position within the PDF (for rasterization of this specific area), their ground-truth label, and their font sequences. This forms our overall dataset.

As we explained in the introduction, Grobid sometimes fails to extract correct paragraphs, meaning that some of the paragraphs identified by Grobid overlap blocks of different category (say, *basic* and *theorem*). We label such paragraphs with the *overlap* class, exclusively used for these outliers in the data.

We fix a validation dataset as a subset of the entire dataset formed of roughly 500 000 samples (paragraph blocks), coming from 3 682 PDF articles selected at random. The training dataset is formed of blocks from all other articles.

For pretraining our language model, the entire content of the training dataset is used, after filtering what appears to be potential personal information (author names within metadata and references, as identified by Grobid) – as we intend to release our language model, we wanted to ensure it contained as little personal information as possible.

The training dataset is split at random into batches of 1 000 PDF articles (consisting of roughly 200 000 samples of blocks each). Classifiers are incrementally fit on successions of batches, in a fixed

⁷https://info.arxiv.org/help/bulk_data_s3.html

Table 1: Overall performance comparison (accuracy and mean F_1 over the three classes *basic*, *theorem*, and *proof*) of individual modality models and multimodal model, with and without the block-sequential CRF model; for each model, the number of batches (1 000 PDF documents, roughly 200k samples) it was trained on is indicated

Modality	Model chosen	Block sequence CRF	#Batches	Accuracy (%)	Mean F_1 (%)
Dummy	—	—	—	59.41	24.85
Font	LSTM 128 cells	no	11	65.00	45.00
		yes	11+1	52.20	50.49
Vision	EfficientNetV2m_avg	no	9	69.43	60.33
		yes	9+1	74.13	69.82
Text	Pretrained RoBERTa-like	no	20	76.45	72.33
		yes	20+1	82.70	80.52
Multimodal	GMU	no	10	76.86	73.87
		yes	10+4	84.38	83.01

order, until convergence is achieved. We never need to use more than a couple of dozen batches.

Once unimodal classifiers are trained, their weights are frozen and serve within the multimodal classifier, which is trained in turn. After every unimodal or multimodal classifier is trained, their weight is used as feature extractors for the CRF sequential model, as explained in Section 3.2.

All experiments were run on a supercomputer with access at any point to 4 NVIDIA (V100 or A100) GPUs.

5 EXPERIMENTAL RESULTS

We now report experimental results on the *basic-theorem-proof* classification problem, first comparing representative unimodal classifiers, with and without block sequential model, to the multimodal classifier. We then delve into more detail at specifics of every unimodal classifier.

We are interested in two main performance metrics: *accuracy* measures the raw accuracy of the classifier on the validation dataset (disjoint with the training dataset); and (unweighted arithmetic) *mean F_1 -measure* of the *basic*, *theorem*, and *proof* classes, which summarizes the precision and recall over each class assigning the same weight to every class. As *basic* is the most common class in the dataset, a *dummy* classifier that would always predict the *basic* class would have an accuracy of 59.41%; but its recall would be 100% on *basic* and 0% on the other classes, while its precision would be 59.41% on *basic* and 0% on the other classes, resulting in a mean F_1 of $\frac{1}{3} \times \frac{2 \times 59.41\%}{59.41\% + 100\%} \approx 24.85\%$. This gives an important comparison point for all other methods; accuracy measures how well the classifier works on the actual unbalanced data, while mean F_1 favors methods that perform well to identify all three classes, which is arguably a better metric.

Overall Results. The results we obtain for the different modalities, with and without the use of the block sequence model, are shown in Table 1. The following lessons can be drawn from these results:

(1) This is a hard task, as the best performance reached is 84% for accuracy and 83% for mean F_1 . Informal experiments we did with some human annotators show that it is actually sometimes hard for a human being to distinguish whether a block is part of a

Table 2: Performance comparison of text models

Model	#Batches	Inf. time (ms/step)	Accuracy (%)	Mean F_1 (%)	#Params
Dummy	—	—	59.41	24.85	—
Pretrained	20	58	76.45	72.33	124M
RoBERTa base	20	62	76.61	71.66	124M
SciBERT base	20	59	76.89	73.00	110M

proof or theorem environment, especially in the middle of it, so it is unsurprising that we cannot reach near-perfect results.

(2) Looking at unimodal models: the font-based model performs rather poorly, though still beating (at least in terms of mean F_1) the dummy model, meaning that some thing is learned; the text-based model is the best performing one, suggesting that textual clues impact more than visual ones for this task.

