Projeto Final - DocVQA

Luiza Amador Pozzobon

Janeiro de 2021

Abstract

O uso dos *Transformers* trouxe ganhos expressivos para o campo de processamento de linguagem natural nos últimos anos e modelos como o *BERT* e o *T5* ganharam notoriedade pelo seu sucesso em explorar a etapa de pré-treino. Este projeto, portanto, explora diferentes estratégias de pré-treinamento de um modelo *T5* para tarefas de visão e linguagem, onde as entradas são imagens e texto e as saídas são texto, com objetivo final de aplicação no conjunto de dados *DocVQA*. Julga-se que a estratégia ideal ainda não foi encontrada e nenhum dos experimentos trouxe resultados promissores.

1 Introdução

O campo de processamento de linguagem natural (*Natural Language Processing* ou *NLP*) sofreu uma revolução nos últimos anos com a ampla utilização do *Transformer* [8] para resolução de problemas como tradução de texto. O sucesso alcançado deve-se, principalmente, ao mecanismo de auto-atenção utilizado na arquitetura, que torna o modelo capaz de atentar-se a todo o contexto de uma sequência de entrada.

Com o uso da arquitetura citada como base, outros modelos obtiveram sucesso em tarefas de linguagem, como o *BERT* [2], que evidenciou a importância da etapa de pré-treino, e o T5 [5], criado com o intuito de unificar a resolução de tarefas relacionadas a texto em um *framework* texto-para-texto. A potência dos mecanismos de atenção ainda foi pouco explorada para imagens, entretanto, e o modelo que ganhou notoriedade até esse momento foi o *Visual Transformer* [3], que obteve 88,55% de acurácia *top-1* para o *dataset ImageNet*, equivalente ao estado da arte. Esse sucesso traz como possibilidade a não necessidade do uso de convoluções para o processamento de imagens com redes profundas.

Quando se fala de problemas que misturam texto e imagem (*Visual and Language Tasks*, ou *V+L*), ainda não há clareza de qual abordagem é bem sucedida, talvez pelo campo ainda não ter estabelecido um *dataset* base para os testes, mas principalmente porque há duvidas sobre como representar conjuntamente os dados de imagem e de texto. Como abordagens para a tarefa, ressaltam-se

dois trabalhos: o *LayoutLM* [10] e o UNITER (*UNiversal Image-TExt Representation*) [1]. O primeiro utiliza tanto *features* da imagem, obtidos com auxílio de uma rede *Faster-RCNN*, como de *OCR* (*Optical Character Recognition*). O segundo, por sua vez, investe em um pré-treino complexo junto a um modelo *T5*.

Neste trabalho, objetiva-se investigar métodos de pré-treino que poderiam auxiliar na tarefa *V+L* de entendimento de documentos com o conjunto de dados *DocVQA* (*Document Visual Question Answering*) [4]. O *dataset* é composto por cerca de 13 mil imagens e 50 mil perguntas de documentos de diversos segmentos industriais. As perguntas podem ser em relação ao conteúdo textual do documento, mas também sobre elementos de *layout* e disposição espacial das informações, então a utilização apenas do *OCR* ou apenas dos elementos da imagem é, provavelmente, insuficiente para resolução da tarefa.

Sendo assim, faz-se uso de um modelo *T5* para tentativa de resolução da tarefa apresentada, que recebe como entrada, concatenados, tanto os *embeddings* da imagem, quanto os da pergunta relacionada a essa. Para resolução dessa tarefa, julga-se necessário que o modelo não só entenda o *layout* do documento, mas também que tenha noções de entendimento textual. Exploram-se estratégias de pré-treino com enfoque em duas tarefas principais: (1) *OCR end-to-end* do texto e (2) resolução de perguntas e respostas sobre o texto da imagem. A primeira estratégia utiliza o *dataset* SQuAD 2.0 (*Stanford Question Answering Dataset*) [6] para resolução das tarefas citadas em cascata. A segunda estratégia faz uso de um *dataset* sintético gerado a partir dos textos do *WikiText*, com algumas mudanças na arquitetura citada, que conta com adição do OCR da imagem, e das tarefas objetivadas, que são na linha de *VQA* e *MLM (Masked Language Modeling)*.

