

A Concise Model for Medical Image Captioning

Notebook for the ImageCLEFmedical Caption Lab at CLEF 2023

Aaron Nicolson^{1,*}, Jason Dowling¹ and Bevan Koopman¹

¹*Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston 4006, Queensland, Australia*

Abstract

We describe our participation in the ImageCLEFmedical Caption task of 2023. The task required participants to automatically compose coherent captions for a set of medical images. To this end, we employed a concise encoder-to-decoder model for caption generation. In addition, we leveraged Self-Critical Sequence Training (SCST) to optimise our model on the primary metric of the competition, BERTScore. CSIRO placed first amongst the participating teams—with a BERTScore of 0.643. The decoder of our best-performing submission was conditioned on the visual features of the medical image via the self-attention rather than the cross-attention. Here, the visual features were mapped to the token embedding space and used to prompt the decoder. Code and model checkpoints are available at https://github.com/aehrc/imageclefmedical_caption_23.

Keywords

Medical image captioning, Multimodal learning, Encoder-to-decoder model

1. Introduction

We detail our participation in the ImageCLEFmedical Caption task of 2023, the 7th edition of the task [1, 2]. Specifically, we participated in the caption prediction subtask. Here, participants were tasked with automatically generating captions for given medical images, where the image could be one of many modalities, e.g., radiography, ultrasonography, computed tomography, magnetic resonance, etc. The development of medical image captioning methods lays the groundwork for potential multimodal medical image analysis tools that could assist with clinical documentation, maintain and improve the consistency, quality, and efficiency of clinical reporting, produce rich textual descriptions from medical images, provide fast and inexpensive second readers, and help reduce teaching time.

For the 7th edition, several issues with the dataset (lemmatization errors and duplicate captions) were amended from the previous edition. The primary evaluation metric for the caption prediction subtask was also changed to a metric that captures the semantic similarity between generated and label captions, namely, BERTScore [3].

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

✉ aaron.nicolson@csiro.au (A. Nicolson); jason.dowling@csiro.au (J. Dowling); bevan.koopman@csiro.au (B. Koopman)

🆔 0000-0002-7163-1809 (A. Nicolson); 0000-0001-9349-2275 (J. Dowling); 0000-0001-5577-3391 (B. Koopman)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