(3) Including the CRF sequential model, despite its relative simplicity, greatly increase the performance of every unimodal or multimodal model, by 5 to 10 points of accuracy or mean F_1 (with a single exception for the accuracy of the font sequence model). The importance of the use of an approach modeling block sequences is thus clear.

(4) The multimodal model outperforms every unimodal model, though the margin with the text model is somewhat low.

Putting these together, we put forward our multimodal model with block-sequential CRF model as a state-of-the-art candidate for identification of theorems and proofs for scientific articles. The level of accuracy and mean F_1 reached, if not perfect, is acceptable for automatic processing of articles and construction of a knowledge base of theorems, which may need to be further manually cleaned and curated. We now look into some more detail at the performance of unimodal models.

Text Modality. To measure the performance of various language models (our language model pretrained on our corpus, RoBERTa, and SciBERT), we evaluate their accuracy as shown in Table 2 and Figure 6 on the validation dataset. All three have similar numbers of parameters (obviously the same for our pretrained and the base version of RoBERTa), have similar inference time, and reach similar levels of accuracy (76.45% to 76.89%) and mean F_1 (71.66% to 73.00%)

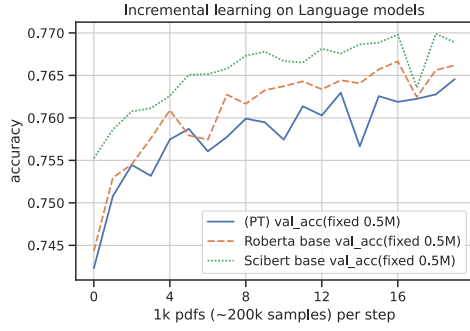


Figure 6: Accuracy of language models on fine-tuning task with respect to number of batches

Table 3: Samples to reach target accuracy for text models

Model	Data size	Samples to 65%	Samples to 70%
Pretrained	11 GB, 197k papers	39 552	141 632
RoBERTa base	160 GB	41 472	186 496
SciBERT base	1.14M papers	36 928	91 200

Table 4: Performance comparison of vision models

Model	#Batches	Inf. time (ms/step)	Accuracy (%)	Mean F ₁ (%)	#Params
Dummy	-	-	59.41	24.85	-
EfficientNetB0	5	49	65.27	46.00	6.9M
EfficientNetB0_max	5	35	58.21	25.00	4.0M
EfficientNetB0_avg	5	34	62.93	39.66	4.0M
EfficientNetB4_avg	5	61	65.87	47.33	17.6M
EfficientNetB7_avg	5	145	61.22	42.33	64.1M
EfficientNetV2s_avg	5	70	59.41	25.00	20.3M
EfficientNetV2m_avg	5	94	64.02	42.66	53.2M
EfficientNetB4_avg	9	88	68.46	54.33	17.6M
EfficientNetV2s_avg	9	71	59.80	27.00	20.3M
EfficientNetV2m_avg	9	92	69.43	60.33	53.2M

and converge after training on 20 batches. SciBERT does have slightly higher performance.

Table 2 shows another side of the picture: to reach a target level of accuracy (say, 65% or 70%), our pretrained model needs much fewer fine-tuning data than the RoBERTa model (trained on a corpus of 15 times more text data). Although SciBERT performs better than our pretrained model on this metric as well, it is important to note that it has been trained on 5.5 times more scientific papers than our pretrained model.

Vision Modality. The performance of a wide variety of vision-based models is displayed in Table 4. For the simplest model, we experiment with different forms of pooling for the last convolutional layer: none, max pooling, or average pooling. We see that no pooling yields performance, on a small model that is quite close and comparable to much larger models. We also notice that average pooling works quite well in most cases.

Font Modality. In addition to our model formed of an LSTM with 128 cells, and in order to investigate potential further gains, we try switching LSTM cells to GRU [4]; and using a Bidirectional LSTM

Table 5: Performance comparison of font models

Model	#Batches	Inf. time (ms/step)	Accuracy (%)	Mean F ₁ (%)	#Params
Dummy	-	-	59.41	24.85	-
LSTM (128)	11	14	65.00	45.00	1.72M
GRU (128)	11	14	61.00	42.71	1.72M
BiLSTM (128)	11	26	68.26	45.66	1.82M

Теорема 3 При $\log_n k = o(n)$ в классе функций P_k^n существует функция, минимальная ДНФ реализация которой содержит не менее N_k конъюнкций, где для N_k выполнено

$$N_k \geq \Omega\left(\frac{nk \log n}{\log k}\right).$$

Доказательство. Указанную оценку можно получить оценив сложность покрытия околу-левых точек функции принимающей нулевые значения на одном из слоев булева куба. ■

Figure 7: Excerpt from an example Russian-language paper [7] with high performance of font model

to capture sequential information across both forward and backward axis. Our results from Table 5 indicate that the bidirectional component in fonts alone does not have a huge impact in deciding the label of the blocks, even if modest gains are observed.

