DATA2VEC IN IMAGE CAPTIONING: Bridging the Gap Between Images and Descriptive Text

By

Nalla Amulya (420211)

Veeramalla Sumukh (420249)

Kondapalli jagadeesh (420149)

Under the supervision of

Dr. Nagesh Bhattu Sristy

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH

TADEPALLIGUDEM-534101,INDIA

2023

DATA2VEC IN IMAGE CAPTIONING: Bridging the Gap Between Images and Descriptive Text

Submitted in the partial fulfillment of the requirements

of the degree of

Bachelor of Technology

By

Nalla Amulya (420211)

Veeramalla Sumukh (420249)

Kondapalli jagadeesh (420149)

Under the supervision of

Dr. Nagesh Bhattu Sristy

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH

TADEPALLIGUDEM-534101,INDIA

2023

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Nalla Amulya	Veeramalla Sumukh	Kondapalli Jagadeesh
420211	420249	420149
Date:	Date:	Date:

Department of Computer Science and Engineering

NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH

Certificate

It is certified that the work contained in the thesis titled "DATA2VEC IN IMAGE CAPTIONING: Bridging the Gap Between Images and Descriptive Text" by "Nalla Amulya, bearing Roll No: 420211" , "Veeramalla Sumukh, bearing Roll No: 420249" and "Kondapalli Jagadeesh, bearing Roll No:420149" has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Dr. Nagesh Bhattu Sristy

Computer Science and Engineering

N.I.T.Andhra Pradesh

May 2023

Abstract

Image captioning is a challenging task in computer vision and natural language processing that involves generating a text describing the content in the image. therefore, this requires the understanding of both image content and textual context. Traditional image captioning models faced difficulty to effectively capture the complex relationships between visual and text features. Those approaces also struggled to integrate these relationships and hence generated some non sensible captions.Our work, therefore, aimes to focus on generating relevant and accurate captions using "data2vec - a generalized selfsupervised approach" [1] that is believed to have the power of learning from latent representations to bridge the gap between images and captions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data2vec model is believed to be a most suitable model for image captioning as it has the ability of generalized learning which transforms both visual and text data into common latent representations. It is suitable for this downstream task beacause of its power to capture the semantic meaning of both image and text which allows it to create embeddings that can very well represent and capture the complete context of the image.

The unique strength of Data2Vec lies in its ability to learn meaningful(contextual) rich and continious representations from both visual and textual inputs.By this, the model can generate captions that accurately describe the visual content in a meaningful manner. Moreover, Data2Vec can effectively generalize to unseen images by leveraging the learned latent representations, making it suitable for real-world applications.

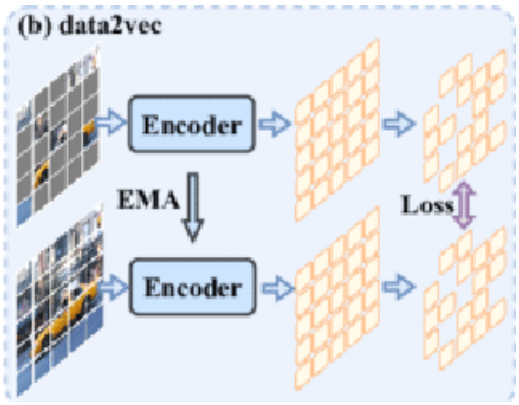

Figure 1.1: data2vec - learning [1]

Data2Vec is a novel model that effectively bridges the gap between visual and textual data for image captioning by the approach of generalized learning across different modalities.

In this, we explore the image caption generation based on data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language [1].

For this Image Captioning Task we use Transformers [17].

Chapter 2

Literature Review

2.1 Traditional methods

Traditional approaches for image captioning relied on handcrafted features and rule-based methods. These approaches often involved designing predefined caption templates or defining specific rules based on linguistic and domain knowledge to generate captions, but they were limited in their ability to capture nuanced details and lacked the flexibility to generate diverse and contextually relevant captions.

2.2 Show and Tell: A Neural Image Caption Generator

The "Show and Tell" paper by Vinyals et al. (2015) [18] achieved image captioning by proposing an end-to-end deep learning model that combined a convolutional neural network (CNN)[19] to extract image features and a long short-term memory (LSTM) network [16] to generate captions based on the extracted features, effectively mapping images to descriptive sentences.

