Let there be a clock on the beach: Reducing Object Hallucination in Image Captioning

Ali Furkan Biten Lluis Gomez Dimosthenis Karatzas
Computer Vision Center, UAB, Spain
abiten, lgomez, dimos@cvc.uab.es

Abstract

Explaining an image with missing or non-existent objects is known as object bias (hallucination) in image captioning. This behaviour is quite common in the state-of-the-art captioning models which is not desirable by humans. To decrease the object hallucination in captioning, we propose three simple yet efficient training augmentation method for sentences which requires no new training data or increase in the model size. By extensive analysis, we show that the proposed methods can significantly diminish our models' object bias on hallucination metrics. Moreover, we experimentally demonstrate that our methods decrease the dependency on the visual features. All of our code, configuration files and model weights are available online[1].

1. Introduction

Thomas Kuhn [24] stated that discoveries in anomalies usually lead to new paradigms. Machine Learning (ML) in its early days relied on hand-coded/crafted features (alas simple, elegant ones are favored) to create models. However, there was an anomaly that performed better than hand crafted features and it led to a paradigm shift called the Unreasonable Effectiveness of Data [16], where they simply advised "to follow the data". Following the data was only the first step in the paradigm change, the second step was the amount of data. The introduction of big datasets like MSCOCO [26] and ImageNet [10] combined with the current advent in compute has resulted deep learning achieving significant feats [25]. Nevertheless, many works are published regarding the various failure cases and shortcuts that are exploited by deep models [14].

The shortcuts can be found especially in Vision and Language tasks such as Image Captioning and Visual Question Answering (VQA) in the form of object hallucination [33], language prior [15], focusing on background [5], spurious correlations [46], action bias [46], and gender bias [17].

[1] https://github.com/furkanbiten/object-bias

UD: A man on a beach with a surfboard
AoA: A man standing on a beach holding a frisbee
Ours (UD): A man standing on a beach near the ocean
Ours (AoA): A man standing on a beach with a clock

Figure 1: Standard approaches to image captioning are known to hallucinate on objects that do co-occur frequently, e.g. beach and frisbee or surfboard. Our method is capable of reducing object bias by normalizing the co-occurrence statistics, resulting in a reduction of hallucinated objects and the correct prediction of lower probability ones.

Solving the problem of object bias in image captioning is important for various reasons. First and foremost, describing an image while failing to correctly identify objects is not desirable to humans [33]. This is especially true for visually impaired people where they prefer correctness over coverage [29] for obvious reasons. Secondly, even though the results of the captioning models are pushed to the limit in evaluation metrics, this does not translate to a decrease in object bias/hallucination [33]. Finally, solving object hallucination is crucial for our models' generalization capabilities, allowing them to adapt easier to unseen domains.

It is obvious that hallucination cannot be corrected by collecting even more data from the same biased world. The co-occurrence patterns will not change or they will be mag-

nified. In other words, these biases do not seem to disappear neither with scaling up the dataset and nor with the increase in model size [14].

In this work, we demonstrate that it is possible to reduce the object bias without needing more data or increase in the model size while not affecting the model's computational complexity and performance. More specifically, we tweak any existing captioning models by providing object labels as an additional input and employ a simple yet effective sampling strategy which consist of artificially changing the objects in the captions, e.g. modifying the sentence "a *person* is playing with a dog" to "a *fork* is playing with a dog". Along with a change in the sentence, in a corresponding way we also replace the object labels provided to the model.

The reason is simple and can be traced back to co-occurrence statistics. By altering the co-occurrence statistics of the objects, we lessen the models' dependence on language prior and visual features as can be seen in Figure 1. Our contributions in this work are as follows:

- A simple method that can be applied to any captioning model to reduce object bias which requires no extra training data or increase in the model parameters.
- We improve the results on the hallucination metric CHAIR [33] while obtaining a boost over our baseline models on image captioning evaluation metrics.
- We demonstrate that our technique works with two commonly used loss functions, cross entropy and REINFORCE [32] algorithm.

2. Related Work

Following the advances of the encoder-decoder framework [8] with attention [4] in machine translation, automatic image captioning took off using similar architectures [39, 45]. The next advance in captioning came from using a pretrained object detector as feature extractor with two types of attention, top-down and bottom-up attention [3]. In parallel, it was demonstrated that training captioning models with the REINFORCE algorithm [42], optimizing the evaluation metrics directly, had benefits over using cross-entropy loss [32]. More recently, with the presentation of Transformers [36], a new family of models [19] achieved state-of-the-art results. Image captioning recently shifted into new directions such as generating diverse descriptions [38, 12, 44] by allowing both grounding and controllability [9, 47, 7] while using various contextual information [6, 35].

