UNDER REVIEW

RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] Reproducing and extending the Hirota et al. CVPR 2022 paper

Marten Türk[1, ID], Luyang Busser[1, ID], Daniël van Dijk[1, ID], and Max J.A. Bosch[1, ID]
[1]University of Amsterdam, Amsterdam, The Netherlands – [1]Equal contribution

Edited by
(Editor)

Reviewed by
Anonymous Reviewers

Received
01 November 2018

Published
–

DOI
–

Reproducibility Summary

Scope of Reproducibility — The main objective of this paper is to reproduce and verify the following claims made in the original paper: (1) According to the LIC metric, all evaluated image captioning models amplify gender and racial bias, (2) the proposed LIC metric is robust against encoders, and (3) captioning model NIC+Equalizer amplifies gender bias beyond baseline.

Methodology — We reproduced the results of the original authors with only minor modifications to the code they made available. We contribute to their research by highlighting a noteworthy limitation in the used data split and propose an integrated gradients method to increase explainability, allowing users to understand predictions better using the Captum library for Pytorch. As for the computational requirements, all experiments were run on a cluster with a NVIDIA Titan RTX GPU and the time required to run a total of 720 models was $\sim$98 hours.

Results — The results we obtained showed the same patterns as in the original authors' work. All our results were in the range of ± 1 LIC score units compared to the original work, which supports the claims on the gender and racial bias amplification, robustness against encoders, and amplification by NIC+Equalizer beyond baseline. As for our contributions, we show that the attribution scores obtained by using integrated gradients follow similar patterns in terms of gender amplification for all evaluated language models, providing additional support for the proposed LIC metric.
During data set analysis we observed a leakage in the original data split being used, resulting in identical captions occurring multiple times in both the training and test set. The removal of already seen captions during training from the test set reduced its size by 62.4% on average and caused a decline in LIC_M scores of approximately 5 units.

What was easy — Reproducing the results using the original provided code offered no difficulties.

What was difficult — Finding a useful angle of contribution to the paper proved to be challenging. After we had decided upon using our selected explainability method, implementing and modifying existing code was more work than expected.

Correspondence should be addressed to Luyang Busser (luyang.busser@student.uva.nl)
The authors have declared that no competing interests exist.
Code is available at https://github.com/martentyrk/mlrc2022hirota. – SWH swh:1:dir:2b9535456facf442ee411ab4cae5d3c4ce1cc29e.
Open peer review is available at https://openreview.net/forum?id=N9Wn91tE7D0.

1 Introduction

The main focus of research in the AI sub field of image-to-text has been to increase the accuracy of caption-generating models, which has led to significant advancement of the state of the art in recent years [1, 2, 3]. However, these image captioning models have been shown to preserve or even increase societal biases present in training data, such as race and gender bias [4]. A standardized bias evaluation metric allows researchers to quantitatively compare the societal biases in models and address this issue as new methods are developed. Hirota et al. [5] state that current bias evaluation metrics and methods have their shortcomings. One such limitation is that they cannot differentiate between bias present in the training data and the amount of bias amplified by the image captioning model.

Therefore, Hirota et al. propose the LIC score, which measures bias in captions based on a classifier's accuracy and confidence in predicting a protected attribute. A set of captions is considered unbiased if the classifier's performance is no better than random chance. A model amplifies bias when the LIC score of the model's generated captions (LIC_M) is higher than the LIC score of the human-made captions of the training data (LIC_D).

This paper aims to reproduce the findings of the original work, verify some of the authors' claims and present additional results by using an explainability method used to understand the predictions of neural networks and furthermore provide insight into how the data set that was used.

2 Scope of reproducibility

To assess their proposed LIC metric, Hirota et al. evaluate nine image captioning models and calculate the associated LIC scores for each. Additionally, they use three different language models as classifiers for the LIC metric. The main goal of this study is to examine and expand upon the following points made in the original paper:

- Bias amplification: According to the LIC metric, all evaluated image captioning models amplify both gender and racial bias present in the data.
- Robustness against encoders: Choosing a different language model as the encoder maintains the same tendency for the LIC scores, and the different captioning models still achieve the same relative results.
- NIC+Equalizer further amplifies gender bias: NIC+Equalizer [6] amplifies gender bias more according to the LIC metric compared to the baseline NIC+ [6]. However, the NIC+Equalizer does not amplify racial bias beyond the baseline.