Despite the low score obtained by the font models, they can still be of use in certain situations. For example, anecdotally, Figure 7 shows an example of a Russian-language article whose blocks are correctly classified by the font-based model, while the text model is not able to use any clues as it was trained on English text.

6 DISCUSSION & FUTURE WORK

One of the strongest modality, independently taken, for our task is clearly text, followed by vision and finally font sequence, with low individual performance (but also highest efficiency and lowest number of parameters – around 70× less than that of our language model).

Though our models are not directly interpretable, we can get some post-hoc explanations of their performance: for the text modality, we can visualize [37] the attention heads of the last layer of the language model to see what the model focuses on. Informal experiments suggest, unsurprisingly, that tokens indicative of these environments (such as the “Theorem” or “Proof” tokens, or numberings) are given a strong weight in determining the label of the paragraphs. For the vision model, we can use the Grad-CAM [32] visualization, which indeed show that some layout-based information is captured by the model.

We consider two main directions for future work: First, multi-modal transformers such as LayoutLM [17, 40, 41] try capturing relationships between modalities early on by using novel losses such as masked visual-language model, text–image alignment and text–image matching, or word–patch alignment. It would be interesting to investigate how to integrate our models within similar architectures so as to capture cross-modal interactions better. Second, our CRF-based approach currently only looks at the previous block to decide the label of the current block; we envision that an attention-based mechanism across both forward and backward access can precisely link certain paragraphs that have a strong connection even when they are several pages apart, as it often happens in scientific papers that are quite lengthy (e.g., theorems that are restated).

REFERENCES

- [1] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González. Gated multimodal networks. *Neural Computing and Applications*, 32, 2020.
- [2] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *EMNLP/IJCNLP*, 2019.
- [3] S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *LREC*, 2008.
- [4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*, 2014.
- [5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [6] D. Ginev and B. R. Miller. Scientific statement classification over arXiv.org. In *LREC*, 2020.
- [7] S. Granin and Y. Maximov. Average case complexity of DNFs and Shannon semi-effect for narrow subclasses of boolean functions. *arXiv:1501.03444*, 2015.
- [8] J.-P. T. Guillaume Jaume, Hazim Kemal Ekenel. Funsd: A dataset for form understanding in noisy scanned documents. In *OST@ICDAR*, 2019.
- [9] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: adapt language models to domains and tasks. In *ACL*, 2020.
- [10] A. W. Harley, A. Ufkes, and K. G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, 2015.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [12] D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs). *arXiv:1606.08415*, 2016.
- [13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8), 1997.
- [14] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran. On the limitation of convolutional neural networks in recognizing negative images. In *ICMLA*, 2017.
- [15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.
- [16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.
- [17] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei. LayoutLMv3: Pre-training for document ai with unified text and image masking. In *ACM MM*, 2022.
- [18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [19] D. Kiela, E. Grave, A. Joulin, and T. Mikolov. Efficient large-scale multi-modal classification. In *AAAI*, 2018.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 2017.
- [21] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.
- [22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.
- [23] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 2020.
- [24] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. In *NIPS*, volume 31, 2018.
- [25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.
- [26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *CVPR*, 2022.
- [27] P. Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *ECDL*, volume 5714, pages 473–474, 09 2009.
- [28] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, and B. Sagot. CamemBERT: a tasty French language model. In *ACL*, 2020.
- [29] S. Mishra, L. Pluvinaud, and P. Senellart. Towards extraction of theorems and proofs in scholarly articles. In *DocEng*, 2021.
- [30] S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee. CORD: A consolidated receipt dataset for post-OCR parsing. In *DI@NeurIPS*, 2019.
- [31] S. Peng, K. Yuan, L. Gao, and Z. Tang. MathBERT: A pre-trained model for mathematical formula understanding. *arXiv:2105.00377*, 2021.
- [32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.
- [33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1), 2014.
- [35] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.
- [36] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *ICML*, 2021.
- [37] J. Vig. A multiscale visualization of attention in the transformer model. In *ACL*, 2019.
- [38] A. Wang, Y. Punksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, volume 32, 2019.
- [39] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders. *arXiv:2301.00808*, 2023.
- [40] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *SIGKDD*, 2020.
- [41] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *ACL/IJCNLP*, 2021.
- [42] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *ICLR*, 2020.
- [43] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *CVPR*, 2018.