Nevertheless, despite the continuous improvement on the classic captioning metrics, there are many biases exploited, that produce biases in the models. To compensate for gender bias where the models are known to prefer a certain gender over the other in specific settings, [17] proposed to tweak the original cross-entropy loss with confidence/appearance loss. Another bias in captioning is related to action bias where certain actions are preferred over others described by [46] where they employ causality [31] into the captioning models. More specifically, their proposed method uses 4 layers LSTM with running expected average on ConceptNet [27] concepts for each word produced in the caption, which it introduces a significant computation overload. Similarly, [1] modifies the images with generative models to reduce the effect of spurious correlations in Visual Question Answering task.

[33] show that contemporary captioning models are prone to object bias. Moreover, they describe that evaluation metrics merely measure the similarity between the ground truth and produced caption, not capturing image relevance. Consequently, they propose two metrics to quantify the degree of hallucination of objects, namely CHAIRs and CHAIRi. CHAIR metrics evaluates how much our models produced wrong object labels at sentence level (hence CHAIRs) and at object level (hence CHAIRi). Surprisingly, the object hallucination problem has not received the attention it deserves. In this work, we try to diminish object bias without enlarging the model size or using extra data. We do so following a simple strategy that can be used with any model that accepts object detection features as inputs.

3. Methods

As mentioned previously, we try to reduce the object bias that exists inherently in existing models. The main cause of object bias is the systematic co-occurence of specific object categories in images of our training datasets, we therefore hypothesize that making the co-occurence statistics matrix more uniform will make our models hallucinate less. Accordingly, we devise a series of data augmentation techniques to achieve this goal.

3.1. A Small Tweak to Any Captioning Model

We would like to start off by giving a concise and general introduction to models in image captioning. After the introduction of top-down bottom-up attention [3], most of existing models for image captioning utilize object-level visual features extracted from an object detection network. More formally, given an image I, a set of bounding box features $V = \{v_1, v_2, ..., v_n\}$ are obtained by passing it through a pretrained object detector $\mathcal{O}$, *i.e.* $V = \mathcal{O}(I)$. These features are combined with an attention mechanism to be later fed into Language Models ($\mathcal{L}$) to generate a sentence $S = \{w_1, w_2, ..., w_k\}$ where most common variants of $\mathcal{L}$ are Transformers [36] and LSTM [18]. This formulation can be seen more clearly on the left part of Figure 2.

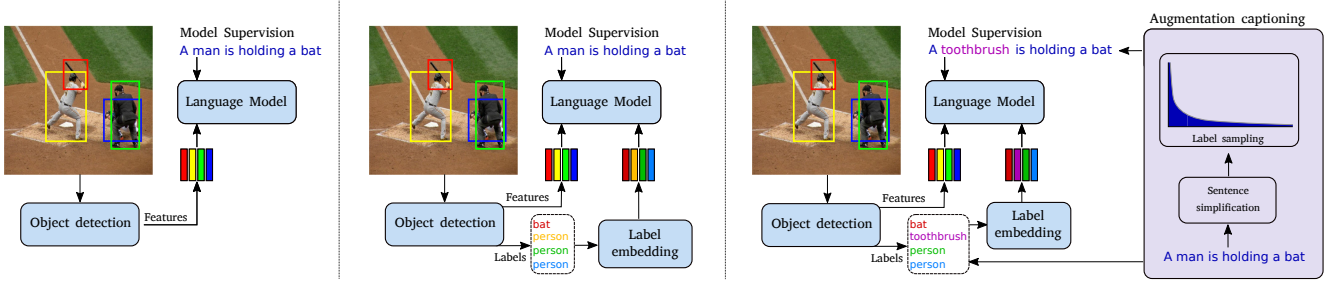

Figure 2: Most current models for image captioning utilize object-level visual features extracted from an object detection network (left diagram). In this paper we propose a simple tweak that consists of providing also the object labels as input (center diagram). The concatenation of label embeddings to visual features allows us to employ data augmentation techniques on the object labels and model supervision (captions) to fix the object bias in our models (right diagram).

$$\begin{aligned} \bar{V} &= Att(V, h) \\ P(w_j),\ h &= \mathcal{L}(w_j|\bar{V}, w_1, w_2, ..., w_{j-1}) \end{aligned} \tag{1}$$

Our tweak to the aforementioned formulation is to simply concatenate the object labels found in an image with bounding box features (middle part of the Figure 2). More formally, we extend the set of bounding box features from V to $\tilde{V} = \{v_1, v_2, ..., v_n, l_1, l_2, ..., l_i\}$ where l_i is the i^{th} embedded object label. After the concatenation, we replace V with $\bar{V}$ and follow exactly the same training procedure outlined in Equation (1).

Concatenation of label embeddings to visual features allows us to employ our data augmentation techniques. Since we use the labels as input to our models, we can directly alter them as we see fit. In the following sections, we describe the strategy behind the augmentation of labels.