3 Methodology

In this section, we will explain our methodology to replicate the results of Hirota et al. and thereafter, our implementation to improve the interpretability and transparency of the results by conducting an analysis using integrated gradients. Our aim is to gain insight into how the models amplify bias.

The code provided by the original authors was executed with only slight modifications to reproduce their results. The only changes made by our team were to calculate the accuracy metrics and develop the necessary code to extend the analysis.

3.1 Model descriptions

The nine image captioning models that were evaluated have been pre-trained and the corresponding generated captions for each model are available on the GitHub project page of the original paper [7]. We will briefly discuss the NIC models and provide a full overview of all nine image captioning models in the Appendix A, as the original paper has discussed these in greater detail. **N**eural **I**mage **C**aption generator (NIC) [8] combines a convolutional neural network (CNN) [9] encoder and a long short-term memory (LSTM) [10] decoder. NIC+ is a version of NIC that is trained on both the MSCOCO and MSCOCO-Bias dataset consisting of images of male/female. NIC+Equalizer is NIC+ with a gender bias mitigation loss forcing the model to predict gender words based only on the area of the person.

Appendix B provides a detailed overview of the three language models used in this study. We utilize one fully-connected layer on top of the LSTM model for classification and two fully-connected layers combined with RELU activation for the BERT [11] models. The architecture of BERT-ft is the same as for BERT-pre, only for the latter the parameters of the language model itself are frozen during training.

3.2 Datasets

Two subsets of the MSCOCO dataset are used, one with 10780 images annotated for gender (male and female) and the other with 10969 images annotated for race (light skin and dark skin). The captions generated by the nine captioning models and human annotators are used to train LSTM or BERT classifiers to calculate the LIC scores.

To balance the datasets, 4152 male entries were removed from the gender annotated dataset, resulting in a balanced dataset of 6628 captions, with 5966 for training and 662 for testing. Similarly, the race annotated dataset was balanced by removing 8804 "light skin" entries, resulting in 2192 images, with 1972 for training and 220 for testing.

The captions were pre-processed by aligning the vocabularies, masking gender words, tokenizing, lower-casing, and transforming the tokens to their encodings. The vocabularies were aligned by replacing the words in human captions with [UNK] that were not present in the caption models' vocabulary. The gender-related words (defined by a word list) are replaced with [MASK] for BERT and "genderword" for LSTM.

Additionally, during our data set analysis, we discovered that all in some instances, captioning models generated identical captions for different images. This is problematic as the classifiers used for calculating the LIC scores only consider the text of the caption, while the original image information is discarded. We will explore and discuss this later in this paper.

Dataset links:
Captions and model vocabularies
Original MSCOCO dataset

3.3 Hyperparameters

Both the LSTM and BERT-pre use a learning rate of $5*10^{-5}$ and are trained for 20 epochs. The BERT-ft model uses a learning rate of $1*10^{-5}$ and 5 training epochs. The Adam optimizer is used for the LSTM and Adamw for the BERT models. Additionally, both the LSTM and BERT models use dropout with a probability of 0.5. We use a batch size of 64 for all models. Each model is trained 10 times with the following seeds: 0, 12, 100, 200, 300, 400, 456, 500, 789 and 1234. Our results are the average of these 10 different training runs. All hyperparameter values can be found in the code adjoined to this paper.

3.4 Experimental setup and code

In order to run the code, the simplest way is to follow the documentation in the README of our GitHub repository, which includes a list of dependencies necessary to train the BERT and LSTM models and simple commands to initiate training.

To eliminate some randomness all experiments were run using the seeds mentioned earlier, which we use to obtain an averaged LIC score and corresponding variance. The metrics include the LIC score, which was introduced by the authors and BLEU-4 [12], METEOR [13], CIDEr [14] and ROUGE-L [15] to measure the accuracy of the captions generated by the models. These scores can be observed in the results section.