2 Trabalhos Relacionados

2.1 Transformers

Os *Transformers* [8], redes que permitiram a revolução na resolução de problemas de linguagem natural nos últimos anos, baseiam-se no uso de mecanismos de atenção e na lógica *encoder-decoder*. O mecanismo fundamental dos *Transformers* é a camada de auto-atenção que permite que o modelo tenha acesso ao contexto completo da sequência observada. A camada de auto-atenção é calculada de acordo com a Equação 1, onde Q, K e V são *embeddings* obtidos a partir da sequência de entrada e d_k é um fator de estabilização da equação obtido a partir da dimensão do modelo d_{model}.

$$Auto - Atenção = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

2.2 Text-to-Text Transfer Transformer

A transferência de aprendizagem, ou *Transfer Learning*, quando um modelo é pré-treinado para uma tarefa em que há abundância de dados antes de ser treinado para uma mais específica, mostrou-se eficiente não só no campo de processamento de imagens, mas também no de linguagem natural. Raffel et al. (2019) exploram a técnica no contexto de *NLP* (*Natural Language Processing*) e comparam arquiteturas, objetivos e métodos de transferência em diversas tarefas relacionadas ao entendimento da linguagem. Como resultado, chegam à arquitetura chamada T5 (*Text-to-Text Transfer Transformer*) [5], que é um *framework* texto-para-texto capaz de performar mais de uma tarefa com o mesmo modelo, função de custo, hiperparâmetros, etc., diferenciadas por uma *string* adicionada à sequência de entrada. Na Figura 1 está o esquemático de entrada e saída do T5.

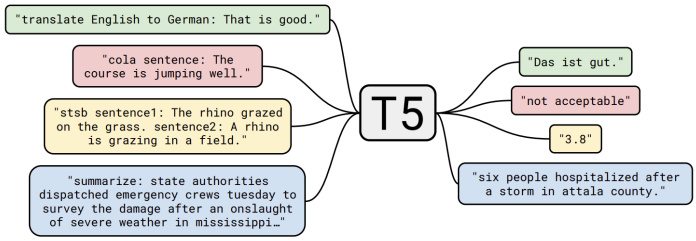


Figure 1: No T5, modelo texto-para-texto, as tarefas são delimitadas pelo prefixo no texto de entrada.

3 Metodologia

Com base nos trabalhos apresentados, como o *BERT* e o *T5*, hipotetiza-se que uma etapa de pré-treino bem desenvolvida é crucial e suficiente para facilitar a resolução de tarefas complexas como as do *DocVQA*. Sendo assim, há um enfoque maior na estratégia de pré-treino e manteve-se o modelo base como um T5 quase inalterado. A seguir, então, são apresentadas a arquitetura e a lógica de pré-treino utilizadas, além de uma análise exploratória preliminar que guiou o trabalho.

3.1 Exploração do DocVQA

Conduziu-se uma breve análise exploratória do conjunto de dados *DocVQA* para compreender a natureza das imagens, principalmente, do texto presente nas imagens, das perguntas e das respostas. Todos os dados foram colhidos a partir do conjunto de validação. Na Tabela 1 estão alguns dos dados observados, bem como na Figura 2 está a distribuição da proporção de $altura/largura$ das imagens. Ainda, nota-se que as perguntas são no geral relacionadas ao texto

do documento ou à interação de elementos não textuais com o texto. A seguir alguns exemplos de perguntas do conjunto de dados:

- "What is the date on the 'form'?"
- "What is the page number?"
- "In which year did Michael receive his PhD?"
- "What is the number written in the credit field ?"
- "What is written within the circle at the bottom of the" page?

Table 1: Especificações do conjunto de validação do *dataset DocVQA*.

