Judging Book by Its Cover and Content

1st Beili Yin
Department of Computer Science
Lakehead University
Thunder Bay, Canada
byin@lakeheadu.ca

2nd Dr. Thiago E. A. de Oliveira
Department of Computer Science
Lakehead University
Thunder Bay, Canada
talvesd@lakeheadu.ca

***Abstract*—The old saying "judging a book by its cover" still holds true when selecting books either in retailers or online book stores. Along with book cover the readers also pay significant amount of attention in book description, which is the concise abstraction of the book content. In this project, Vision Transformer [7] and Generative Pre-trained Transformer 2 (GPT-2) [8] are utilized to learn from a non-biased book cover dataset, and the fine-tuned models are used to predict the reader's preference from randomly scraped books in corresponding genres within the dataset. The selected books are demonstrated on a customized web front-end.**

***Index Terms*—Recommendation System, Natural Language Processing, Computer Vision, Deep Learning, Vision Transformer.**

I. Introduction

Book covers are usually the very first impression to its readers and they often convey important information about the content of the book. Book genre classification based on its cover would be utterly biennial to many modern retrieval systems, considering that the complete digitization of books is an extremely expensive task [1]. At the same time, it is also an extremely challenging task due to the following reasons: First, there is a wide variety of book genres, many of which are not concretely defined and often intersect. Second, book covers, as graphic designs, vary in many different ways such as colors, styles, textual information, etc., even for books of the same genre. Third, book cover designs may vary due to many external factors such as country, culture, target reader populations, etc. With the growing competitiveness in the book industry, the book cover designers push the cover designs to its limit in the hope of attracting sales. The cover-based book classification systems become a particularly exciting research topic in recent years.

Considering judging books solely by their covers is unrealistic in consumers' preference and shopping behaviour. After the title and the book cover, the description is the most important book marketing material. The book description goes most prominently on the back cover, and the top of the online book store page (such as Amazon, it's below the price and above the book reviews). It's crucial it be compelling, because readers also make buying decisions from the book description. In other words, the book description is the pitch to the readers about why they should buy the book.

A genre is the information that a brick-and-mortar bookstore clerk uses to figure out where to put the book in their bookstore. It's the information an electronic retailer uses to make the book easy to find in their electronic databases online. And it's the information that librarians use to figure out where to put the book in their library. A book within a specific genre can reflect on the color palette and design style as well as strong association with high frequency words that differentiate itself from books in another genre.

With these in mind, building a recommender based on information extracted from both book cover and description

can yield more satisfactory result than judging books by either of these factors alone.

II. Related Workds

A. LetNet and AlexNet

Researches on the visualization of book covers are on-going efforts. Brian Kenji Iwana et al. [2] sifted through a wide variety of book covers and styles, overcame the difficulty of non-scripting and misleading covers, to determine if genre information could be obtained based on visual aspects of book covers.

The book covers in the dataset they present lack distinguishable features. For the convolution neuro networks of their choice, which are LetNet and AlexNet, has to rely on color alone to classify covers. In general, most cover images have either a strong activation toward a single class or are ambiguous and could be part of many classes at once.

All of the images are resized to fit 227px by 227px by 3 color channels for the input of the AlexNet and 56px by 56px by 3 color channels for LeNet. The pre-trained AlexNet with transfer learning resulted in a test set Top 1 classification accuracy of 24.7%, 33.1% for Top 2, and 40.3% for Top 3 which are 7.4, 5.0, and 4.0 times better than random chance respectively. As comparison, using the modified LeNet, we had a Top 1 accuracy of 13.5%, Top 2 accuracy of 21.4%, and Top 3 accuracy of 27.8%. The AlexNet performed much better on this dataset than the LeNet.

The classification of books based on the cover image is in general, a difficult task. The paper reveals that many books have cover images with few visual features or ambiguous features causing for many incorrect predictions. While uncovering some of the design rules found by the CNN, books can have misleading covers. In addition, because books can be part of multiple genres, the CNN had a poor Top 1 performance.