3.2. Sentence Simplification

A first step to all our data augmentation methods is sentence simplification. By sentence simplification we refer to removing adjectives that are used in the captions for objects in the scene. As an example, we would like to modify the sentence "A small black cat is sitting on top of an old table" into "A cat is sitting on top of a table.". The reasons are twofold, one of which is that there are adjectives that can not hold true for every object, e.g. "small" and "black" can be used for a cat but this will not be correct when cat is artifically changed with another object such as elephant or banana. Secondly, simplifying the sentence in this way provides another variation of sentences, acting as type of a regularizer for captioning models to exploit language prior existing in the dataset.

To achieve this goal, we first analyze every caption with a Part-Of-Speech (POS) and find all the noun phrases corresponding to sentences. However, these noun phrases do not necessarily have to refer to objects found in an image. That is why, we make use of synonyms list for object classes that exist in the dataset (e.g. 80 objects in MSCOCO) and filter the noun phrases that include the object name or its synonyms. As a final step, we replace the whole noun phrase with the root of the phrase.

3.3. Augmentation of Sentences

After simplifying the sentences, we employ different sampling strategies to pick which object to replace. More formally, given a sentence containing objects o_i and o_j, we sample object o_k to replace o_j according to the distribution $P(o_k|o_i)$. Now, we explain in detail which distributions we use to augment the sentences.

3.3.1 Uniform Sampling

The choice of uniform sampling is inspired by our hypothesis on creating a uniform object label co-occurrence matrix. In its most simplest form, we make use of uniform distribution for sampling, where

$$P(o_k|o_i) = P(o_k) = 1/N. \tag{2}$$

In other words, every object has an equal probability to be sampled where dataset statistics are disregarded. The next two distribution takes into account the discarded dataset statistics.

3.3.2 Inverse Multinomial Sampling

The most accessible statistics one can obtain regarding any given dataset is the co-occurrence matrix $M \in R^{N \times N}$ where M_{ij} refers to the co-occurrence statistics of objects o_i and o_j and N is the number of objects. We define a new distribution which considers dataset statistics called inverse multinomial by making use of M where

$$P(o_k|o_i) = \frac{1}{\tilde{M}_{ik}} \text{ where } \tilde{M}_{ik} = \frac{M_{ik}}{\sum_k M_{ik}} \tag{3}$$

With inverse multinomial, we sample object o_k if the occurence is low with object o_i. On the other hand, if object o_k and o_i co-occurs frequently in the dataset, then the probability of selecting o_k will be quite low.

3.3.3 Updating Co-Occurence Matrix

Although inverse multinomial sampling increases the chance of low frequency pairs to be sampled, it does not prevents creating a new bias for low frequency pairs. To circumvent the problem, we determine to keep track of the matrix M and constantly update according to the sampled pair. More formally, the distribution is defined as:

$$P(o_k|o_i) = \frac{1}{\tilde{M}_{ik}} \text{ where } \tilde{M}_{ij} = \frac{M_{ij}}{\sum_j M_{ij}} \qquad (4)$$
$$M_{ik} = M_{ik} + 1, M_{ij} = M_{ij} - 1$$

By keeping track of co-occurence statistics in training diminishes the prospect of models finding a shortcut as well as allowing faster convergence to a uniform M.

4. Experiments

4.1. Dataset and Baseline Models

MSCOCO: [26]. We use the most commonly used captioning dataset, MSCOCO [26]. We follow the literature on using the 'Karpathy' split [21]. The split contains 113,287 training images with 5 captions each and 5k images for validation and testing.

Evualation Metrics: To evaluate caption quality, we report the standard automatic evaluation metrics; CIDEr [37], BLEU [30], METEOR [11], SPICE [2]. Moreover, we include the new metric called SPICE-U [41] which is a variant of SPICE where it rewards for uniqueness of sentences. Finally, we provide the hallucination metrics CHAIRs [33] and CHAIRi [33] for sentence and object level, respectively. In CHAIR metrics, lower is better.

UpDown (UD): [3]. The bottom-up and top-down attention model utilizes the salient image regions proposed by object detector pretrained on VG [23] and then weighting the regions by employing an attention mechanism calculated according to Language Models' hidden state.

AoA: [19]. The attention on attention model extends the conventional Transformers [36] model by including another attention to determine the relevance between attention results and queries. When we train with object labels given as inputs, we refer those models as UD-L and AoA-L.

4.2. Implementation Details

All our models are implemented on top of publicly available code[2]. We use Adam [22] optimizer with batch size 10 and learning rate 0.0002 and 0.0005 for UpDown [3] and AoA [19], respectively. Both models are trained for 30 epochs and we kept the best models according to best score on validation set on Cider-D [37]. We generate sentences with no beam search and both models use visual features provided by [3]. For embedding the object labels, we utilize FastText [20].