2.3 Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

The "Show, Attend and Tell" paper by Xu et al. (2015) [20] achieved image captioning by introducing an attention mechanism in their model. The attention mechanism allowed the model to dynamically focus on different regions of the

image while generating captions, enabling more accurate and detailed descriptions.

2.4 CPTR: Full TransformerNetwork For Image Captioning

The "CPTR: Full TransformerNetwork For Image Captioning" paper [12] achieved image captioning by adapting the transformer [17] architecture, originally designed for sequence-to-sequence tasks, to handle image captioning. They combined a transformer architecture encoder for image feature extraction and a transformer network for caption generation, enabling the model to effectively capture long-range dependencies and generate coherent and contextually relevant captions. They used pretrained models ViT [8] and BEiT [3] as encoders for image feature extraction

Conclusion: Image captioning has been a rapidly evolving field with significant contributions from various research groups. The development of attention mechanisms, semantic embeddings, and transformer-based architectures has significantly improved the performance of image captioning models. Further, Data2vec enhances the encoding of image features and captures semantic relationships, leading to more contextually relevant, diverse, and coherent captions.

Chapter 3

Problem Statement

Traditional approaches to image captioning had difficulty in effectively bridging the gap between visual and text data, leading to limited performance in generating accurate and coherent captions. There is need to address this challenge and also to capture the semantic relationships between image and text. The goal is to explore the effectiveness of data2vec, a generalized self supervised learning approach, in image captioning to bridge this gap and thus enhancing the overall performance of image captioning systems.

Chapter 4

Methodology

The architecture of the base data2vec model is depicted in Figure 4.1

4.1 Data2vec model

Algorithms and learning objectives till now are specifically focused on a single modality . Data2vec is a general self supervised framework[1] that uses a same learning method for the three modalities like speech , NLP and computer vision. In this model, the idea is to predict the latent representations[9][5] of the complete input data from a masked view. A self distillation method is used using a transformer architecture .[17]

4.1.1 Learning process and Training method

There are two modes student mode and teacher mode of transformer architecture and these transformers are trained in the following steps:
1) The teacher mode deals with building representations of full input data. These representations serve as targets for the student mode which is referred to as learning mode.
2) In student mode a masked version of same input is given and student mode learns by predicting the targets given from teacher mode.

The weights of the teacher are updated by EMA of student mode parameters[10][9][5]. The target representations generated are continuous and contextualized. This is due to self attention[17] which makes these representations more richer in context than the targets based only on local information.

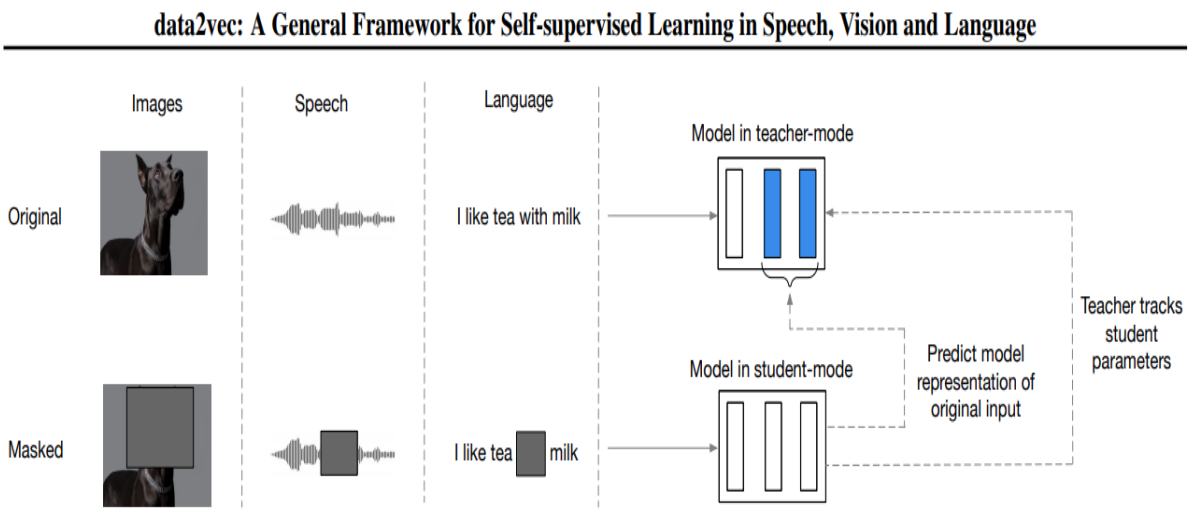

Figure 4.1: Architecture [1]