Three factors are vital to the inference: colors, objects and texts. With limited features, color can associate with many book genres. Common objects are identifiable since the structures and layouts of each genre have impacts on classification. The quality and font properties of the cover texts are captured by AlexNet indicates that there are correlations between text styles and book genres.

B. Cold Start Problem of a Recommendation System

With no user history and preference, a recommendation system often encounters cold start problem. Data sparsity is also a negative drawback that drag down a recommender's performance. Madhusmita Kalita and Dr. Thiago E. A. de Oliveira [5] specifically designed a hybrid model to combat the issues. Unsupervised clustering method, i.e., K-Means were used to extract dominate hues from the book covers, which were passed to convolution autoencoder for visual feature generation. Finally, Gaussian Mixture Model is designated for transferring visual feature into possibilities. The final performance of the hybrid model saw 4.5% improvement than the baseline model.

C. Multi-modal information Issue

The issues with identifying analogous information to customer's preferences are that it is often multi-modal data in nature. Visual messages rarely are presented alone, they often accompanied by texts, sounds, streamline medias, or other modalities. Recent advances in self-attention neural networks such as transformers, have led to progressive improvements in both Natural Language Processing and Computer Vision tasks.

D. Researches on Vision Transformer

Before Vision Transformers, or ViT began to show promising results on image classification tasks in terms of both speed and accuracy, the Convolutional Neural Networks almost dominates the area. Despite Convolutional Networks' promising performance, their architectures generally have limitation in modeling explicit long-range spatial relations. However, unlike the benefit of deep layers can bring to CNNs, as the transformer goes deeper, the attention maps gradually become similar and even much the same after certain layers. Daquan Zhou et al. [3] proposed re-attention blocks between each attention layers to diversity their cosine similarity and

addressed the self-attention layers saturation problem. Furthermore, the results show that training deep vision transformer can achieve satisfactory performance without extra datasets and augmentation policies.

Another issue addressed by Ze Liu et al. [4] was the overfitting and gradient exploding/vanishing problem that haunts the general transformer architecture. By constructing a hierarchical representation by starting from small-sized patches (outlined in gray) and gradually merging neighboring patches in deeper Transformer layers, the Swim Transformer noticeably surpass the counterpart DeiT architectures with similar complexities; Compared with RegNet and EfficientNet, the Swin Transformer achieves a slightly better speed-accuracy trade-off on medium datasets such as ImageNet-1K. On large object detection tasks (COCO) and semantic segmentation (ADE20K) Swim Transformer had higher performance trade-offs than previous state-of-the-art models.

E. Deep Multi-modal Networks on Book Covers

Chandra Kundu and Lukun Zheng [6] took the advantage of both CNNs and LTSM into two multi-models to classify images and texts directly extracted from the book covers.The texts are extracted from the book covers using Google Cloud Vision API. Firstly, the book covers were solely tested with image-only models and text-only models. LeNet, AlexNet, VGGNet-16, MobileNet-V1, MobileNet-V2 , Inception-V2, and ResNet-50, LSTM for texts. Then comparing image-only model performances with that of the multi-models.

Texts were extracted by Google Cloud Vision API, which were directly captured texts from book cover images. In addition to these image-based models, they also evaluated two text-based models for this task. One is recurrent neural networks (RNN) with Long Short-Term Memory (LSTM). It allows the network to accumulate past information and thus be able to learn the long-term dependencies which are very common in textual data. The other one is Universal Sentence Encoder. It is a pre-trained sentence embedding and designed to leverages the encoder from Transformer which is a multi-attention head that helps the model "attend" to the relevant information.

Among the image-based models, ResNet-50 achieved the best performance with a top-1 accuracy of 19.6% and a top-3 accuracy of 49.0%. The RNN-LSTM model achieved a top-1 accuracy of 41.5% and a top-3 accuracy of 61.6%. The DCCA concatenation model achieved a top-1 accuracy of 48.9% and a top-3 accuracy of 72.3%

III. Methodology: Transformer Hybrid

A. Dataset Preparation

In this project, we used Book Covers Dataset collected by Luka Anicin. This dataset contains 33 book categories and each contains close to 1000 images. It is balanced judging by the book cover quantity of each class.