We use both of the commonly used training losses employed by the literature, namely cross-entropy and REINFORCE [32]. For every variant of our model, we randomly choose to use original sentences or augmented sentences according to flip of a coin as ground-truth. All the models trained with our augmentation are fine-tuned to allow faster convergence and to see if we can reduce the "learned" biases of our models. Finally, we always use the ground truth object labels as input to our models and use X101-FPN from Detectron2 [43] library to obtain object labels for testing. All the code, model weights and configuration file necessary for the hyper-parameters will be released upon acccceptance.

4.3. Comparison to State of Art

We present the results of our models as well as the state-of-the-art model results in 1. First and foremost, UD-VC ad AoA-VC (row 1.1, 1.2) uses the features extracted from state of the art object detector while concatenating with the original features provided by UpDown [3], *i.e*. they use 2 FasterRCNN architecture in their model training. While UD-DIC (row 1.3) uses 4 deep LSTMs [18] to find matching between produced words and ConceptNet [27] labels. Moreover, UD-MMI (1.4) and AoA-MMI (1.5) trains an LSTM without any visual features to detect the common and not unique sentences and later to be used at inference time. From aforementioned models, we observe that beating state-of-the-art results or increase in the model size or even using better features does not result in our models hallucinating less.

Remark 1 *Increase in the model size (parameters) or boost in the image captioning metrics does not result in decrease in CHAIR metrics.*

Subvariant of this conclusion can be also seen in REINFORCE [32] training. It is common practice in captioning community to train the models first with cross entropy and then with self-critical loss [32] on CIDER-D [37]. While this training ensures a significant boost on the automatic metrics especially on CIDER, it makes our models hallucinate more (can be seen in row 1.7, 1.8, 1.11, 1.13 and 1.18).

Remark 1.1 *Self-Critical training leads to increase in the captioning metrics while making models hallucinate more.*

[2] https://github.com/ruotianluo/self-critical.pytorch

Table 1: Results of image captinoning models on Karpathy test split. * numbers are provided by [33] with beam search 5. B-4: Bleu-4, M: Meteor, C: Cider, S: Spice, S: Spice-U, CHs: CHAIRs, CHi: CHAIRi, UD: UpDown, AoA: Attention on Attention, Uni: Uniform Sampling, Inv: Inverse Multinomial Sampling, Occ: Co-occurence Updating. In CHAIR metrics, lower is better.

		Cross Entropy							Self Critical						
	Model	**B-4** ↑	**M** ↑	**C** ↑	**S** ↑	**CHs** ↓	**CHi** ↓	**S-U** ↑	**B-4** ↑	**M** ↑	**C** ↑	**S** ↑	**CHs** ↓	**CHi** ↓	**S-U** ↑
1.1	UD-VC [40]	39.5	29	130.5	-	10.3	6.5	-	-	-	-	-	-	-	-
1.2	AoA-VC [40]	39.5	29.3	131.6	-	8.8	5.5	-	-	-	-	-	-	-	-
1.3	UD-DIC [46]	38.7	28.4	128.2	21.9	10.2	6.7	-	-	-	-	-	-	-	-
1.4	UD-MMI [41]	22.77	28.84	106.42	20.72	7.8	-	25.27	-	-	-	-	-	-	-
1.5	AoA-MMI [41]	27.18	30.39	128.15	22.81	9.28	-	26.53	-	-	-	-	-	-	-
1.6	DiscCap [41]	21.58	27.42	110.9	20.27	10.84	-	24.52	-	-	-	-	-	-	-
1.7	LRCN [13]*	-	23.9	90.8	17.0	17.7	12.6	-	-	23.5	93.0	16.9	17.7	12.9	-
1.8	FC [32]*	-	24.9	95.8	17.9	15.4	11	-	-	25	103.9	18.4	14.4	10.1	-
1.9	Att2In [32]*	-	25.8	102	18.9	10.8	7.9	-	-	25.7	106.7	19	12.2	8.4	-
1.10	UD [3]*	-	27.1	113.7	20.4	8.3	5.9	-	-	27.7	120.6	21.4	10.4	6.9	-
1.11	NBT [28]*	-	26.2	105.1	19.4	7.4	5.4	-	-	-	-	-	-	-	-
1.12	GAN [34]*	-	25.7	100.4	18.7	10.7	7.7	-	-	-	-	-	-	-	-
1.13	UD	33.2	26.9	108.4	20.0	10.1	6.9	24.05	36.5	27.8	121.5	21.3	11.9	7.7	23.85
1.14	UD-L	34.4	27.3	112.7	20.7	6.4	4.1	24.68	37.7	28.6	124.7	22.1	5.9	3.7	25.41
1.15	UD-L + Uni	34.2	27.2	112.4	20.6	6.3	4.0	24.61	37.6	28.7	125.2	22.3	**5.8**	**3.7**	25.54
1.16	UD-L + Inv	34.3	27.3	112.6	20.7	6.2	4.0	24.05	37.8	28.7	125.4	22.3	5.9	3.8	25.60
1.17	UD-L + Occ	33.9	27.0	110.7	20.3	**5.9**	**3.8**	24.52	37.7	28.7	125.2	22.2	**5.8**	**3.7**	25.58
1.18	AoA	33.7	27.4	111.0	20.6	9.1	6.2	24.57	38.8	28.7	127.2	22.4	9.6	6.1	24.68
1.19	AoA-L	33.1	27.0	110.0	20.3	7.1	4.4	24.30	35.9	28.0	119.6	21.7	7.8	4.8	24.81
1.20	AoA-L + Uni	34.1	27.2	111.4	20.5	**6.2**	**3.9**	24.58	35.1	27.8	117.7	21.4	7.3	4.5	24.58
1.21	AoA-L + Inv	34.3	27.3	112.0	20.6	6.5	4.1	24.93	35.7	28.0	119.2	21.8	7.5	4.6	24.93
1.22	AoA-L + Occ	34.3	27.1	111.3	20.5	**6.2**	**3.9**	24.57	34.5	27.5	116.0	21.1	**7.0**	**4.3**	24.20