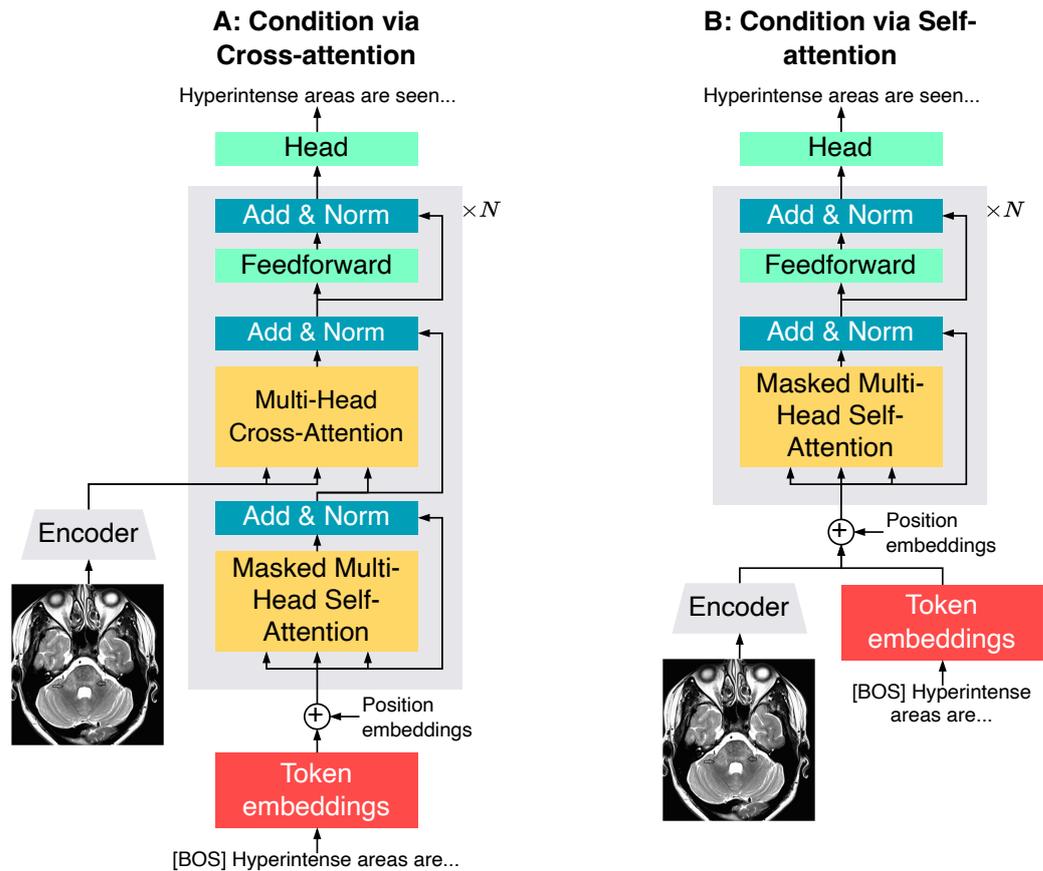


Figure 1: Decoder conditioned on the visual features of the image via **A)** the cross-attention, and **B)** the self-attention. The visual features are extracted with the encoder. CC BY [Muacevic et al. (2022)]. N is the number of Transformer blocks. [BOS] is the beginning-of-sentence special token.

Our proposed approach for the caption prediction subtask builds upon our participation in previous editions where we used encoder-to-decoder model [4, 5]. As in the previous edition, we employ the Convolutional vision Transformer (CvT) [6] as the encoder and DistilGPT2 [7] as the decoder, forming the CvT2DistilGPT2 encoder-to-decoder model [8]. The novelty for this edition lies in the use of reinforcement learning to optimise the model for the primary metric and the means of conditioning the decoder on the visual features. For reinforcement learning, we employed BERTScore as the reward for Self-Critical Sequence Training (SCST) [9].

Motivated by the Pre-trained Language Model (PaLM) with continuous observations of modalities Embodied into the token embedding space (PaLM-E) [10], we investigated a different method of conditioning the decoder on the visual features extracted via the encoder. The standard approach of conditioning is through the cross-attention of the decoder. However, as with PaLM-E, the visual features can be mapped to the token embedding space and used to prompt the decoder. This has the advantage of requiring no cross attention, as shown in Figure 1. However, this at the cost of increasing the input sequence length, and thus the size of the

self-attention matrices of each head. We aim to determine if there is a performance difference between conditioning via the self-attention, rather than the cross-attention.

2. Dataset

The dataset for the task is an updated and extended version of the Radiology Objects in COntext (ROCO) dataset [11], which was formed from figures in open-access biomedical journal articles from PubMed Central. All images in the dataset were accompanied by a caption, which form the labels for the caption prediction task. Each caption was pre-processed by removing links from the captions. The splits for the dataset are as follows:

- **Training set:** 60 918 images and their corresponding captions.
- **Validation set:** 10 437 images and their corresponding captions.
- **Test set:** 10 473 images and their corresponding captions.

3. Methodology

3.1. Models

The two encoder-to-decoder models that we used for our submissions are shown in Figure 1. CvT-21 was the encoder (specifically, the `microsoft/cvt-21-384-22k` checkpoint) [6, 8].¹ Layer normalisation was applied to its last hidden state, followed by a projection to the decoder’s hidden state size. Each image was resized using bilinear interpolation so that its smallest side had a length of 384 and its largest side maintained the aspect ratio. Next, the resized image was cropped to a size of $\mathbb{R}^{3 \times 384 \times 384}$. The crop location was random during training and centred during testing. During training, the image was rotated around its centre where the angle of rotation was sampled from $\mathcal{U}[-5^\circ, 5^\circ]$. Finally, the image was standardised using the mean and standard deviation provided with the CvT-21 checkpoint.

