Interactions between visual encoder architectures and attention mechanisms in image captioning: analysis of synergy, antagonism, and computational efficiency

Mateusz Bartosiewicz and Marcin Iwanowski

Institute of Control and Industrial Electronics, Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warsaw, Poland

Abstract. We present a systematic empirical study demonstrating that image captioning quality is governed less by architectural novelty and more by non-additive interaction effects between the visual encoder and the attention mechanism. By evaluating all pairwise combinations of seven CNN backbones and four representative attention paradigms on MS COCO dataset under a controlled protocol, we reveal pronounced patterns of synergy and antagonism. Our results show that some encoder–attention pairs significantly outperform expectations, while architectural mismatches substantially degrade performance. Crucially, we demonstrate that carefully matched classical encoder–decoder models can remain highly competitive while reducing parameter counts by over 40%, offering favourable accuracy–efficiency trade-offs. We conclude that interaction-centric optimisation, rather than blind architectural scaling, provides a viable path for sustainable, resource-constrained AI applications.

Keywords: Image captioning, Attention mechanisms; CNN backbones; Encoder–decoder interaction; Computational efficiency

1 Introduction

Automatic image captioning lies at the intersection of computer vision and language modelling, where a visual encoder maps an image into a spatial feature representation and a decoder generates a sentence conditioned on these features. Classical encoder–decoder pipelines based on CNN encoders and recurrent decoders remain widely used in practical systems, particularly when constraints on latency, memory footprint, or training resources preclude large-scale pretraining.

A key enhancement to the encoder–decoder paradigm is visual attention, which computes a time-varying visual context aligned with the decoder state and improves grounding by focusing on image regions relevant to the currently generated word [13]. Over the last decade, numerous attention formulations have been proposed–including soft attention [13], spatial variants [1], adaptive gating with a visual sentinel [8], and meta-attentional filtering such as Attention-on-Attention (AoA) [4]. In parallel, CNN backbones have evolved from early

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41

architectures to families with markedly different depth, connectivity patterns, and feature statistics [2,3,10].

Despite this breadth, attention mechanisms and visual encoders are typically evaluated in isolation or relative to a single fixed baseline. This implicitly suggests that captioning performance is driven primarily by the standalone "strength" of a component or by architectural novelty. However, captioning systems are modular and the decoder interacts with the encoder through attention; therefore, performance may be governed by *non-additive interaction effects* between the encoder feature distribution and the functional properties of the attention mechanism. In other words, an attention mechanism that performs well with one encoder may be suboptimal with another, even when both are strong individually.

In this work, we adopt an interaction-centric computational perspective and systematically quantify synergy and antagonism between visual encoders and attention mechanisms under a controlled protocol on MS COCO [7] dataset. We train all pairwise combinations of seven CNN backbones (Resnet, Densenet, InceptionV3, Regnet) and four representative attention methods (soft, spatial, adaptive, and self-attention with AoA filtering). Beyond aggregate captioning metrics (CIDEr, BLEU, ROUGE-L) [12,9,6], we analyse training dynamics, parameter efficiency, and inference-time sensitivity to beam search.

The contributions of this paper are threefold: (i) a controlled empirical study that isolates encoder–attention interaction effects across a comprehensive grid of configurations; (ii) evidence that compatibility, rather than component complexity alone, explains large fractions of performance variance, yielding clear cases of synergy and antagonism; (iii) practical guidelines identifying Pareto-efficient encoder–attention pairings that retain competitive accuracy while reducing model size and computational cost.

The remainder of the paper is organised as follows. Section 2 formulates the problem and introduces the interaction-centric evaluation perspective. Section 3 provides a functional characterisation of attention mechanisms and encoder feature distributions. Section 4 details the experimental setup and optimisation protocol. Section 5 reports the quantitative results, analyses interaction patterns, and discusses computational efficiency and limitations. Finally, Section 6 concludes the study.

2 Motivation and problem formulation

Contemporary image captioning research frequently emphasises architectural novelty, introducing increasingly complex attention mechanisms or larger encoder backbones. In most empirical studies, these components are evaluated independently or against a fixed reference baseline, implicitly suggesting that overall performance is an additive function of encoder capacity and attention design.