The next point we would like to move to is regarding our methods starting from 1.13. Simply by adding the object labels as input we notice an improvement on CHAIR metrics for both of the models. This progress can also be observed on classic image captioning metrics for UpDown model. Furthermore, we note that addition of labels also reaches the reported numbers in 1.10 while significantly diminishing the object bias on both sentence and object level. Finally, we see that this simple technique of concatenating object labels and visual features already gives state-of-the-art results in object hallucination by around 1 to 4%.

Remark 2 *Merely concatenating the labels with visual features results in the decline of hallucination of our models while beating state-of-the-art models on CHAIR metrics.*

Before we focus on our augmentation techniques, we want to point out that reducing object hallucination from 10% to 6% is not an equivalent to reducing it from 6% to 2%. The reason is that there are 2 different elements affecting the hallucination, one of which is the dataset bias which what we are trying to solve and the other one is the noisy and incorrect FasterRCNN features. From the next section, we see that our methods upper bound is around 2-3%, suggesting that the rest of the hallucination is mostly coming from the visual features. That said, it can be seen that we even better the results of row 1.14 and row 1.19 in comparison to our proposed techniques by around 0.5 to 1%.

Remark 3 *We demonstrate that our proposed techniques can reduce the object bias on the same model architectures.*

Furthermore, we remark that we always obtain the best results on CHAIR metrics with the Co-Occurrence Updating technique although we usually obtain a decrease in the other common metrics. We also see that Inverse Multinomial sampling results in the best performance in the classic captioning metrics. Moreoever, Co-Occurence Updating always achieves the best CHAIR scores out of all the different sampling.

Remark 3.1 *Inverse Multinomial achieves the best scores on standard captioning metrics while Co-Occurence updating performs the best on CHAIR metrics.*

Finally, we report the recently introduced metric SPICE-U [41] where it evaluates how unique and informative a caption is. We take interest in the said metric since our concern was that proposed augmentation can make captioning models produce more repetitive or less informative captions because of the sentence simplification. As can be observed from the 1, even in the cases where we have a drop on standard image captioning metrics, we still improve on SPICE-

Table 2: Results on Karpathy Test split. The numbers are obtained by using ground truth object labels instead of using object detector.

		Cross Entropy						Self Critical					
Model	**Aug**	**Bleu-4** ↑	**METEOR** ↑	**CIDEr** ↑	**SPICE** ↑	**CHAIRs** ↓	**CHAIRi** ↓	**Bleu-4** ↑	**METEOR** ↑	**CIDEr** ↑	**SPICE** ↑	**CHAIRs** ↓	**CHAIRi** ↓
UD	-	**34.6**	**27.4**	112.9	20.8	4.5	2.8	37.9	28.7	125.9	22.3	**3.5**	**2.2**
UD	U	**34.6**	**27.4**	113.4	20.8	4	2.5	38.0	28.9	126.2	22.5	3.7	2.3
UD	IM	34.5	**27.4**	**114.0**	**20.9**	3.9	2.4	**38.0**	**28.8**	**126.4**	**22.5**	3.9	2.4
UD	Occ	34.0	27.1	111.6	20.5	**3.6**	**2.2**	**38.0**	**28.8**	**126.4**	**22.5**	**3.5**	**2.1**
AoA	-	33.4	27.2	111.4	20.5	4.4	2.7	36.2	28.3	121.3	22.0	4.3	2.6
AoA	U	34.4	27.3	112.5	20.7	**2.7**	**1.6**	35.5	**28.0**	**119.2**	**21.7**	3.9	2.3
AoA	IM	**34.6**	**27.4**	**113.4**	**20.8**	3.1	1.9	**36.1**	**28.3**	**121.0**	**22.0**	3.9	2.3
AoA	Occ	34.4	**27.4**	113.0	20.7	**2.7**	**1.6**	34.9	27.7	117.4	21.3	**3.7**	**2.2**

U. In row 1.14-1.17, we even have 2% improvement in self-critical training. This is quite encouraging especially compared to SOTA numbers on row 1.4-1.6 where we even beat the numbers without the need of training an extra LSTM.