DistilGPT2, along with its tokenizer, was used for decoding [7]. Greedy search and beam search with four beams were employed during validation and testing, respectively. The maximum number of tokens for the labels and the generated captions was 256. During testing, a penalty was applied during caption generation to the probability of tokens to prevent trigrams from appearing more than once in a caption (the penalty was realised by setting a token’s probability to zero). Before training, both the CvT-21 and DistilGPT2 checkpoints were used to warm-start the encoder and decoder, respectively.

For the model conditioned via the cross-attention (CA) in Figure 1 A (CvT2DistilGPT2-CA), randomly initialised multi-head cross-attention modules were added to each Transformer block. Here, the visual features from the encoder were passed as the keys and values to the cross-attention heads. For the model conditioned via the self-attention (SA) in Figure 1 B (CvT2DistilGPT2-SA), the visual features and token embeddings were concatenated before adding the position embeddings. The visual features occupied the first 576 positions with the token embeddings occupying the remaining positions (DistilGPT2 accommodates up to 1 024

¹<https://huggingface.co/microsoft/cvt-21-384-22k> (last accessed 03/07/2023).

Table 1

Scores for each of the runs of team CSIRO. The primary metric is highlighted in grey.

Team Name	Run	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
CvT2DistilGPT2-CA	1	0.622	0.243	0.305	0.206	0.090	0.213	0.815
+ SCST (BERTScore)	2	0.641	0.246	0.315	0.159	0.080	0.207	0.814
CvT2DistilGPT2-SA	3	0.619	0.235	0.306	0.192	0.084	0.197	0.813
+ SCST (BERTScore)	4	0.643	0.245	0.314	0.161	0.080	0.203	0.815

positions). Here, the encoder learns to map the visual features to the token embedding space, à la PaLM-E [10]. The mapped visual features were then used to prompt the decoder.

Training: Two stages of training were performed: Teacher Forcing (TF) [12], followed by SCST. Gradient descent optimisation was performed with AdamW [13] with a mini-batch size of 32 at an initial learning rate of $5e-5$ for TF and $5e-6$ for SCST. The models were trained with NVIDIA Tesla P100 16 GB GPUs and automatic mixed precision. For TF, early stopping with a patience of eight was employed. For SCST, three epochs were completed and validation was performed every $\frac{1}{10}$ of an epoch. The validation BERTScore was the monitored metric for early stopping and checkpoint selection. For SCST, the baseline was generated with greedy search, while the sample was produced with top- k sampling ($k = 50$).

3.2. Metrics

The primary metric for this edition of the caption prediction subtask was BERTScore with `microsoft/deberta-xlarge-mnli` as the model [3]. ROUGE-1 was the secondary metric [14]. The following metrics were also included in the evaluation: METEOR [15], CIDEr [16], BLEU-1 [17], BLEURT (BLEURT-20) [18], and CLIPScore [19]. For all metrics, both the generated and label captions were pre-processed by converting to lower-case, replacing numbers with ‘number’, and removing punctuation.

4. Results & Discussion

The results for each of our four submissions are shown in Table 1. As expected, employing BERTScore as a reward for SCST improved the BERTScore. Choosing BERTScore as the reward positively impacted the scores for ROUGE and BLEURT, while negatively impacting the scores for BLEU and METEOR, and had no noticeable impact on CIDEr and CLIPScore. When comparing the method of conditioning on the visual features, conditioning via the cross-attention outperformed conditioning via the self-attention on six out of the seven metrics with TF (CvT2DistilGPT2-CA vs. CvT2DistilGPT2-SA). However, for SCST, conditioning via the self-attention performed better than conditioning via the cross-attention for three of the metrics, while they performed equally for METEOR (CvT2DistilGPT2-CA + SCST (BERTScore) vs. CvT2DistilGPT2-SA + SCST (BERTScore)). This indicates that there is no substantial difference between their performance.