This model doesn't predict the targets like words, visual tokens or speech units. Instead it predicts the contextualised latent representations from entire input. And these predictions are done on multiple network layers instead of only top layer of transformer network(generalization on all modalities)

4.1.2 Method and model architecture

The standard transformer architecture is used along with a self distillation [15] method in it. Though the learning algorithm is the same but input encodings are modality specific like ViT strategy encoding mechanism for image as sequence of patches 16 into 16 pixels [8] each which is fed as input to linear transformation. Text as subward units which are them taken with as respective embedding vectors.

Masking: Replacing with a learned MASK embedding token and feed the sequence to the Transformer network:
1) Computer vision - Block wise masking strategy.[3]
2) Speech - spans of latent speech representations.[2]
3) NLP- mask token.[7]

The predictions by student mode for target representations happens only for the time stamps which are masked. Due to self attention used in the transformer network these target representations are contextualized which capture the information from the sample for a particular time step.

Teacher parameterization: Teacher parameterization: (Exponentially moving average - EMA):

$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2/\beta & |y_t - f_t(x)| \leq \beta \\ (|y_t - f_t(x)| - \frac{1}{2}\beta) & \text{otherwise} \end{cases}$$

Figure 4.2: Objective function [3]

1) $x_t - model\ parameters$
2)$EMA_t - weights\ of\ model\ in\ target\ mode$
3)$\alpha - \alpha_0\ to\ \alpha_e\ over\ first\ n\ updates$

$$EMA_t = \alpha \cdot x_t + (1 - \alpha) \cdot EMA_{t-1}$$

Objective Function: Given contextualized training targets yt, we use a Smooth L1 loss to regress these targets.where b controls the transition from a squared loss to an L1 loss, depending on the size of the gap between the target yt and the model prediction ft(x) at time-step t.

4.2 Image Captioning

4.2.1 Data Preprocessing

Data pre-processing and cleaning is an important part of the whole model building process. Understanding the data helps us to build more accurate models.After extracting zip files you will find below folders...

1) Flickr8k_Dataset: Contains a total of 8092 images in JPEG format with different shapes and sizes. Of which 6114 are used for training and 1529 for test.

2)Flickr8k_text : Contains text files describing train_set ,test_set. Flickr8k.token.txt contains 5 captions for each image i.e. total 40460 captions. The captions are concatinated with start and end symbols.

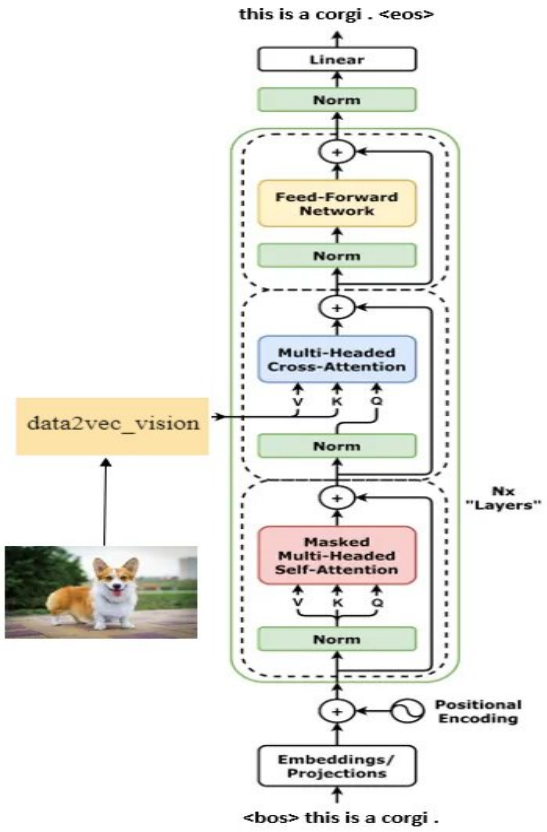

Figure 4.3: Image captioning architecture[1]

4.2.2 Encoding: Data2vec feature extraction

The next step is to extract features from the images using a data2vec Vision model. This is typically pre-trained on a large dataset such as ImageNet1K [6]. Imagenet is a standard dataset used for classification. It contains more than 14 million images in the dataset, with little more than 21 thousand groups or classes.The output embeddings of this model are the image feature extracted vectors.