We not only reproduce the results in this research, but also provide new insight by visually displaying the significance of specific words in captions that influence the predictions. This is possible using the integrated gradients method discussed in the next section. To implement this method we adapted code from the following article [16] and also wrote additional code to support the use of integrated gradients for the LSTM model.

3.5 Integrated Gradients

First introduced in the paper by Merity, Neskar and Socher in 2017 [17], the integrated gradients provide explainability to a wide variety of machine learning models. For this particular research we will look into the language models that predict the protected attribute using this integrated gradient technique. For this, we use the Captum package for Pytorch [18].

The method works with so-called attribution scores, which are based on integrated gradients. Attribution score shows how important a particular feature was for the model's prediction. This method computes the attribution based on the gradient of the model's output with respect to the embedding of the input. For this we generate a baseline input, then in a step-wise manner reconstruct the original input from this baseline and calculate the gradients for each step of reconstruction.

Finally, we obtained the attribution score for each token in the caption, which enabled us to visually understand the importance of each token in the model's prediction. This analysis allowed us to identify which words may contain implicit biases based on gender or race, and to quantify the extent of such biases. Furthermore, the sum of all feature attributions and the confidence score of the model's prediction were displayed, providing us with a more nuanced understanding of the potential amplification of biases inherent in the model.

3.6 Computational requirements

All experiments were performed on a cluster that is equipped with the Intel Xeon Gold 5118 Processor which has 12 cores and 24 threads. The GPU of the high performance computer was the high-end NVIDIA Titan RTX.
The total run time to train and test eighteen models per classifier, 10 seeds each was $\sim$98h.

4 Results

This section is split into two parts: reproducing the results of the paper and our contribution. The reproducibility mainly included experimenting with the LIC scores introduced by Hirota et al, which were in most cases successfully obtained.

The contributions include an additional in depth qualitative and a quantitative analysis of the bias introduced by the models. The second contribution focuses on dissecting the data set used to run all the experiments in the paper.

4.1 Results reproducing the original paper

The upcoming section will cover the experiments that were conducted to explore the claims made by the authors.

Testing the models for gender bias amplification – The first experiment aimed to verify the consistency of the LIC_M and LIC_D scores with those reported in the paper. The aim of this experiment was to verify the claim that all models evaluated amplified bias, leading to a higher LIC_M score than the baseline ($LIC_M = 25$).

As expected, all models regardless of the classifier used, produced higher scores then the baseline, aligning with the authors' claims. The original paper found that the NIC model introduced the least amount of bias among all classifiers, and the results of our experiments confirmed this observation, as shown in Table 1 and Table 2.
However, our results deviated from the original findings, as for us, the OSCAR model demonstrated the worst LIC score in comparison to the NIC+Equalizer when a pre-trained BERT was used for classification.

	BERT-pre			BERT-ft		
Model	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
NIC [8]	43.2 ± 1.0	41.1 ± 0.8	2.1	47.3 ± 1.9	48.0 ± 1.0	-0.7
SAT [19]	44.0 ± 1.5	41.4 ± 0.9	2.6	47.9 ± 1.4	47.5 ± 1.3	0.4
FC [20]	46.4 ± 1.6	40.2 ± 0.8	6.2	48.9 ± 1.9	45.7 ± 1.2	3.2
Att2in [20]	45.5 ± 1.1	40.8 ± 0.8	4.7	47.9 ± 1.8	46.6 ± 1.1	1.3
UpDn [21]	48.5 ± 1.1	41.4 ± 0.8	7.1	52.1 ± 1.1	47.3 ± 1.1	4.8
Transformer [22]	47.5 ± 1.1	42.0 ± 1.0	5.5	54.1 ± 1.6	48.4 ± 1.1	5.7
OSCAR [23]	48.1 ± 1.1	40.8 ± 0.8	7.3	52.4 ± 1.5	47.4 ± 1.2	5
NIC+ [6]	46.8 ± 1.1	40.8 ± 0.8	6.0	49.5 ± 1.5	47.7 ± 1.1	1.8
NIC+Equalizer [6]	49.3 ± 0.8	42.9 ± 0.8	6.4	54.6 ± 1.4	47.4 ± 1.2	7.2

Table 1. Gender LIC scores per model for BERT. The lower the score, the better. Red/green denotes the worst/best among all models.