The leaderboard for the competition is shown in Table 2. Run 4 (CvT2DistilGPT2-SA + SCST (BERTScore)) was compared to the runs of the other participants as it scored the highest

BERTScore. Team CSIRO ranked first based on the primary metric (BERTScore), with a score of 0.643. Team CSIRO also attained the highest CLIPScore, the third highest ROUGE, BLEURT, and CIDEr scores, the fourth highest METEOR score, and the sixth highest BLEU score. The lower rank of Run 4 for METEOR and BLEU could be attributed to optimising with SCST with BERTScore as the reward.

Shown in Figures 2 and 3 are generated reports for given medical images from the validation set. Here, we inspect the impact of SCST on the generated reports (compared to only using TF), as this had the largest impact on performance. Here, we use CvT2DistilGPT2-SA. Shown in Figure 2 are examples where SCST outperforms TF (in terms of the BERTScore). For image 000414, both TF and SCST identify that there is contrast. SCST identifies the correct plane and provides more details about the modality. However, neither identifies the empty sella. For image 002044, both identify the modality and body part correctly. TF does not identify the opacity. While SCST correctly identifies the opacity, the location was incorrect. In Figure 3 are examples where TF outperforms SCST. For image 008243, both identify the modality. TF identifies that there is an aneurysm of the Internal Carotid Artery (ICA), but incorrectly identifies the left ICA instead of the right ICA. TF also identifies that there is damage to the right ICA (pseudoaneurysm) which is semantically similar to what was described in the label (aneurysmal rupture). SCST incorrectly identifies the artery and the abnormality. For image 000193, TF identifies the right coronal artery, which is connected to the mitral valve. While SCST identifies the body part, it introduces a false positive abnormality and identifies the wrong artery. It should be noted that SCST identified calcification in three out of the four examples shown in Figures 2 and 3, which indicates that SCST could increase hallucinations. While this is a small sample of the differences between SCST and TF, it is clear that SCST did not improve performance across all examples.

Table 2

Leaderboard for the caption prediction subtask of ImageCLEFmedical Caption 2023. The primary metric used to rank the participants is highlighted in grey.

Team Name	Run	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
CSIRO	4	0.643	0.245	0.314	0.161	0.080	0.203	0.815
closeAI2023	7	0.628	0.240	0.321	0.185	0.087	0.238	0.807
AUEB-NLP-Group	2	0.617	0.213	0.295	0.169	0.072	0.147	0.804
PCLmed	5	0.615	0.253	0.317	0.217	0.092	0.232	0.802
VCMi	5	0.615	0.218	0.308	0.165	0.073	0.172	0.808
KDE-Lab Med	3	0.615	0.222	0.301	0.156	0.072	0.182	0.806
SSN MLRG	1	0.602	0.211	0.277	0.142	0.062	0.128	0.776
DLNU CCSE	1	0.601	0.203	0.263	0.106	0.056	0.133	0.773
CS Morgan	10	0.582	0.156	0.224	0.057	0.044	0.084	0.759
Clef-CSE-GAN-Team	2	0.582	0.218	0.269	0.145	0.070	0.174	0.789
Bluefield-2023	3	0.578	0.153	0.272	0.154	0.060	0.101	0.784
IUST NLPLAB	6	0.567	0.290	0.223	0.268	0.100	0.177	0.807
SSNSheerinKavitha	4	0.544	0.087	0.215	0.075	0.026	0.014	0.687