	Média	**Mediana**	**Máximo**
Quantia de palavras no OCR	199	159	1176
Altura da imagem (px)	2084	2207	7184
Largura da imagem (px)	1776	1706	6921

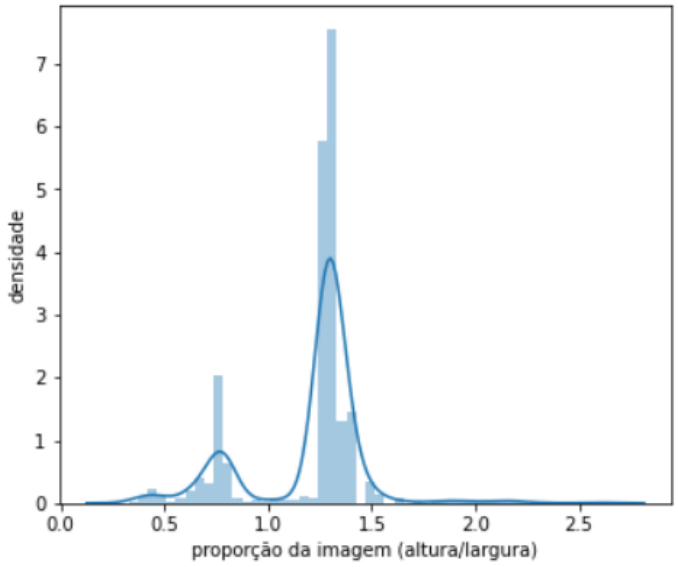


Figure 2: Distribuição da proporção de altura por largura das imagens do dataset DocVQA.

3.2 Arquitetura

Como arquitetura principal e preliminar, fez-se uso de modelo *T5-base* completo que recebe como entrada, concatenados, os *embeddings* da imagem e os da pergunta relacionada a essa.

Os *embeddings* de imagem são obtidos a partir de uma camada convolucional de 384 canais, com tamanho de *kernel* 28, *stride* 24 e *padding* 4, que gera *patches* de 32×24 pixels. Esses *patches* são então redimensionados para obtermos *embeddings* de tamanho $N \times 384 \times d_{model}$, sendo N o tamanho do *batch* e $d_{model} = 768$ para o caso do *T5-base*. De acordo com a exploração do *dataset* *DocVQA* apresentada, definiu-se o formato de $3 \times 768 \times 576$ para as imagens de entrada, visto que no conjunto de dados as imagens no geral possuem uma altura maior que a largura.

As perguntas relacionadas às imagens tem tamanho de sequência máximo de 20 e seus *embeddings* são obtidos a partir daqueles já conhecidos pelo modelo (pela camada *shared* do modelo da biblioteca utilizada, *HuggingFace* [9]). Somam-se *embeddings* posicionais aos *embeddings* de texto. A seguir, concatenam-se os *embeddings* das imagens aos das perguntas, resultando em *embeddings* finais de tamanho $N \times 404 \times 768$. Na estratégia de pré-treino com o dataset *WikiText*, houve ainda a concatenação do *OCR* da imagem, com tamanho de sequência de 128 *tokens*, o que resultou em uma entrada máxima no T5 de $N \times 532 \times 768$. Na Figura 3 está a arquitetura proposta.

3.3 Estratégias de pré-treino

Conforme mencionado, a hipótese é que para o sucesso da tarefa de perguntas e respostas de documentos no conjunto de dados *DocVQA* o modelo necessite de habilidades que vão desde a leitura do texto (*OCR*), até o entendimento da disposição das informações na página e no próprio texto. Sendo assim, estruturaram-se duas estratégias de pré-treino, uma utilizando o conjunto de dados *SQuAD* 2.0 e outra com o *WikiText*, mas com foco em duas tarefas: (1) leitura do texto da imagem (*OCR end-to-end*) e (2) *Visual Question Answering*, ou perguntas e respostas sobre o texto da imagem.