	LSTM Gender bias ↓			Model accuracies ↑			
Model	LIC_M	LIC_D	LIC	BLEU-4	CIDEr	METEOR	ROUGE-L
NIC [8]	43.2 ± 1.5	39.5 ± 0.9	3.7	61.8	33.5	37.9	35.8
SAT [19]	44.7 ± 1.2	39.2 ± 0.9	5.5	71.6	70.1	47.5	45.9
FC [20]	45.8 ± 0.8	37.8 ± 0.8	8	71.1	68.6	46.1	45.3
Att2in [20]	45.7 ± 0.9	38.2 ± 0.9	7.5	73.7	75.2	49.1	47.7
UpDn [21]	47.5 ± 0.8	39.0 ± 1.0	8.5	76.6	84.6	51.9	49.2
Transformer [22]	48.3 ± 1.1	39.8 ± 0.7	8.5	73.9	74.9	50.2	47.0
OSCAR [23]	48.6 ± 0.9	39.1 ± 0.5	9.5	79.6	100.4	54.6	52.0
NIC+ [6]	46.3 ± 1.5	39.3 ± 0.5	7	69.6	59.5	45.0	44.2
NIC+Equalizer [6]	51.9 ± 0.7	39.5 ± 0.8	12.4	68.7	56.1	43.9	43.4

Table 2. Gender LICs per model for LSTM. Red/green denotes the worst/best performance for a certain model compared to all models tested. For bias, lower is better, for accuracies, higher is better. The unbiased model baseline is $LIC_M = 25$ and $LIC = 0$

Testing the models for racial bias — The aim of the second experiment was to confirm the claims made about models which included captions referring to race. The claim stated that all models amplify racial bias and furthermore, that racial bias LIC scores are closer to random chance than gender bias models. Our results were in correlation with the ones obtained by Hirota et al. and showed that racial bias is less amplified since all models produce LIC scores that are closer to the baseline, as can be seen in Table 3.

	LSTM		
Model	LIC_M	LIC_D	LIC
NIC [8]	33.3 ± 1.8	27.8 ± 1.3	5.5
SAT [19]	31.2 ± 2.3	26.7 ± 1.0	4.5
FC [20]	33.6 ± 1.0	26.2 ± 0.6	7.4
Att2in [20]	35.2 ± 2.3	26.8 ± 0.7	8.4
UpDn [21]	34.1 ± 2.8	26.9 ± 0.5	7.2
Transformer [22]	33.2 ± 2.2	27.1 ± 0.7	6.1
OSCAR [23]	32.9 ± 1.8	26.9 ± 1.2	6
NIC+ [6]	34.8 ± 1.4	27.6 ± 1.1	7.2
NIC+Equalizer [6]	33.2 ± 2.2	26.7 ± 0.7	6.5

Table 3. Race LIC scores per model for LSTM according to LIC_M, LIC_D and LIC. Captions are not masked.

4.2 Results beyond original paper

In addition to validating the claims of the original authors, we also aimed to contribute to their work, increasing the explainability of the results using the integrated gradients method. The experiments focused on the NIC, NIC+ and NIC+Equalizer models, as they represent a range of the best and worst models and provide a baseline for comparison. Furthermore, we show findings relating to the MSCOCO data set used for evaluations.

Qualitative integrated gradients results — Figure 1 showcases two examples of the integrated gradients method. The results show the captions of the human annotators and the NIC+ and NIC+Equalizer models for two different images. For each caption, the true label of the corresponding image can be seen (zero for male and one for female) as well as the label predicted by the classifier together with the confidence score. A score lower than 0.5 will be classified as female and a higher score as male. The total attribution score of a sentence is the sum of the attribution scores of the individual words. In the figure, words highlighted with brighter green indicate a more positive attribution score (contributing towards female prediction) and words in darker red indicate a more negative attribution score (contributing toward male prediction).