Remark 4 *Our techniques can improve or at least stay the same as the base model on producing informative and unique captions.*

4.4. What if we have perfect label extractor?

We try to figure out the upper bound for our techniques. In other words, since it is known that object detectors are far from providing the perfect labels, we test our methods with the ground truth annotations of object labels to see the full performance of our different methods, given in Table 2. We use the same models provided in Table 1.

First conclusion is that we see an improvement on all the metrics with the usage of ground truth. This is quite expected since we have trained with the ground truth annotations.

Remark 5 *With the perfect object detector, we can improve on all the metrics.*

One important remark is that the gap between the models with labels and models trained with our augmentation is much bigger. Particularly, for UpDown we see the gap becomes 0.9% and 0.6% while for AoA, it is 1.7%, 1.1% on CHAIRs and CHAIRi, respectively. This suggests that our proposed augmentation will reach even higher values with the advances on object detector's performance.

Remark 5.1 *Our proposed methods can achieve higher performance simply by obtaining more precise labels.*

Finally, it can be appreciated that in all of the models whether trained with cross-entropy or self-critical, Co-Occurrence Updating always accomplishes the best scores on CHAIR metrics, confirming our hypothesis on creating a uniform co-occurence matrix causing a decrease on object bias.

Remark 5.2 *By making the co-occurrence matrix uniform causes our models to have* ***the least*** *object bias.*

			FRCNN		Ground Truth	
	Vis Feat	**Labels**	**CHAIRs**	**CHAIRi**	**CHAIRs**	**CHAIRi**
UD-L	✓	✗	9.2	6.6	-	-
UD-L + Uni	✓	✗	9.4	6.7	-	-
UD-L + Inv	✓	✗	9.2	6.6	-	-
UD-L + Occ	✓	✗	9.8	7.1	-	-
UD-L	✗	✓	35.8	29.1	35.7	28.7
UD-L + Uni	✗	✓	26.1	18.8	24.7	17.3
UD-L + Inv	✗	✓	29.2	21	28	19.8
UD-L + Occ	✗	✓	20.2	13.6	17.1	11.2

Table 3: Results on Karpathy Test split. We either provide to our models only visual features or object label embeddings.

4.5. Data augmentation effect on the models

Our next set of experiments is done to find out what is the proposed augmentation provides to the model. To take a stab at the problem, we decided to zero out either the visual features or the object labels at inference time to see how much importance they have on hallucination. Our numbers can be seen in Table 3. Primarily, we appreciate that the results are much better when using visual features than when using object labels. This is anticipated and can be thought as taking away the "eyes" of the model. However, we identify that visual features holds more significance for UD-L than the models trained with the augmentation (UD-L+Occ and UD-L+Uni).

Remark 6 *The proposed training leads to models to put more emphasis on the labels while reducing the dependence on visual features.*

Moreover, it can be appreciated that our model trained with Co-Occurence Updating puts less importance on the visual features or utilizes it to a lesser degree than the other models. This point is especially reinforced when we examine the zeroing out the visual features. We recognize that models trained with our augmentations utilizes much more the provided labels, in which from UD-L to UD-L+Occ,