From a computational systems perspective, such an assumption is restrictive. An image captioning model is inherently modular: the visual encoder produces a spatial feature distribution, and the attention mechanism transforms this distribution into a context vector conditioned on the decoder state. The behaviour

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41

of the complete system therefore depends not only on the standalone quality of each module but also on their statistical and functional compatibility. Differences in spatial granularity, channel diversity, and feature homogeneity may substantially alter how attention weights are computed, filtered, and propagated to the language model.

This observation motivates an interaction-centric formulation. We treat the captioning architecture as a two-factor modular system composed of:

- a visual encoder E_i, defining the structure and statistics of visual features,
- an attention mechanism A_j, defining the aggregation and filtering strategy applied to these features.

For each encoder–attention pairing, we measure captioning performance $P(E_i, A_j)$ under a unified optimisation protocol. Rather than analysing components in isolation, we decompose performance into three conceptual terms:

$$P(E_i, A_j) = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

where μ denotes the global mean performance, α_i the main effect of encoder E_i, β_j the main effect of attention mechanism A_j, and γ_{ij} a non-linear interaction term capturing encoder–attention compatibility.

The central research question of this study is therefore:

To what extent is captioning performance determined by the interaction effects between the visual encoder and the attention mechanism, rather than by their individual capacities alone?

To answer this question, we conduct a controlled full-factorial evaluation across all pairwise combinations of selected CNN backbones and attention mechanisms on MS COCO. By isolating interaction effects from architectural novelty, the study aims to establish a computational framework for analysing modular compatibility in deep vision–language systems, with implications for both performance optimisation and efficiency-aware design.

3 Theoretical Background–Encoders and Attention Mechanisms

In encoder–decoder captioning systems, attention mechanisms function as structured operators that transform spatial visual feature maps into time-dependent context vectors. We evaluate four representative functional paradigms: soft attention, which performs deterministic smoothing of the spatial feature distribution [13]; spatial attention, which refines localisation precision while preserving compulsory visual grounding [1]; adaptive attention, which utilises a gating mechanism to selectively rely on visual input or linguistic memory [8]; and Attention-on-Attention (AoA), which employs a meta-filter to perform quality-aware filtering of aggregated features [4].

Simultaneously, the visual encoder defines the statistical structure–characterized by abstraction, heterogeneity, and channel diversity of the feature maps consumed by the attention mechanism. We analyse four distinct architectural families:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41

Residual networks (Resnet), providing stable, hierarchical representations [2]; DenseNet, whose extensive feature reuse results in high channel diversity [3]; InceptionV3, which emphasises multi-scale parallel processing [11]; and Regnet, representing design-space-optimised architectures with highly homogeneous feature statistics [10]. These functional differences suggest that captioning performance is governed by the interaction between encoder feature statistics and attention aggregation strategies.

4 Experimental setup

To isolate interaction effects, we evaluated all pairwise combinations of the four attention mechanisms [13,1,8,4] and seven CNN backbones [2,3,11,10]. Images were resized according to each backbone's requirements, and a learnable linear adaptation layer projected the varying spatial feature dimensions into a unified hidden state space. We employed two training paradigms: soft, spatial, and adaptive attention models were trained in two stages (freezing the encoder initially to prevent gradient shock and preserve semantic integrity) , whereas self-attention models with the AoA module were trained end-to-end from scratch. All models were optimised by minimising the cross-entropy loss using the Adam optimiser ($\beta_1 = 0.9, \beta_2 = 0.999$). We used a base learning rate of 4×10^{-4} for the decoder and attention modules, and a reduced rate of 1×10^{-4} for CNN fine-tuning. Models were trained for a maximum of 50 epochs. To stabilise convergence, we integrated an adaptive scheduler that decayed the learning rate if the validation BLEU-4 stagnated, alongside an early stopping mechanism. Experiments were conducted on the MS COCO dataset utilising the standard Karpathy split [5]. Caption quality was primarily assessed using CIDEr [12], supplemented by BLEU [9] and ROUGE-L [6]. For inference, we evaluated beam search decoding with widths $k \in \{1, 2, 3, 5, 8\}$ to analyse the robustness of encoder–attention interactions across different search spaces.

5 Results and Discussion

The quantitative results (summarised in Table 1 and Fig. 1) confirm that captioning performance is governed by non-linear interaction effects between visual encoders and attention mechanisms rather than being a strictly additive function of their individual capacities.