(a) Image ID: 531602

Legend: Negative Neutral Positive

	True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
Human:	0	1 (0.74)	1	1.75	a little genderword brushing genderword teeth in the bathroom
NIC+:	0	1 (0.73)	1	1.75	a young genderword brushing genderword teeth with a toothbrush .
NIC+Eq:	0	1 (0.76)	1	2.18	a young genderword brushing genderword teeth with a toothbrush .

(b) Image ID: 76632

Legend: Negative Neutral Positive

	True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
Human:	0	0 (0.03)	1	-1.92	a baseball player throwing a ball at a person
NIC+:	0	0 (0.35)	1	-0.98	a group of people on a field playing baseball .
NIC+Eq:	0	0 (0.04)	1	-1.40	a baseball player holding a bat next to home plate .

Figure 1. Integrated gradients example on two sets of image captions generated by a human, NIC+, and NIC+Equalizer respectively. The images that the captions describe can be seen on the left together with their corresponding image ID in the MSCOCO dataset.

Quantifying the gender bias — In order to better understand which of the two genders gets amplified and how much, we ran our experiments on NIC, NIC+ and NIC+Equalizer, since they provide an overview of the best and the worst model and additionally NIC+ provides a baseline to NIC+Equalizer. In order to quantify bias we averaged the attribution scores for both genders as can be seen in Table 4. The average was obtained after running the model for 10 different seeds. These scores were obtained using the LSTM model. Unfortunately executing the algorithm on the BERT models was time consuming, hence, we are able to show the attribution scores for BERT-ft for only one seed, the results can be seen in Appendix C in Table 8.

	LSTM		
Model	Female	Male	Sum
NIC [8]	1.12 ± 0.0	1.10 ± 0.0	2.22
NIC+ [6]	1.20 ± 0.0	1.18 ± 0.0	2.38
NIC+Equalizer [6]	1.31 ± 0.0	1.22 ± 0.0	2.53

Table 4. Averages of attribution scores for the LSTM model.

Duplicates in the data set and the effect on LIC and attribution scores — During our examination of the original code and methodology employed in the research, we encountered certain challenges. One of these challenges was the presence of duplicate entries in both the training and testing datasets. This limitation was brought to light through the gracious provision of access to the code and dataset by the original authors.

To mitigate the issue, we implemented a solution that involved the elimination of all captions in the test set that overlapped with the captions in the training set. Subsequently, we re-calculated the LIC_M scores to assess the impact of this modification. The results of this recalculation are shown in Table 5.

	LSTM duplicate removal		
Model	LIC_M	Unseen	Difference
NIC [8]	40.0 ± 1.9	231 ± 17	-3.2
SAT [19]	41.4 ± 2.3	296 ± 14	-3.4
FC [20]	38.8 ± 2.5	133 ± 11	-7
Att2in [20]	39.1 ± 2.1	214 ± 13	-6.6
UpDn [21]	42.9 ± 1.7	279 ± 14	-4.6
Transformer [22]	43.5 ± 0.8	452 ± 16	-4.8
OSCAR [23]	46.0 ± 1.8	373 ± 15	-2.6
NIC+ [6]	38.6 ± 3.3	133 ± 10	-7.7
NIC+Equalizer [6]	39.2 ± 2.6	137 ± 7	-12.7

Table 5. LIC_M scores averaged over ten seeds after removing training samples from the test set. The average number of non-training samples are displayed in the ”Unseen” column and the difference with our original LIC_M is displayed with ”Difference”. The original test set size was 662 samples for each model.

5 Analysis

The following section will elaborate more on the results mentioned earlier. First, the observed discrepancy in our reproduced results between NIC+Equalizer and Oscar with the pre-trained BERT as language encoder, is reasonable. In the original paper the observation is made that there is a trade-off between bias amplification and model performance. Since OSCAR was the best performing model in terms of accuracy metrics, the higher LIC score is in line with this observation.