However, this dataset has only half the samples per class, and 3 more classes in comparison rendering the classification task is expected to be more difficult than the researches done in the related works section.

Along with the images comes with all meta information for every book cover: image - URLs of book covers. Use this cover to download images yourselves if you need. name - Title of a book. author - Author of a book. format - Physical format of a book (i.e., paperback) bookdepositorystars - Book's rating found on the bookdepository.com (NOTE: Due to difference between scraping and download time of the dataset, this information might be different from one on the website) price - Book's current price found on the bookdepository.com (NOTE: Due to difference between scraping and download time of the dataset, this information might be different from one on the website) currency - Currency of prices found in the dataset. old_price -Book's old price (if exists) found on the bookdepository.com (NOTE: Due to difference between scraping and download time of the dataset, this information might be different from one on the website) isbn -ISBN number of a book. category -Category of a book found on the bookdepository.com img_paths -Book's cover local path (after scraping).

Unfortunately, the meta information does not include book description. Google Books API is utilized to complete the dataset. Book descriptions are capture based on their titles and authors. For those books that do not have description available in Google Cloud, their titles serve as substitution since a book title often carries distinguishable words from one genre to another.

The dataset is split into training set and validation set, for the purpose of performance evaluation. The test set is random books captured by Google Books API. The ratio of training and validation is 0.85:0.15. Samples of the dataset and test data are displayed in Fig1 and Fig2.

For the purpose of this project, we deem it to be beneficial to demonstrate succinct prediction result on our customized web front-end, only book title, description, category and image are being trained by the proposed Transformer models and displayed by the web page.

B. Vision Transformer (base-sized model)

The Vision Transformer (ViT) [7] is a transformer encoder model (BERT-like) pre-trained on a large collection of images in a supervised fashion, namely ImageNet-21k, at a resolution of 224x224 pixels.

Images are presented to the model as a sequence of fixed-size patches (resolution 16x16), which are linearly embedded. One also adds a [CLS] token to the beginning of a sequence to use it for classification tasks. One also adds absolute position embeddings before feeding the sequence to the layers of the Transformer encoder. Note that this model does not provide any fine-tuned heads, as these were zero'd by Google researchers. However, the model does include the pre-trained pooler, which can be used for downstream tasks (such as image classification).

By pre-training the model, it learns an inner representation of images that can then be used to extract features useful for downstream tasks: if you have a dataset of labeled images for instance, you can train a standard classifier by placing a linear layer on top of the pre-trained encoder. One typically places