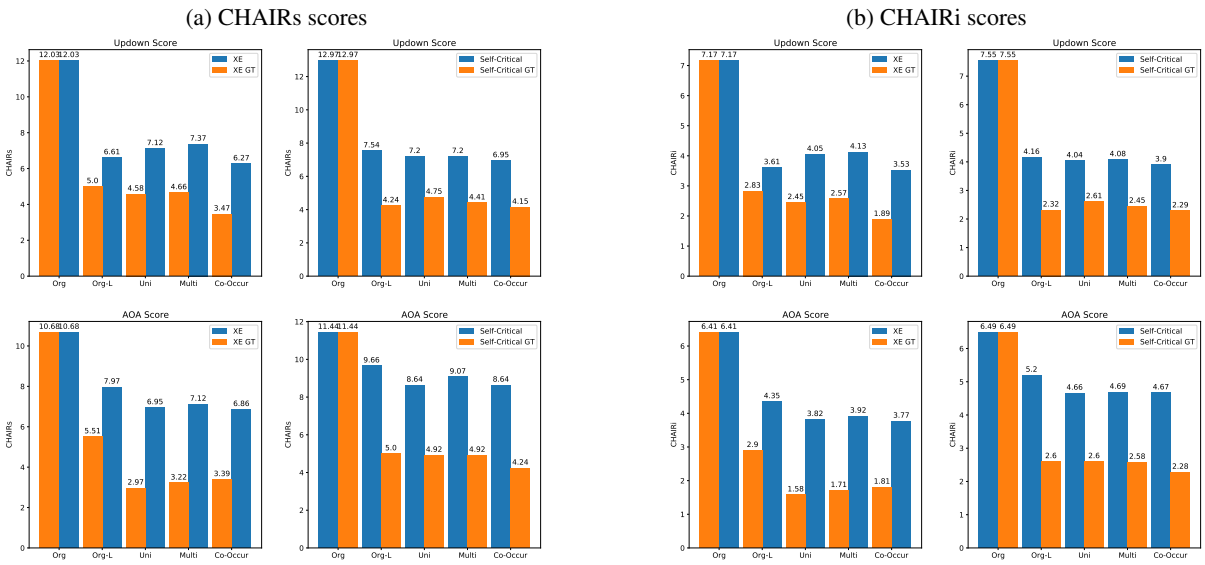

Figure 3: Bar plot on low frequency pairs. We provide all the models we trained with object detector labels and ground truth labels. We select the sentences which contain objects pairs that has less than 200 co-occurrence.

there is a 15% improvement. Another evidence for the statement is that in UD-L from object detection labels to ground truth, there is simply 0.1%, 0.4% improvement. Furthermore, we can even see that this gap grows even more when the ground truth is used as input to our models. We notice that the same difference when used ground truth raises to 18% and 17% for CHAIRs and CHAIRi, respectively.

Remark 6.1 *Co-Occurrence Updating exploits the labels the most out of other 3 models.*

4.6. Captioning with uncommon object pairs

To further investigate our proposed formulation, we provide Figure 3 for all of our models in 1.13- 1.22. In Figure 3, we calculated the CHAIRs (Figure 3a) and CHAIRi (Figure 3b) for pairs of objects with low co-occurrence. For this we filtered those images of the MSCOCO dataset with pairs of objects with a co-occurrence lesser than 200. This accounts for 23.6% of the MSCOCO test set. It can be recognized that original models UD (1.13) and AoA (1.18) have a much higher object bias on low frequency pairs than the other ones, an increase around 2% for both models on CHAIRs and 0.2%, 0.3% on CHAIRi for UD and AoA. In addition, we see much better numbers on UD-L and AoA-L, so simple concatenation of labels lowers the object bias. Additionally, with the utilization of perfect labels (orange bars in Figure 3), we appreciate that we obtain even better numbers on low frequency object pairs than the overall numbers calculated in Table 2. This suggests that our proposed augmentation can handle well on low frequency object pairs whether trained with cross entropy or self-critical.

As well, we notice that the gap between the original models and Co-Occurrence Updating is bigger on the low frequency pairs. Therefore, our hypothesis on making co-occurrence matrix as uniform as possible making object bias lower holds valid.

4.7. Ablation Study

	SS	CHAIRs [33]	CHAIRi [33]
UD-L + Uni	✗	6.3	4.1
UD-L + Inv	✗	6.3	4
UD-L + Occ	✗	6.5	4.2
UD-L + Uni	✓	6.3	4
UD-L + Inv	✓	6.2	4
UD-L + Occ	✓	**5.9**	**3.8**

Table 4: Ablation results on sentence simplification.

Our final experimentation is on the analysis of sentence simplification. To see if the suggested formulation for sentence simplification has any effect on object bias, we decided to run UpDown model with and without sentence simplification. Our results can be found in Table 4.

As can be appreciated from the Table 4, sentence simplification does not seem to have a lot of effect on Uniform and Inverse Multinomial sampling. Even though we always get better results on using sentence simplification, we only obtain 0.1% which can be accounted for randomness.

However, sentence simplification has a significant effect on Co-Occurrence Updating. Our conjecture regarding this phenomena is that since Co-Occurrence Updating se-

UD: A dog is sitting in the grass near a lake	**UD**: A man is jumping a street on a skateboard	**UD**: A young child holding a remote control in his hand	**UD**: A group of people on a beach with a kite	**UD**: A woman is looking at a cell phone
AoA: A dog running in a field near a body of water	**AoA**: A man is in the air on a skateboard	**AoA**: A baby holding a remote control in its hand	**AoA**: A man standing under a beach umbrella on top of a beach	**AoA**: A woman holding a cell phone in her hand
Ours (UD): A horse is sitting in the grass near a lake	**Ours (UD)**: A man is doing a trick on a traffic light	**Ours (UD)**: A baby is holding a cell phone in its mouth	**Ours (UD)**: A man is standing on the beach with a surfboard	**Ours (UD)**: A person is holding a pair of scissors
Ours (AoA): A horse running in a field near a body of water	**Ours (AoA)**: A man is jumping the air on a traffic light	**Ours (AoA)**: A little girl holding a cell phone in her hand	**Ours (AoA)**: A group of people standing on top of a beach	**Ours (AoA)**: A woman is a pair of scissors in a brown