4 Conjuntos de dados

4.1 SQuAD 2.0

O *Stanford Question Answering Dataset* foi construído para auxilio na resolução de tarefas de entendimento de linguagem natural, mais especificamente para a de perguntas e respostas acerca de um fragmento textual. A versão 2.0 combina os dados presentes no *SQuAD* 1.0 [7] com mais de 50.000 amostras cuja resposta para a pergunta não está presente no texto, resultando em mais de 150.000 amostras no conjunto de dados. Conforme os autores, para um sistema performar bem no *SQuAD* 2.0, é necessário que esse não só seja capaz de en-

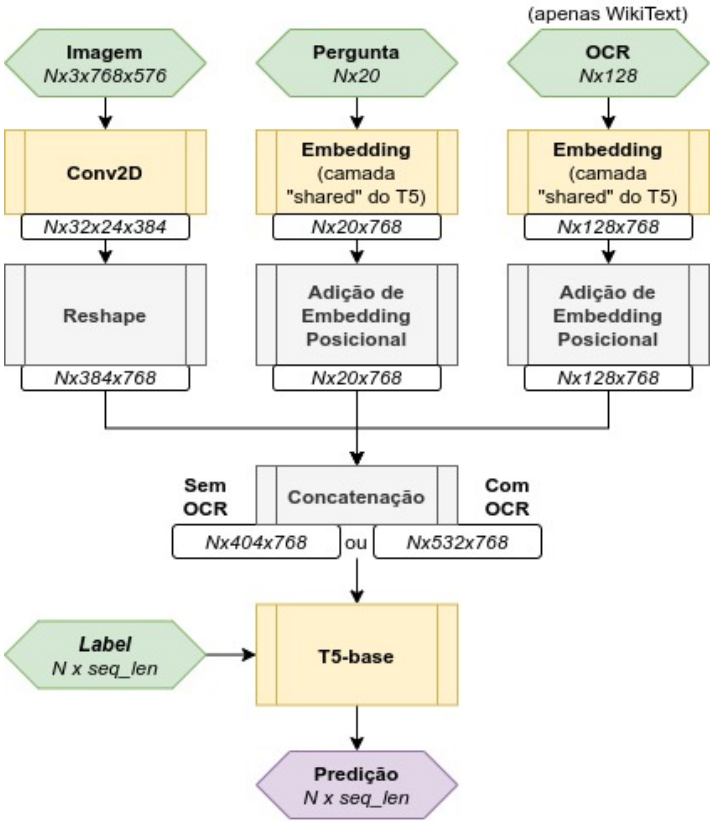


Figure 3: Arquitetura dos modelos e fluxo dos dados nos experimentos.

contrar a resposta no parágrafo fornecido, mas também que consiga discernir se a resposta está presente nele ou não [6].

Para ambas as tarefas, constrói-se um *dataset* de imagens sintéticas com o texto disposto em uma imagem de fundo branco. O ponto âncora do texto varia horizontalmente de 10 a 30 pixels e verticalmente de 10 a $0.4 \times image\_height$ pixels e a fonte, Arial, pode variar de 10 a 16 pt. Em uma mesma época de treino, os textos das imagens se repetem $k = 1, 2, ...$ vezes, com variações de tamanho de letra e ancoragem, visto que cada imagem possui k perguntas associadas a elas.

Na tarefa de OCR, o texto da imagem e aquele previsto pelo modelo tem tamanhos máximos de 128 *tokens*. Na tarefa de perguntas e respostas sobre o texto, mantiveram-se as mesmas configurações da tarefa anterior, mas adicionou-se um ruído gaussiano de -30 a 30 na imagem. Nesta etapa, espera-se que o modelo já tenha a habilidade de ler o texto da imagem e que agora passe a "entendê-lo" suficientemente bem para responder as perguntas propostas pelo conjunto de dados.

Table 2: Perguntas geradas no dataset sintético e mascarado de perguntas e respostas do WikiText.

Perguntas presentes no dataset	**Tradução**
Which are all the masked tokens?	Quais são todos os tokens mascarados?
What is the first masked token?	Qual é o primeiro token mascarado?
What is the word after/before the first/last covered token?	Qual é a palavra após o primeiro/depois do último token mascarado?
How many masked words are there?	Quantas palavras mascaradas existem?
What is the first/last word of the sentence?	Qual é a primeira/última palavra da sentença?
What is the whole sentence?	Qual é a sentença completa?