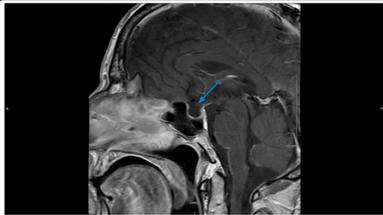
Identifier	000414	002044
Image		
Label	Contrast-enhanced T1-weighted sagittal image of the brain, 1 month after initial presentation. The arrow shows a mostly empty sella.	Chest x-ray of patient 2 with right middle to lower lung opacity.
Generated (TF)	MRI of the brain with contrast showing a large pituitary stalk (blue arrow) and a large cerebellar pedunculated mass (red arrow).	Chest X-ray showing bilateral interstitial infiltrates
BERTScore (TF)	0.585	0.715
Generated (SCST)	Sagittal T1-weighted MRI scan of the brain showing a calcification in the left cerebellum (blue arrow).	Chest X-ray showing a left pulmonary opacity.
BERTScore (SCST)	0.714	0.797

Figure 2: Images 000414 (CC BY-NC [Murvelashvili et al. (2021)]) and 002044 (CC BY-NC [Ogamba et al. (2021)]) from the validation set, their corresponding labels, and the corresponding generated captions for both CvT2DistilGPT2-SA with TF and additionally with SCST. The BERTScores for each generated report are also given.

5. Conclusion

In this work, we detailed our participation in the caption prediction subtask of ImageCLEFmedical Caption 2023. By leveraging SCST with the primary metric, BERTScore, team CSIRO was able to rank first amongst participating teams. We also investigated conditioning the decoder on the visual features via the cross-attention or self-attention. The results indicate that there is no substantial difference in performance between the two configurations. This demonstrates that there is no penalty when removing the cross-attention and instead using the self-attention to condition the decoder on the visual features. While the selected metrics for this edition have improved the evaluation process considerably, they are still general-domain metrics. The evaluation process could be improved by including domain-specific metrics that better capture the semantic similarity between captions. Such a metric could be derived from domain-specific encoders, such as PubMedBERT [20] or CXR-BERT (general) [21].²

²<https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-general> (last accessed 02/07/2023).

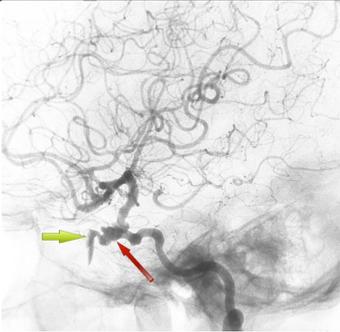
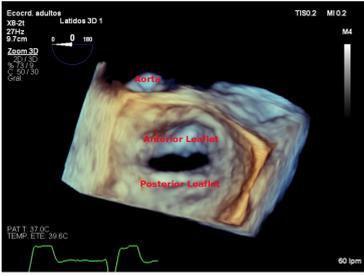
Identifier	008243	000193
Image		
Label	DSA lateral view showing an aneurysm of the right ICA (red arrow) along with the site of the aneurysmal rupture (green arrow). DSA: digital subtraction angiogram; ICA: internal carotid artery	Zoom 3D focused on the mitral valve as view from the left atrium with the aorta placed superiorly (surgical view).
Generated (TF)	Angiography of the left internal carotid artery showing a large aneurysm (red arrow) and a pseudoaneurysm of the right internal carotid artery (yellow arrow).	The image of the right coronary artery.
BERTScore (TF)	0.703	0.635
Generated (SCST)	Angiography of the brain showing a calcification in the left vertebral artery (red arrow).	Coronal scan of the heart showing a calcification in the left anterior coronary artery.
BERTScore (SCST)	0.629	0.631

Figure 3: Images 008243 (CC BY [Muacevic et al. (2021)]) and 000193 (CC BY [Ruiz et al. (2021)]) from the validation set, their corresponding labels, and the corresponding generated captions for both CvT2DistilGPT2-SA with TF and additionally with SCST. The BERTScores for each generated report are also given.

Acknowledgments

This work was partially funded by CSIRO's Machine Learning and Artificial Intelligence Future Science Platform.