4.2.3 Tying Layers

Here we use cross attention layers where key and value matrices are taken from data2vec vision model and query matrix from decoder.

4.2.4 Decoder

Since there are 5 captions for each image and we have preprocessed and encoded them in below format

“startseq “ + caption + “ endseq”

The reason behind startseq and endseq is,

1)startseq : Will act as our first word when feature extracted image vector is fed to decoder. It will kick-start the caption generation process.

2)enseq : This will tell the decoder when to stop. We will stop predicting word as soon as endseq appears or we have predicted all words from train dictionary whichever comes first.
The extracted features are then fed to transformer decoder that generates the captions for the images.The transformer decoder is trained to generate a sequence of words conditioned on image features.The transformer decoder is mainly built from attention layers. It uses self-attention to process the sequence being generated, and it uses cross-attention to attend to the image.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Transformer decoder contains a causal self attention layer (CausalSelfAttention), where each output location can attend to the output so far,A cross attention layer (CrossAttention) where each output location can attend to the input image and A feed forward network (FeedForward) layer which further processes each output location independently.Softmax and Relu activation functions are used. Output - A multiclass-classification over the output vocabulary.

Softmax:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K$$

Relu:

$$Relu(z) = max(0, z)$$

4.2.5 Training Phase

During training, the correct input is given to the decoder at every time-step, even if the decoder made a mistake before. The transformer decoder is now trained using a loss function. The loss function we use here is "cross entropy" [13]loss function that predicts the similarity between the predicted and the ground truth

captions.

Cross entropy loss:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\log(\hat{y}_{it})$$

We also use "Adam optimizer"[11] to update the weights of the model.

$$\begin{aligned} m_t &= \beta_1 \cdot m_{t-1} + (1-\beta_1)\cdot g_t \\ v_t &= \beta_2 \cdot v_{t-1} + (1-\beta_2)\cdot g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\ \theta_{t+1} &= \theta_t - \frac{\eta}{\sqrt{\hat{v}_t}+\epsilon}\cdot \hat{m}_t \end{aligned}$$

4.2.6 Evaluation

The evaluation metric [4] we used for this process is "accuracy"[4] and "bleu score" [14][4], provides a direct measure of the correctness of the generated caption.Accuracy provides a direct measure of the model's ability to generate captions that are semantically relevant and coherent.

4.2.7 Inference

Once the model is trained, we then used it to generate captions for new images. This involves feeding the image through the data2vec encoder to extract features, and then using the Transformer decoder to generate the caption.

Chapter 5

Experimental Setup

In this experiment, we are working on the Flickr8k Dataset. For this experiment we use datasamples with each image having 5 corresponding captions. A tokenizer is then used to divide the caption into tokens. We experiment with data2vec vision base as encoder Base containing L = 12 Transformer blocks with H = 768 hidden dimension (with 4 × H feed-forward inner-dimension). EMA updates are performed in fp32 for numerical stability . We embed images of 224x224 pixels with patch size is setting to 16. Each patch is linearly transformed and a sequence of 196 representations is input to a standard Transformer.The data2vec vision model is used to extract the features.

The input data(caption) is tokenized using a byte-pair encoding of 50K types and the model learns an embedding for each type.As batch size we use 64 for thos architecture. We use pre-trained data2vec moedl to extract embedded image feature of dimension 1024.for pretraining data2vec model, we use 0 = 0.999, e = 0.9999 and n = 100, 000, K = 10 and set = 4. The model is optimized with Adam over 1M updates using a tri-stage . If the pre-trained XLM model features are not available, then we can get pre-trained model features from MLM objective function. Finally, we use argmax to generate the sentence.
To evaluate our model, we use BLEU metric.

Chapter 6

Datasets and Results

6.1 Datasets

Flickr8k is one of the largest dataset used for the image captioning task.A new benchmark collection for sentence-based image description and search, consisting of 8,000 images that are each paired with five different captions which provide clear descriptions of the salient entities and events. ... The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations.