4.1.1 WikiText

Foram selecionadas 1 milhão de amostras do *WikiText* para o conjunto de treino, mil para validação e mil para teste. A tarefa foi estruturada a partir da união de duas: (1) o *Masked Language Modeling (MLM)* conforme apresentado pelo *BERT* e explorado similarmente pelo *T5*, onde é feita a predição das palavras mascaradas de uma sequência de texto [2, 5]; e (2) a tarefa de perguntas e respostas (VQA).

Das 128 palavras máximas da sequência de entrada, 15% são escolhidas aleatoriamente para o mascaramento. Dessas, 10% podem ser inalteradas (não mascaradas), 10% trocadas por um *token* aleatório e 80% mascaradas com "[MASK i]", onde i vai de 0 até o número total de palavras mascaradas na sequência menos um. Produzem-se perguntas e respostas acerca do texto modificado, que é alocado em uma imagem com ruído gaussiano de -30 a 30 e âncora de texto variável. Para cada amostra, um par de pergunta e resposta é sorteado aleatoriamente e o conjunto de dados tem tarefas relacionadas tanto a *OCR* quanto a *Visual Question Answering*.

Na Tabela 2 observam-se as possibilidades de perguntas do *dataset* sintético e no Apêndice A estão os exemplos do mascaramento, das perguntas e respostas geradas e uma amostra do conjunto de dados.

5 Experimentos

A seguir são expostas as métricas utilizadas para validação dos modelos, bem como os resultados obtidos para cada método de pré-treino. Os hiperparâmetros dos experimentos são visualizados no Apêndice A.

5.1 Métricas

Duas métricas foram utilizadas para a avaliação dos modelos, conforme apresentado junto ao conjunto *SQuAD* 1.0 [7]. Essas ignoram pontuações e artigos da língua inglesa (*a, an, the*).

Exact Match. Métrica que mede a percentagem de predições que correspondem exatamente àquelas do conjunto de *labels*.

Pontuação F1. As predições e os *labels* são tratados como uma bolsa de *tokens* e a pontuação F1 é calculada conforme a Equação 2.

$$F1 = 2\frac{precision \cdot recall}{precision + recall} \tag{2}$$

5.2 Pré-treino com o SQuAD 2.0

Idealizou-se que o pré-treinamento para a tarefa de *OCR* seria de grande valia para a tarefa de *VQA*, especialmente se utilizando o mesmo conjunto de dados, que por sua vez seria crucial para resolução do DocVQA. Sendo assim o segundo pré-treino dessa estratégia é iniciado a partir do melhor *checkpoint* do primeiro.

5.2.1 Tarefa de OCR

Na tarefa de *OCR*, o modelo treinou por quase 5 dias e obteve pontuação F1 de validação máxima de 20% e *exact match* sempre zero, conforme Figura 4. Observando amostras de predições de validação, nota-se que o modelo provavelmente memorizou algumas amostras de texto, as quais devem ter aumentado a pontuação F1 (mas não completamente, pois não conseguiu nenhuma pontuação na métrica de *exact match*), enquanto que para a maioria das outras amostras o texto predito nada tem a ver com o que está na imagem.

O modelo foi capaz de realizar um *overfit* para um teste preliminar com dois *batches* e atingiu pontuação F1 de 96%. Devido a esse resultado, julga-se que a falta de sucesso no treino completo possa estar nos parâmetros associados, no tempo de treino ou no *dataset* em questão, que é parcialmente sintético.

5.2.2 Tarefa de VQA

Mesmo com a falta de sucesso da tarefa anterior, foi realizado o experimento de *VQA* para o *dataset SQuAD 2.0*. Fez-se uso do melhor *checkpoint* de *OCR* para treinar o modelo por mais um dia e, embora a perda tenha reduzido bruscamente para níveis muito inferiores aos da etapa anterior, nenhuma das métricas acompanhou essa mudança e ambas permaneceram próximas de zero de acordo com a Figura 5.

5.3 Treino no DocVQA

Como o pré-treino para a tarefa de *VQA* não resultou em pontuações maiores que zero, o treino no *DocVQA* foi realizado a partir do melhor *checkpoint* de pré-treino de *OCR*. Entretanto, conforme o esperado devido às baixas pontuações, após quase dois dias de treino, o modelo não foi bem sucedido na tarefa final, atingindo no máximo 1,5% de pontuação F1 de validação e o *exact match* permaneceu próximo de zero, como na Figura 6.