5.1 Synergy, Antagonism, and Efficiency

The strongest synergistic effect occurs when Regnet is paired with self-attention and the AoA module, achieving a peak CIDEr of 117.60 (a 6.27% relative improvement over the Resnet101 baseline). This demonstrates a high compatibility between Regnet's homogeneous features and AoA's meta-attentional filtering, which suppresses noise within uniform distributions. Conversely, severe antagonism emerges when Regnet is paired with soft attention (92.74 CIDEr, an 11.53%

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41

Table 1. Change of the value of the CIDEr metric considering various attention mechanisms. Columns %**Δ** C and %**Δ** p **(mln)** present the difference in the CIDEr value compared to the reference model, which is first row of each group, in italics.

Attention type	**Backbone**	**C**	p **(mln)**	%Δ **C**	%Δ p **(mln)**
Adaptive	*Resnet152*	*104.48*	*68.29*	–	–
	Regnet16	**106.97**	**91.73**	**2.38**	**34.32**
	InceptionV3	105.02	32.15	0.51	-52.91
	Densenet201	103.00	28.35	-1.42	-58.49
	Densenet161	100.54	36.96	-3.78	-45.88
	Densenet121	100.26	16.48	-4.04	-75.86
	Resnet101	99.54	52.64	-4.73	-22.91
Spatial	*Resnet152*	*102.38*	*68.29*	–	–
	Regnet16	**105.89**	**91.73**	**3.43**	**34.32**
	Densenet201	103.18	28.35	0.78	-58.49
	InceptionV3	102.46	32.15	0.08	-52.91
	Densenet161	100.40	36.96	-1.93	-45.88
	Resnet101	99.79	52.64	-2.53	-22.91
	Densenet121	99.77	16.48	-2.55	-75.86
Soft	*Resnet101*	*104.83*	*59.33*	–	–
	Densenet201	**108.96**	**34.66**	**3.94**	**-41.58**
	Resnet152	108.52	72.13	3.52	21.57
	InceptionV3	108.12	35.99	3.14	-39.33
	Densenet161	107.52	41.26	2.57	-30.46
	Densenet121	103.37	17.35	-1.39	-70.76
	Regnet16	92.74	85.13	-11.53	43.48
Self	*Resnet101*	*110.66*	*129.87*	–	–
	Regnet16	**117.60**	**168.93**	**6.27**	**30.08**
	Resnet152	108.92	145.51	-1.57	12.05
	Densenet201	100.65	105.33	-9.04	-18.89
	Densenet121	98.45	94.20	-11.04	-27.47
	Densenet161	95.30	114.01	-13.89	-12.22
	InceptionV3	84.51	112.48	-23.63	-13.39

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41

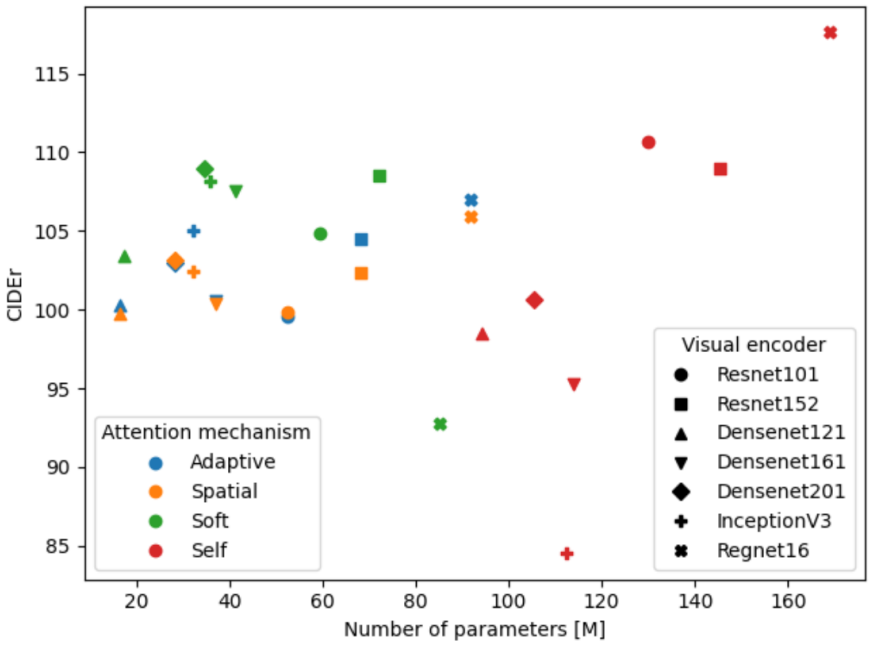

Fig. 1. Relationship between captioning performance and model complexity. The non-linear distribution of points highlights that performance is driven by architectural compatibility rather than parameter count alone.

drop), as deterministic averaging struggles without a filtering gate to process redundant representations. Similarly, self-attention with AoA collapses when paired with InceptionV3 (84.51 CIDEr, a 23.63% drop), proving that high-capacity attention cannot compensate for fundamental architectural mismatch.