In Table 4 we observe that the NIC+Equalizer model amplifies bias the most, which can also be noticed by looking at the sums of the attribution scores, where we observe that the scores are highest for NIC+Equalizer. These scores verify the claim of the authors that NIC+Equalizer amplifies bias beyond the baseline NIC+ from a different perspective, using integrated gradients.

Finally, the results of the removal of duplicates, as presented in Table 5, indicate a decline in the LIC_M scores compared to the original results, despite similar training sets and models. The test sets decrease in size on average by 68%, with the FC and NIC+ models having the most substantial reductions of around 80%. This reduction in the test set size could potentially compromise the reliability of our results, as the decreased volume might result in increased variance and sensitivity to outliers. However, the decrease in the LIC_M scores across all models suggests that the initial estimation of bias amplification may still have been overestimated.

6 Discussion

In this study, we aimed to replicate the findings presented in the work of Hirota et al. and extend it by incorporating the integrated gradients method developed by Merity et al. Our results indicate that the integrated gradients method provides valuable insight into the decision-making process of classification models and has the potential to mitigate the problem of bias amplification in captioning models in future studies.

However, it is worth noting that the computational expense of this method, particularly when using BERT as the language encoder, is a limitation. The calculation of attribution scores and visualization took an excessive amount of time, which made it infeasible in this study to perform more attribution score analysis using BERT.

Additionally, the presence of duplicate captions in the training and test sets raises concerns regarding the validity of the original results. Our brief experiment indicated a decline in the average LIC_M scores by approximately 5 units, highlighting the potential for overfitting and overconfident predictions. Therefore, future research should address the issue of duplicate captions in the train and test splits and re-evaluate the reliability of these results to ensure the validity of the findings.

6.1 Communication with original authors

Thanks to clear writing and accessible code/data, we were able to reproduce results without contacting the authors. We did however contact the authors to inform them of our work. Hirota's response promised to remove the data leakage in the human captions and as for the duplicates in the outputs of the captioning models, Hirota did not find it reasonable to remove those duplicates. The reason being that generating similar captions is something that captioning models tend to do in comparison to humans and this is one of the reasons for bias.

References

1. C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju. "Improving Image Captioning with Conditional Generative Adversarial Nets." In: **Proceedings of the AAAI Conference on Artificial Intelligence** 33.01 (Aug. 2019), pp. 8142–8150. DOI: 10.1609/aaai.v33i01.33018142. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4823.
2. M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara. **From Show to Tell: A Survey on Deep Learning-based Image Captioning**. 2021. DOI: 10.48550/ARXIV.2107.06912. URL: https://arxiv.org/abs/2107.06912.
3. S. Herdade, A. Kappeler, K. Boakye, and J. Soares. "Image Captioning: Transforming Objects into Words." In: **Advances in Neural Information Processing Systems**. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/680390c55bbd9ce416d1d69a9ab4760d-Paper.pdf.
4. J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. **Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints**. 2017. DOI: 10.48550/ARXIV.1707.09457. URL: https://arxiv.org/abs/1707.09457.
5. Y. Hirota, Y. Nakashima, and N. Garcia. **Quantifying Societal Bias Amplification in Image Captioning**. 2022. DOI: 10.48550/ARXIV.2203.15395. URL: https://arxiv.org/abs/2203.15395.
6. L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. "Women also Snowboard: Overcoming Bias in Captioning Models (Extended Abstract)." In: (2018). DOI: 10.48550/ARXIV.1807.00517. URL: https://arxiv.org/abs/1807.00517.
7. Y. Hirota, Y. Nakashima, and N. Garcia. **lick caption bias**. https://github.com/rebnej/lick-caption-bias. 2022.

8. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and Tell: A Neural Image Caption Generator." In: (). DOI: 10.48550/ARXIV.1411.4555. URL: https://arxiv.org/abs/1411.4555.
9. Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning." In: **nature** 521.7553 (2015), pp. 436–444.
10. S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory." In: **Neural Computation** 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.
12. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation." In: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.
13. M. Denkowski and A. Lavie. "Meteor Universal: Language Specific Translation Evaluation for Any Target Language." In: **Proceedings of the Ninth Workshop on Statistical Machine Translation**. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 376–380. DOI: 10.3115/v1/W14-3348. URL: https://aclanthology.org/W14-3348.
14. R. Vedantam, C. L. Zitnick, and D. Parikh. "CIDEr: Consensus-based Image Description Evaluation." In: (2014). DOI: 10.48550/ARXIV.1411.5726. URL: https://arxiv.org/abs/1411.5726.
15. C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries." In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.
16. R. Winastwan. **Interpreting the prediction of Bert Model for Text Classification**. Jan. 2023. URL: https://towardsdatascience.com/interpreting-the-prediction-of-bert-model-for-text-classification-5ab09f8ef074.
17. S. Merity, N. S. Keskar, and R. Socher. "Regularizing and optimizing LSTM language models." In: **arXiv preprint arXiv:1708.02182** (2017).
18. A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: **Advances in Neural Information Processing Systems 32**. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
19. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." In: **Proceedings of the 32nd International Conference on Machine Learning**. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 2048–2057. URL: https://proceedings.mlr.press/v37/xuc15.html.
20. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. "Self-critical Sequence Training for Image Captioning." In: (2016). DOI: 10.48550/ARXIV.1612.00563. URL: https://arxiv.org/abs/1612.00563.
21. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering." In: (2017). DOI: 10.48550/ARXIV.1707.07998. URL: https://arxiv.org/abs/1707.07998.
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. **Attention Is All You Need**. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.
23. X. Li et al. "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks." In: (2020). DOI: 10.48550/ARXIV.2004.06165. URL: https://arxiv.org/abs/2004.06165.

Appendix

A Image captioning models

Model	Description	Type
NIC [8]	**N**eural **I**mage **C**aption generator combines a convolutional neural network (CNN) encoder and a long short-term memory (LSTM) decoder	CNN encoder-LSTM decoder model
SAT [19]	**S**how, **A**ttend and **T**ell: Neural image caption generation with visual attention	CNN encoder-LSTM decoder model
FC [20]	Encodes input images using a deep CNN and embeds it through a linear projection with word one-hot vectors that are embedded linearly	CNN encoder-LSTM decoder model
Att2in [20]	Attention model that dynamically re-weights the input spatial (CNN) features to focus on specific regions of the image	CNN encoder-LSTM decoder model
UpDn [21]	Combined bottom-**up** and top-**down** attention mechanism that enables attention to be calculated at the level of objects	CNN encoder-LSTM decoder model
Transformer [22]	Network architecture based solely on attention mechanisms	Transformer-based model
OSCAR [23]	**O**bject-**S**emanti**c**s **A**ligned P**r**e-training, which uses object tags detected in images as anchor points to ease the learning of alignments	Transformer-based model
NIC+ [6]	Version of NIC that is trained on both the MSCOCO and MSCOCO-Bias dataset consisting of images of male/female	CNN encoder-LSTM decoder model
NIC+ Equalizer [6]	NIC+ with a gender bias mitigation loss forcing the model to predict gender words based only on the area of the person	CNN encoder-LSTM decoder model

Table 6. A full overview of all nine evaluated image captioning models

B Language models

Model	Pre-trained/Fine-tuned	Total parameters	Trainable parameters	Hidden dimension size
LSTM [10]	Fine-tuned	2,372,157	Same as total	256
BERT-pre [11]	Pre-trained	109,681,666	199,426	256
BERT-ft [11]	Fine-tuned	109,681,666	Same as total	256

Table 7. Overview of the language models

C Attribution scores for the fine-tuned BERT

	BERT-ft		
Model	Female	Male	Sum
NIC [8]	1.08	0.86	1.95
NIC+ [6]	0.98	1.13	2.11
NIC+Equalizer [6]	1.07	1.09	2.16

Table 8. Attribution scores for the BERT-ft model run on a seed 0.