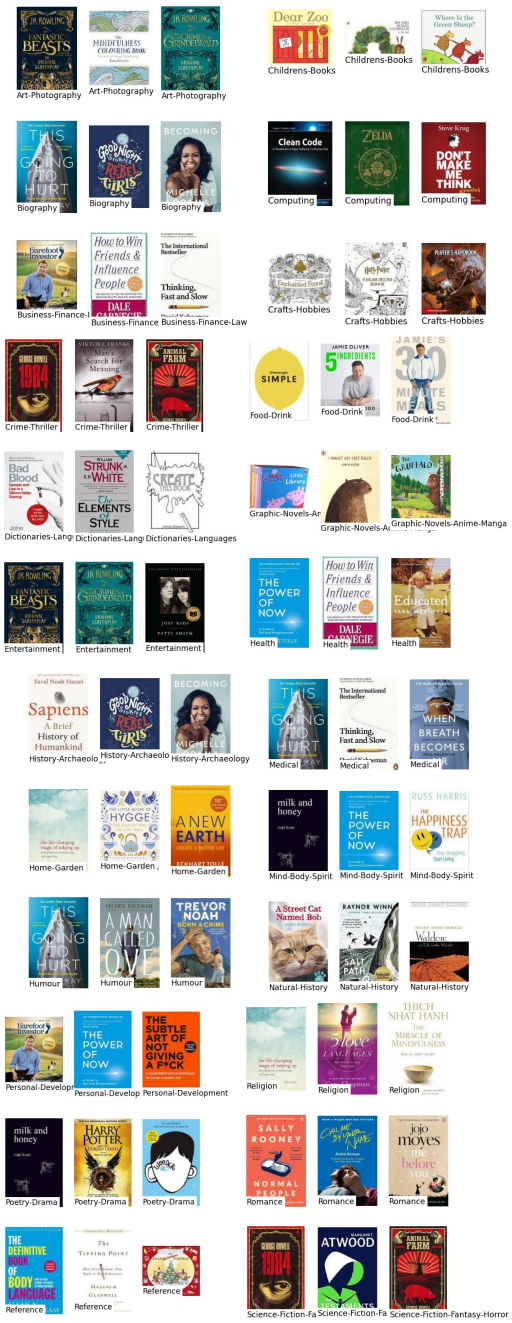

Fig. 1. Book Cover Dataset Samples

a linear layer on top of the [CLS] token, as the last hidden state of this token can be seen as a representation of an entire image. Shown in 3

The best results are obtained with supervised pre-training, which is not the case in NLP. The authors also performed an experiment with a self-supervised pre-training objective, namely masked patched prediction (inspired by masked lan-

Fig. 2. Test Set Samples

guage modeling). With this approach, the smaller ViT-B/16 model achieves 79.9% accuracy on ImageNet, a significant improvement of 2% to training from scratch, but still 4% behind supervised pre-training.

Vision Transformer can achieve state-of-the-art performance when training from scratch on large database. The dataset considered can only be identified as small. To fully facilitate

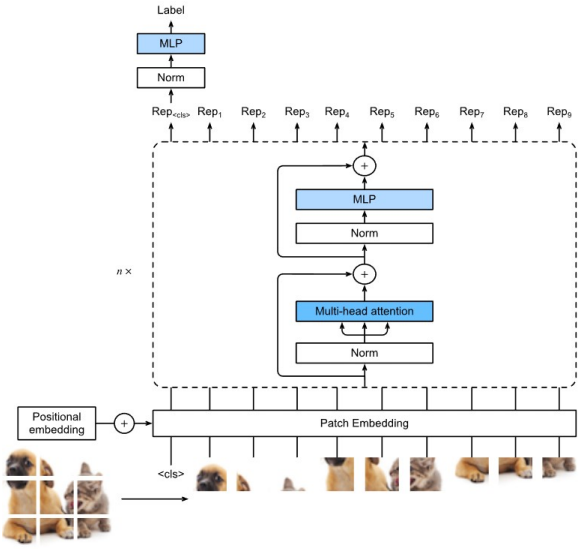

Fig. 3. Vision Transformer Architecture

its potential in both accuracy and training speed, we choose ViT base-sized model to fine-tune upon. The model for this project is pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224, which is processed by pre-trained ViT feature extractor.

C. GPT-2

GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages [8]. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data.

GPT-2 displays a broad set of capabilities, including the ability to generate conditional synthetic text samples of unprecedented quality, where we prime the model with an input and have it generated a lengthy continuation. In addition, GPT-2 outperforms other language models trained on specific domains (like Wikipedia, news, or books) without needing to use these domain-specific training datasets.

The Layer Normalization from the Transformer is moved up to the input of each sub-block (pre-norm) 4.

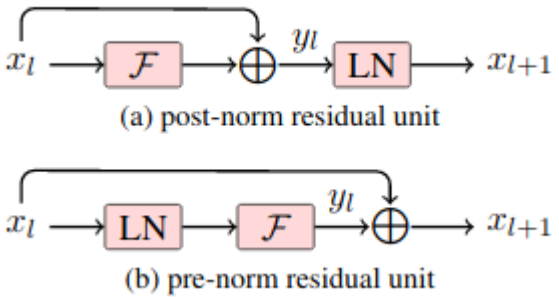

Fig. 4. Transformer Layer Normalization

An additional Layer Normalization is added at the end of the final self-attention block. The weights of the residual layers are scaled by a factor 1/N (at initialization), where N is the number of residual layers. The vocabulary and the maximum sequence length are expanded to 50,257 and 1024 respectively.