Figure 4: Some qualitative samples from our baselines and Co-Occurence Updating models, referred as ours.

lects more numbers of various pairs than the other two samplings, the models find a correlation between the adjectives and the replaced objects. As an example, usage of little or cute is usually adopted for boys or girls. When we replace the phrase "cute little boy" first with "cute little broccoli" and later with "cute little clock". The model will learn to associate "cute little" phrase first with broccoli and then with clock. However, in Uniform Sampling the models will merely discard this association because of the uniformity in nature and in Inverse Multinomial, only a handful of pairs will be associated with the phrase. Which is why we don't see a lot of disruption in Uniform and Inverse Multinomial.

4.8. Qualitative Results

Last but not least, we present some interesting qualitative samples in Figure 4. Our first remark is that our models outperform the baselines in two ways, one of which is the deletion of hallucinated objects. This behaviour can be observed in the third and forth column where the baseline models predicted a surfboard, frisbee, beach umbrella or kite. These examples show the strong language prior that our models exploit.

On the other hand, our models also outperform the baselines in that they not only delete the incorrect objects but also replace it with the correct one. As an example, in the first (or second) column of Figure 4, while baseline models predict a dog (skateboard), our models corrects it to a horse (traffic light). One important remark is that sentences' verb or action prediction stays the same, *e.g.* sitting, running, jumping, in which it calls for an augmentation technique for actions as well.

Finally, we see that even in the case of wrongly produced captions (see fifth column), our models can still identify the correct object however they are constrained by the language models.

5. Conclusion

Since describing an image with a failure to correctly identify objects is not desirable to humans, we focus at the object bias in image captioning models. To reduce object hallucination in image captioning, we propose 3 different sampling techniques to augment sentences to be treated as ground truth to train image captioning models. By extensive analysis, we show that the proposed methods can significantly diminish our models' object bias on hallucination metrics. Also, we demonstrate that our methods can achieve much higher scores with the advances on object detectors. Moreover, we identify that our suggested techniques makes the models depend less on the visual features and by making co-occurrence statistics of objects uniform, and resulting in models generalizing better. But more importantly, we show that it is possible to decrease the object bias without needing extra data/annotations or increase in the model size or the architecture. Our hope is that this study incites more research on simple but effective methods to train deep models while keeping the model complexity untouched.

Acknowledgments

This work has been supported by projects PID2020-116298GB-I00, the CERCA Programme / Generalitat de Catalunya, AGAUR project 2019PROD00090 (BeARS) and PhD scholarship from UAB (B18P0073).

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 2016. 4.1

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998*, 2017. 2, 3.1, 4.1, 4.2, 4.3, 1

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. pages 1–15, 2014. 2

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018. 1

[6] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019. 2

[7] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9962–9971, 2020. 2

[8] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014. 2

[9] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1

[11] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Workshop on Statistical Machine Translation*, 2014. 4.1

[12] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David A Forsyth. Diverse and controllable image captioning with part-of-speech guidance. 2018. 2

[13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 1

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020. 1, 1

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 1

[16] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. 1

[17] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 1, 2

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 3.1, 4.3

[19] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019. 2, 4.1, 4.2

[20] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016. 4.2

[21] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017. 4.1

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4.2

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 4.1

[24] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012. 1

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014. 1, 4.1

[27] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. 2, 4.3

[28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 1

[29] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999, 2017. 1

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. BLEU: a method for automatic evaluation of machine translation. *Annual Meeting on Association for Computational Linguistics*, 2002. 4.1

[31] Judea Pearl. *Causality*. Cambridge university press, 2009. 2

[32] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 4.2, 4.3, 1, 1

[33] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1, 1, 2, 4.1, 1, **??**, **??**

[34] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017. 1

[35] Alasdair Tran, Alexander Mathews, and Lexing Xie. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13035–13045, 2020. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2, 3.1, 4.1

[37] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 4.1, 4.2, 4.3

[38] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2017. 2

[39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[40] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 1, 1

[41] Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. Towards unique and informative captioning of images. *arXiv preprint arXiv:2009.03949*, 2020. 4.1, 1, 1, 1, 4.3

[42] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 2

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4.2

[44] Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12637–12646, 2021. 2

[45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, 2015. 2

[46] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 1, 2, 1

[47] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*, pages 211–229. Springer, 2020. 2