5.4 Pré-treino com o WikiText

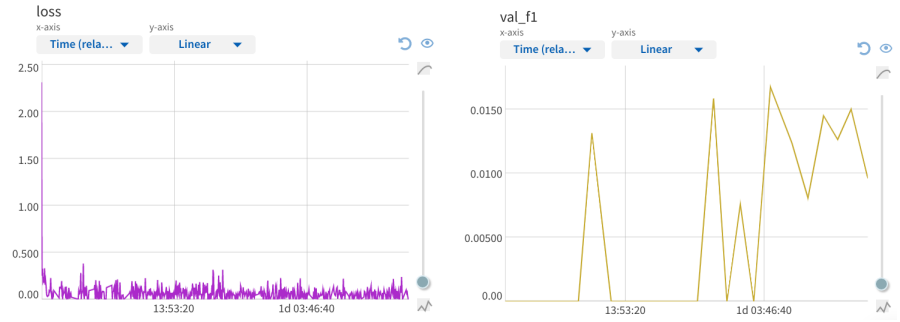
O pré-treino com o conjunto *WikiText* utiliza a arquitetura com adição do *OCR* do texto e faz uso das imagens, perguntas e respostas sintéticas produzidas

especificamente para esse experimento, conforme apresentado. Na Figura 7 observa-se a progressão de perda, de pontuação F1 e de *Exact Match*. A primeira métrica teve seu máximo em 1,8% e a segunda em 1,5%.

Diferentemente das outras tarefas, aqui a perda não regrediu de forma agradável e permaneceu ruidosa durante todo o período de treino. O modelo, por sua vez, sempre retornava a mesma resposta repetida n vezes, o que explica a similaridade dos gráficos das métricas. É necessária uma investigação mais criteriosa do *dataset* já que é onde estão, provavelmente, os erros e *bugs* introduzidos ao sistema.

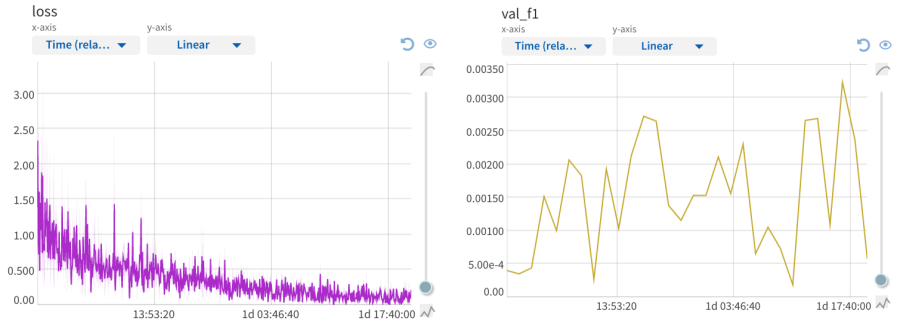
(a) Loss (b) Pontuação F1

Figure 4: Gráficos de perda e pontuação F1 de validação durante o pré-treino para a tarefa de OCR com o dataset SQuAD 2.0.



(a) Loss (b) Pontuação F1

Figure 5: Gráficos de perda e pontuação F1 de validação durante o pré-treino para a tarefa de VQA com o dataset SQuAD 2.0.



(a) Loss (b) Pontuação F1

Figure 6: Gráficos de perda e pontuação F1 de validação durante o treino no dataset DocVQA.

(a) Loss (b) Pontuação F1 (c) Exact Match

Figure 7: Gráficos de perda, pontuação F1 e exact match de validação durante o treino no dataset WikiText de imagens sintéticas e com mascaramento.