IV. Implementation

Vision Transformer fine-tuning parameters: Random seed: 42 Learning rate: 2e-5 Precision: 16 Max epochs: 4 Image resize: 224 X 224

GPT-2 parameters: Random seed: 123 Batch size: 8 Max length: 200 Max epochs: 4

Note that the local setup for the experiment has memory limitation, thus the corpus length GPT-2 learns from the book description is also being reduced to 200. Both models reach fine-tuning threshold at the 4th epoch. To avoid over-training, the performance is evaluated at max epoch being set to 4. For experiment result consistency, both models were trained in local personal computer environment. The setup is as follows:

GPU: NVIDIA GeForce RTX 3070

RAM: 16.0 GB

CPU: Intel(R) Core(TM) i5-10600KF

SSD: WDC WDS100T2B0C-00PXH0

V. Experiment Result

The pre-trained Vision Transformer base-sized model with transfer learning resulted in the validation set accuracy: 25.68%, test set accuracy: 19.68%. The pre-trained GPT-2 model with transfer learning resulted in the validation set accuracy: 38.08%, test set accuracy: 37.12%. More detailed evaluation report is demonstrated in TableV and Fig 4.

class	**precision**	**recall**
Art-Photography	0.16	0.08
Biography	0.15	0.14
Business-Finance-Law	0.35	0.45
Childrens-Books	0.35	0.39
Computing	0.51	0.62
Crafts-Hobbies	0.5	0.49
Crime-Thriller	0.39	0.43
Dictionaries-Languages	0.48	0.47
Entertainment	0.42	0.58
Food-Drink	0.58	0.63
Graphic-Novels-Anime-Manga	0.65	0.61
Health	0.24	0.22
History-Archaeology	0.32	0.43
Home-Garden	0.46	0.4
Humour	0.29	0.29
Medical	0.33	0.39
Mind-Body-Spirit	0.37	0.36
Natural-History	0.41	0.52
Personal-Development	0.19	0.19
Poetry-Drama	0.29	0.32
Reference	0.24	0.14
Religion	0.43	0.43
Romance	0.44	0.5
Science-Fiction-Fantasy-Horror	0.41	0.39
Science-Geography	0.3	0.24
Society-Social-Sciences	0.04	0.01
Sport	0.56	0.58
Stationery	0.44	0.51
Teaching-Resources-Education	0.39	0.44
Technology-Engineering	0.37	0.32
Teen-Young-Adult	0.39	0.32
Transport	0.66	0.61
Travel-Holiday-Guides	0.47	0.49
accuracy	0.39	0.39
macro avg	0.38	0.39
weighted avg	0.38	0.39

TABLE I

GPT-2 Validation Result

VI. Prediction Demonstration

The purpose of this project is to recommend book based on the dataset of our choice. The prediction of both transformers can significantly narrow down the books from randomly scrapped books from Google Cloud API. Here we use the intersection of both model prediction and displayed the final results in a customized web front-end. Since no backend was developed the, front end will need to load the prediction data manually. The web page is shown in Fig5.

VII. Discussion and Conclusion

In this project, we present a recommender of both vision and corpus transformer to predict the reader's book preference based on book cover and description which associated with genre. The experiment shows it is possible to learn the relationship between both factors and book genres. A new dataset is experimented which adds variety and difficulty to the classification task. The test dataset is randomly selected books for the purpose of simulating real recommendation rules.

For Vision Transformer base-sized model, the baseline accuracy at 25.68% outperforms the AlexNet (24.7%) [2] with half samples per class and 3 more classes in total. The test data which has more outlier can still yield reasonable prediction result.

For GPT-2, the validation set accuracy at 38.08% comes close to the test set accuracy at 37.12%. The text-based transformer model proofs to be more robust than the vision-based transformer model when it comes to deciding book genre. Vision Transformer can work as a supplement when it comes to book recommender system when corpus learning models are also considered.

Future research can be put into implementing the advanced Vision Transformer [10] and the latest GPT variant (GPT-3 for now [9]). Along with the matured ViT studies, attention-based autoencoder can also be incorporated into book cover image preprocessing and optimizing. The dataset can be expanded and the missing book description being completed. We hope a more stable recommender will come to in fruition to better represent the reader's preference.