References

- [1] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcia Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu,

- J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023)*, Springer Lecture Notes in Computer Science (LNCS), Thessaloniki, Greece, 2023.
- [2] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection, in: *CLEF2023 Working Notes, CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2023.
- [3] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, 2019. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- [4] A. Nicolson, J. Dowling, B. Koopman, AEHRC CSIRO at ImageCLEFmed Caption 2021, in: *Proceedings of the 12th International Conference of the CLEF Association*, Bucharest, Romania, 2021.
- [5] L. Lebrat, A. Nicolson, R. S. Cruz, G. Belous, B. Koopman, J. Dowling, CSIRO at ImageCLEFmedical Caption 2022, in: *Proceedings of the 13th International Conference of the CLEF Association*, Bologna, Italy, 2022.
- [6] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CvT: Introducing Convolutions to Vision Transformers, in: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2021, pp. 22–31. URL: <https://ieeexplore.ieee.org/document/9710031/>. doi:10.1109/ICCV48922.2021.00009.
- [7] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020. URL: <http://arxiv.org/abs/1910.01108>. doi:10.48550/arXiv.1910.01108, arXiv:1910.01108 [cs].
- [8] A. Nicolson, J. Dowling, B. Koopman, Improving Chest X-Ray Report Generation by Leveraging Warm-Starting, 2022. URL: <http://arxiv.org/abs/2201.09405>. doi:10.48550/arXiv.2201.09405, arXiv:2201.09405 [cs].
- [9] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-Critical Sequence Training for Image Captioning, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, 2017, pp. 1179–1195. URL: <http://ieeexplore.ieee.org/document/8099614/>. doi:10.1109/CVPR.2017.131.
- [10] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence, PaLM-E: An Embodied Multimodal Language Model, 2023. URL: <http://arxiv.org/abs/2303.03378>. doi:10.48550/arXiv.2303.03378, arXiv:2303.03378 [cs].
- [11] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, in: D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, P. Jannin (Eds.), *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 180–189. doi:10.1007/978-3-030-01364-6_20.
- [12] R. J. Williams, D. Zipser, A Learning Algorithm for Continually Running Fully Recurrent

- Neural Networks, *Neural Computation* 1 (1989) 270–280. doi:10.1162/neco.1989.1.2.270, conference Name: Neural Computation.
- [13] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, 2018. URL: <https://openreview.net/forum?id=Bkg6RiCqY7>.
- [14] C.-Y. Lin, E. Hovy, Automatic evaluation of summaries using N-gram co-occurrence statistics, in: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, volume 1, Association for Computational Linguistics, Edmonton, Canada, 2003, pp. 71–78. URL: <http://portal.acm.org/citation.cfm?doid=1073445.1073465>. doi:10.3115/1073445.1073465.
- [15] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: <https://aclanthology.org/W05-0909>.
- [16] R. Vedantam, C. L. Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, 2015, pp. 4566–4575. URL: <http://ieeexplore.ieee.org/document/7299087/>. doi:10.1109/CVPR.2015.7299087.
- [17] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Association for Computational Linguistics, Philadelphia, Pennsylvania, 2001, p. 311. URL: <http://portal.acm.org/citation.cfm?doid=1073083.1073135>. doi:10.3115/1073083.1073135.
- [18] T. Sellam, D. Das, A. Parikh, BLEURT: Learning Robust Metrics for Text Generation, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: <https://aclanthology.org/2020.acl-main.704>. doi:10.18653/v1/2020.acl-main.704.
- [19] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. URL: <https://aclanthology.org/2021.emnlp-main.595>. doi:10.18653/v1/2021.emnlp-main.595.
- [20] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, *ACM Transactions on Computing for Healthcare* 3 (2021) 2:1–2:23. URL: <https://dl.acm.org/doi/10.1145/3458754>. doi:10.1145/3458754.
- [21] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, O. Oktay, Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing, in: S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2022, pp. 1–21. doi:10.1007/978-3-031-20059-5_1.