6 Conclusão

Neste relatório foram apresentados experimentos de pré-treino para tentativa de resolução da tarefa de perguntas e respostas acerca dos documentos do *dataset DocVQA*. Embora nenhuma das tentativas tenha sido bem sucedida, ainda é de entendimento da autora a necessidade de encontrar a(s) tarefa(s) de pré-treino corretas para a resolução da tarefa. De acordo com o que vem sendo apresentado na literatura, essa etapa é crucial para a resolução de problemas complexos, como é o caso do *DocVQA*, que une leitura e entendimento de texto a entendimento de *layout* e de disposição de elementos não textuais.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Docvqa: A dataset for vqa on document images. *arXiv preprint arXiv:2007.00398*, 2020.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[6] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

[10] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020.

A Apêndice

A.1 VQA Sintético com WikiText

Nas Figuras 8 e 9 estão, respectivamente, uma amostra do *dataset* sintético gerado e as possíveis perguntas e respostas geradas para um exemplo simples de sequência de texto.

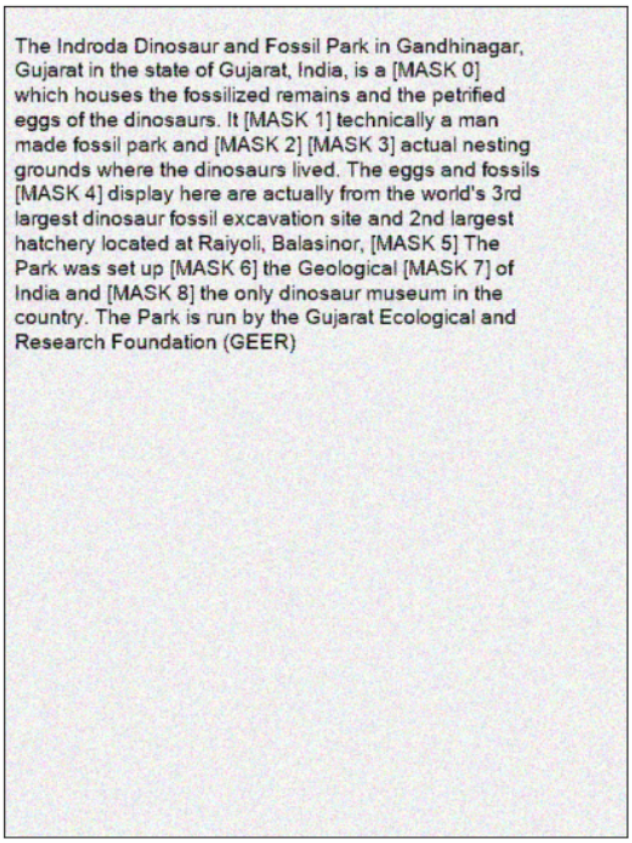


Figure 8: Exemplo de uma amostra do conjunto de dados sintético de perguntas e respostas gerado com o WikiText. As imagens geradas para os experimentos do SQuAD são similares a esta, com ou sem o ruído de fundo, mas sem o mascaramento das palavras.
Pergunta: "Which are all the masked tokens?"
Resposta: "park, is, not, the, on, Gujarat., by, Survey, is"

```
Original text and mask:  ['oi' 'como' 'está' 'você'] [0 1 0 1]
{
  "Which are all the masked tokens?": "como, você",
  "What is the first masked token?": "como",
  "What is the word after the first covered token?": "está",
  "How many masked words are there?": "2",
  "What is the first word of the sentence?": "oi",
  "What is the whole sentence?": "oi como está você"
}
```

Figure 9: Exemplo de uma sequência de texto tokenizada, a máscara e as possibilidades de perguntas e respostas geradas.

A.2 Hiperparâmetros

Os hiperparâmetros para todos os experimentos quase não foram modificados e seus valores são observados na Tabela 3.

Table 3: Hiperparâmetros utilizados nos experimentos.

Modelo base	t5-base
Otimizador	Adafactor
Batch Size - Treino	2
Batch Size - Validação	2
Tamanho de sequência de saída (Squad-OCR)	256
Tamanho de sequência de saída (Squad-VQA)	256
Tamanho de sequência de saída (Squad-DocVQA)	128
Tamanho de sequência de saída (WikiText)	128
Tamanho de sequência das perguntas	20
Learning Rate	5e-4 ou 1e-4