A EXTRA MATERIAL FOR SECTION 2 (UNIMODAL MODELS)

A.1 Text Modality

We investigate the difference in base vocabulary of different language models: Figure 8 compares the overlap of vocabulary between various language models – the one we are proposing has a maximum of 33% overlap with others, including SciBERT which is trained on scientific text, suggesting the relevance of a pretrained model with vocabulary specific to our corpus.

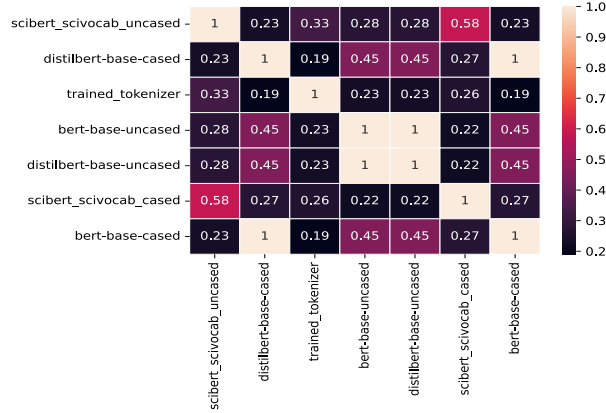


Figure 8: Vocabulary overlap among popular language models (BERT, DistilBERT, SciBERT, in cased or uncased variants and our pretrained model (labeled as trained_tokenizer here))

We report pretraining results on two pretraining configurations (see Table 6): a BERT-like model in addition to the RoBERTa-like model described in the main text. As a quantitative measure of the quality of the pretraining, we report the perplexity of the pretrained language model on the MLM task, similar to Table 3 in the RoBERTa paper [LOG⁺19]. We show the evolution of the MLM loss in Figure 9. For a qualitative analysis, we intentionally picked up samples that require specific vocabulary understanding on the MLM task, see Table 7.

Language model	Batch size	Steps	Learning rate	Perplexity (epoch #10)	Time per epoch (h)
BERT-like (110M)	256	47 773	2×10^{-5}	3.034	11
RoBERTa-like (124M)	256	47 773	2×10^{-5}	2.857	16

Table 6: Pretraining configurations (on arXiv dataset)

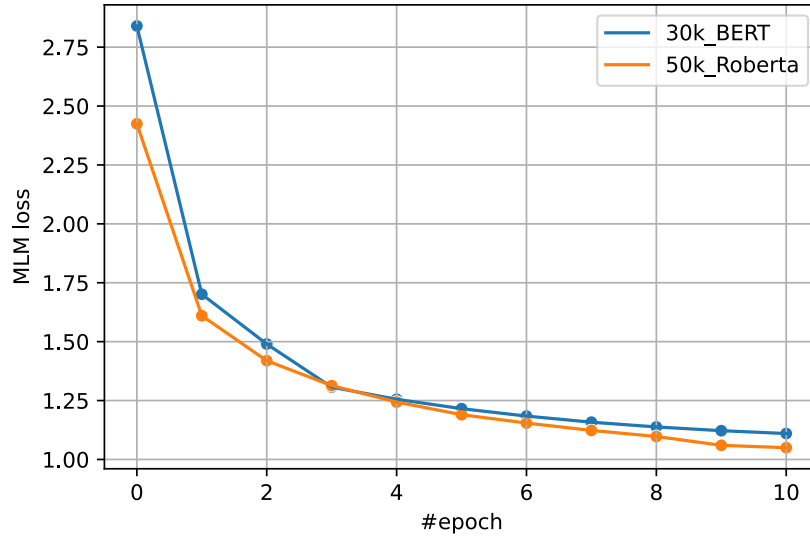


Figure 9: MLM loss for two pretrained models, as a function of the pretraining epoch

Masked sentence	Pretrained BERT-like model	BERT model
This concludes the [MASK].	proof lemma claim theorem case thesis	game story film play episode novel
We show this by [MASK].	induction . definition a lemma contradic- tion	ourselves accident name themselves ear hand
By [MASK]’s inequality.	jensen holder young minkowski cauchy	fourier brown russell fisher newton
The [MASK] is definite positive.	inequality case slimit sum function	result sign value answer form
In particular any field is a [MASK].	. 1 group f field	field theory domain variety category
To determine the shortest distance in a graph, one can use [MASK]’s algorithm.	dijkstra grover tarjan newton hamilton	shannon newton taylor wilson moore
An illustration of the superiority of quan- tum computer is provided by [MASK]’s algorithm.	grover shor dijkstra yao kitaev	turing newton shannon maxwell einstein
One of the ways of avoiding [MASK] is using cross validation, that helps in esti- mating the error over test set, and in decid- ing what parameters work best for your model.	overfitting error errors misspecification noise	errors error this bias uncertainty