S.No	Task	Dataset size
1	Trainig	7.2k
2	Validation	1.1k

Table 6.1: dataset used.

link for flickr8k dataset: https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_Dataset.zip
https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_text.zip

It is small in size. So, the model can be trained easily on low-end laptops/desktops.Data is properly labelled. For each image 5 captions are provided.The dataset is available for free.

6.2 Results

BLEU stands for Bilingual Evaluation Understudy.
It is an algorithm, which has been used for evaluating the quality of machine translated text. We can use BLEU to check the quality of our generated caption.BLEU tells how good is our predicted caption as compare to the provided 5 reference captions.
1)BLEU is language independent
2)Easy to understand
3)It is easy to compute.
4)It lies between [0,1]. Higher the score better the quality of caption

$$modified\ ngram\ precision = \frac{max\ number\ of\ times\ ngram\ occurs\ in\ reference}{total\ number\ of\ ngrams\ in\ hypothesis}$$

BLEU-1(B1) uses the unigram Precision score. BLEU-2(B2) uses the geometric average of unigram and bigram precision. BLEU-3(B3) uses the geometric average of unigram, bigram, and trigram precision.

S.No	Model Name	B1	B2	B3	B4
1	VGG16+LSTM	0.47	0.28	0.19	0.08
2	InceptionV3+LSTM	0.48	0.27	0.18	0.08
3	Multi-feature	0.50	0.30	0.20	0.09
4	Attention model	0.51	0.31	0.22	0.10

Table 6.2: Results.

The advantage of BLEU is that the granularity it considers is an n-gram rather than a word, considering longer matching information. The disadvantage of BLEU is that no matter what kind of n-gram is matched, it will be treated the same.

Chapter 7

Future Scope

1)Till now we used data2vec model only in the encoder end . The model performance may be observed and improved by using data2vec text_base replacing the transformer decoder in the current work.i.e, future work may include performing image captioning task using data2vec as both encoder and decoder.
2)In data2vec model, despite the unified learning regime, we still use modality-specific feature extractors and masking strategies. Our approach still uses modalityspecific input encoders and we adopt modality-specific masking strategies which future work may unify.
3)Performing image captioning task effectively by Multilingual Image Captioning: We can work on extending the Data2Vec model to support multilingual image captioning. Idea is to train the model on multilingual datasets and develop mechanisms to generate captions in multiple languages based on the detected language of the input image or user preferences.

Chapter 8

Conclusion

Self-supervised learning techniques enable the model to learn rich representations from the visual content alone, without requiring explicit human annotations.Therefore, this approach thus reduces manual labelling of data. In this approach we have tried to understand and implement the image captioning task and we have came to the conclusion that the image caption generated using data2vec method lead to better performance results when compared to the models dedicated a specified modality. The concept of generalized self-supervised learning over vision and NLP helped in an effective decrease in the gap between visual and textual data for image captioning.The integration of the data2vec model in image captioning lead to improved caption quality. The Evalution results from the original paper confidently proves that this method gives robust, high-quality and more contexualized captions for the corresponding images.

Chapter 9

Acknowledgement

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that we now have the opportunity to express our gratitude to all of them. We owe our sincere gratitude to our project guide DR. Nagesh Bhattu Sristy , Department of Computer Science and Engineering, National Institute of Technology Andhra Pradesh, who took a keen interest and guided us all along, till the completion of our project work by providing all the necessary information. We avail ourselves of this proud privilege to express our gratitude to all the faculty of the Department of Computer Science and Engineering at NIT Andhra Pradesh for emphasizing and providing us with all the necessary facilities throughout the work. We offer our sincere thanks to all our friends, fellow mates, and other persons who knowingly or unknowingly helped us to complete this project.

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[4] Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. A critical analysis of metrics used for measuring progress in artificial intelligence. *arXiv preprint arXiv:2008.02577*, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] 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.

[8] 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.

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021.

[13] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*, 2023.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[15] Minh Pham, Minsu Cho, Ameya Joshi, and Chinmay Hegde. Revisiting self-distillation. *arXiv preprint arXiv:2206.08491*, 2022.

[16] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[19] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.

[20] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.