Beyond peak accuracy, our analysis highlights significant opportunities for computational efficiency. Densenet and InceptionV3 encoders demonstrate a degree of self-sufficiency due to their rich spatial details and multi-scale processing. When paired with computationally lightweight soft attention, Densenet201 and InceptionV3 achieve highly competitive CIDEr scores (108.96 and 108.12, respectively) while reducing the parameter count by roughly 40% compared to heavier baselines. These Pareto-optimal configurations provide a viable path for sustainable AI and resource-constrained edge deployments.

5.2 Optimisation Dynamics and Inference Robustness

Analysis of training trajectories reveals that self-attention models trained end-to-end converge faster (averaging 18.9 epochs to peak performance) compared to the two-stage training of soft attention models (22.1 epochs). At inference time, while beam search ($k \in \{2, 3\}$) consistently outperforms greedy decoding and improves absolute scores, it does not alter the relative performance ranking

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41

of encoder–attention pairs. This stability confirms that the observed synergistic and antagonistic effects are intrinsic to the architectures and not artefacts of the decoding search space.

Limitations. We acknowledge that this study focuses on classical CNN encoders and recurrent decoders. Future work will extend this interaction-centric framework to fully attentive paradigms (e.g., ViT and Transformer decoders) to verify the persistence of these compatibility patterns.

6 Conclusions

This study demonstrates that image captioning performance is governed by non-linear synergy and antagonism effects dictated by the compatibility between visual feature distributions and attention strategies. Our results lead to three critical findings: (i) architectural compatibility supersedes individual component capacity, as advanced mechanisms like AoA require specific, homogeneous representations (e.g., Regnet) to avoid performance degradation; (ii) efficiency is an optimisation problem, where well-paired, established architectures can reduce parameter counts by over 40% without sacrificing accuracy – the non-linear relationship between complexity and performance (Fig. 1) highlights the diminishing returns of blind scaling; and (iii) the robustness is achieved not only by isolated modules, but also by careful calibration of the entire encoder-decoder inference chain.

As the field moves toward Large Visual Language Models, the principles of modular compatibility established here remain essential. While VLMs offer broad generalisation, specialised tasks requiring precise, controllable, and efficient visual grounding will continue to rely on interaction-aware architectures. This work provides a methodological blueprint for such designs, advocating for a shift from component-centric novelty to interaction-centric optimisation in deep vision–language systems.

Acknowledgments. The research was carried out on devices co–funded by WUT within the Excellence Initiative: Research University (IDUB) programme.

References

1. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6298–6306. IEEE Computer Society, Los Alamitos, CA, USA (2017). https://doi.org/10.1109/CVPR.2017.667
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE Computer Society, Los Alamitos, CA, USA (2016). https://doi.org/10.1109/CVPR.2016.90

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41

3. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. IEEE Computer Society, Los Alamitos, CA, USA (2017). https://doi.org/10.1109/CVPR.2017.243
4. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4633–4642. IEEE Computer Society, Los Alamitos, CA, USA (2019). https://doi.org/10.1109/ICCV.2019.00473
5. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137. IEEE Computer Society, Los Alamitos, CA, USA (2015). https://doi.org/10.1109/CVPR.2015.7298932
6. Lin, C.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
7. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
8. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3242–3249 (2017). https://doi.org/10.1109/CVPR.2017.345
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 311–318. ACL '02, Association for Computational Linguistics, USA (2002). https://doi.org/10.3115/1073083.1073135
10. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10425–10433 (2020). https://doi.org/10.1109/CVPR42600.2020.01044
11. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308
12. Vedantam, R., Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575. IEEE Computer Society, Los Alamitos, CA, USA (2015). https://doi.org/10.1109/CVPR.2015.7299087
13. Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. p. 2048–2057. ICML'15, JMLR.org (2015)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_41