Table 7: Inference of BERT-like pretrained model on selected MLM tasks

1277	B EXTRA MATERIAL FOR SECTION 3 (MULTIMODAL AND SEQUENTIAL MODELS)	1335
1278	We show in Figure 10 the detailed architecture of the multimodal model.	1336
1279		1337
1280		1338
1281		1339
1282		1340
1283		1341
1284		1342
1285		1343
1286		1344
1287		1345
1288		1346
1289		1347
1290		1348
1291		1349
1292		1350
1293		1351
1294		1352
1295		1353
1296		1354
1297		1355
1298		1356
1299		1357
1300		1358
1301		1359
1302		1360
1303		1361
1304		1362
1305		1363
1306		1364
1307		1365
1308		1366
1309		1367
1310		1368
1311		1369
1312		1370
1313		1371
1314		1372
1315		1373
1316		1374
1317		1375
1318		1376
1319		1377
1320		1378
1321		1379
1322		1380
1323		1381
1324		1382
1325		1383
1326		1384
1327		1385
1328		1386
1329		1387
1330		1388
1331		1389
1332		1390
1333		1391
1334		1392

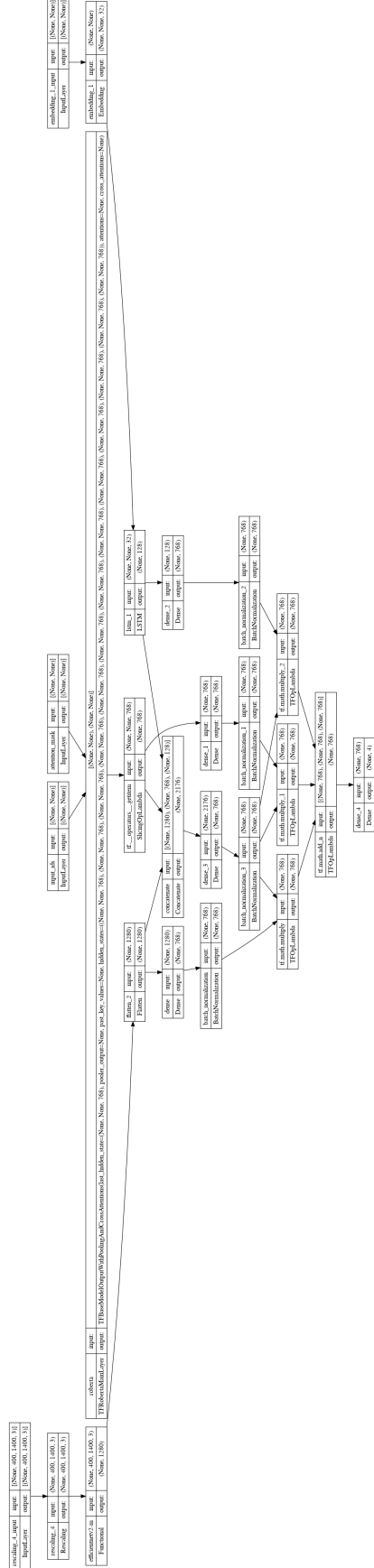


Figure 10: Architecture diagram of the multimodal model

Proof. By Theorem 3 , relative to a suitable oracle $A_{\mathcal{D}}$ (in fact, a *random* oracle suffices), there exists a signature scheme \mathcal{D} , such that any quantum chosen-message attack against \mathcal{D} must make superpolynomially many queries to $A_{\mathcal{D}}$. The oracle $A_{\mathcal{S}}$ will simply be a concatenation of $A_{\mathcal{M}}$ with $A_{\mathcal{D}}$. Relative to $A_{\mathcal{S}}$, we claim that the mini-scheme \mathcal{M} and signature scheme \mathcal{D} are *both* secure—and therefore, by Theorem 16 , we can construct a secure public-key quantum money scheme \mathcal{S} .

Theorem 4. *Suppose the ETH holds with constant c . Then for every $\alpha, \beta \in \mathbb{N}$ there exists a $\gamma = O(\alpha + \beta)$ such that*

Figure 11: Grad-CAM visualizations of some sample blocks

C EXTRA MATERIAL FOR SECTION 6 (DISCUSSION & FUTURE WORK)

An example of use of Grad-CAM for visualization of the vision model is in Figure 11.

REFERENCES FOR THE APPENDIX

[LOG⁺19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.