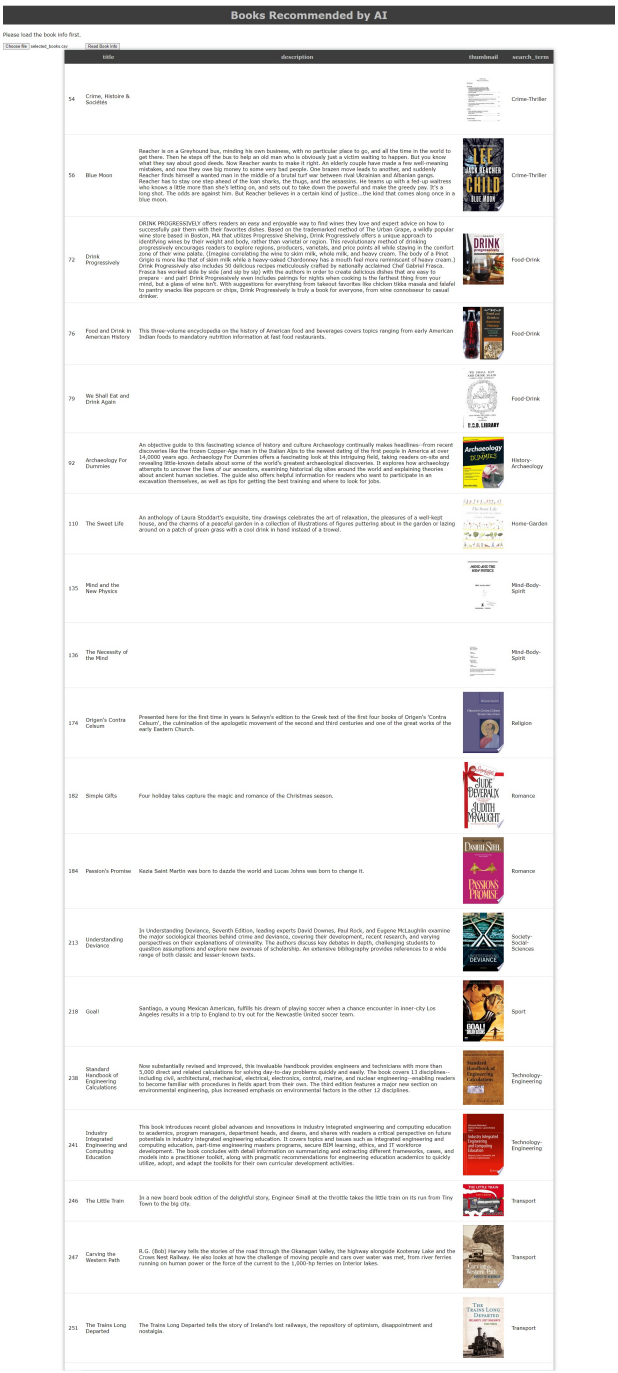

Fig. 5. Book Recommender Front End

VIII. Acknowlagement

All book cover images are copyright https://www.bookdepository.com/ and Google Books Cloud.

The display of the images is transformative and are used as fair use for academic purposes.

The book cover database is available at https://www.kaggle.com/datasets/lukaanicin/book-covers-dataset.

REFERENCES

[1] Grellet F, Francoise G. Developing reading skills: A practical guide to reading comprehension exercises. Cambridge university press; 1981 Sep 30.

[2] Iwana BK, Rizvi ST, Ahmed S, Dengel A, Uchida S. Judging a book by its cover. arXiv preprint arXiv:1610.09204. 2016 Oct 28.

[3] Zhou D, Kang B, Jin X, Yang L, Lian X, Jiang Z, Hou Q, Feng J. Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886. 2021 Mar 22.

[4] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030. 2021 Mar 25.

[5] Madhusmita Kalita, Dr. Thiago E. A. de Oliveira. A Hybrid Book Recommendation Model incorporating Visual Cues with User-Item Interaction

[6] Kundu C, Zheng L. Deep multi-modal networks for book genre classification based on its cover. arXiv preprint arXiv:2011.07658. 2020 Nov 15.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.

[8] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019 Feb 24;1(8):9.

[9] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[10] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. ACM computing surveys (CSUR). 2022 Sep 14;54(10s):1-41.