Quality-agnostic Image Captioning to Safely Assist People with Vision Impairment

Lu Yu[1*] , **Malvina Nikandrou**[2] , **Jiali Jin**[1] , **Verena Rieser**[2,3†]

[1]Tianjin University of Technology, Tian jin, China
[2]Heriot-Watt University, Edinburgh, United Kingdom
[3]Alana AI
luyu@email.tjut.edu.cn, {mn2002,v.t.rieser}@hw.ac.uk

Abstract

Automated image captioning has the potential to be a useful tool for people with vision impairments. Images taken by this user group are often noisy, which leads to incorrect and even unsafe model predictions. In this paper, we propose a quality-agnostic framework to improve the performance and robustness of image captioning models for visually impaired people. We address this problem from three angles: data, model, and evaluation. First, we show how data augmentation techniques for generating synthetic noise can address data sparsity in this domain. Second, we enhance the robustness of the model by expanding a state-of-the-art model to a dual network architecture, using the augmented data and leveraging different consistency losses. Our results demonstrate increased performance, e.g. an absolute improvement of 2.15 on CIDEr, compared to state-of-the-art image captioning networks, as well as increased robustness to noise with up to 3 points improvement on CIDEr in more noisy settings. Finally, we evaluate the prediction reliability using confidence calibration on images with different difficulty / noise levels, showing that our models perform more reliably in safety-critical situations. The improved model is part of an assisted living application, which we develop in partnership with the Royal National Institute of Blind People.

1 Introduction

Vision and Language technologies, such as image captioning, have the potential to support people who suffer from visual impairment to live more independent lives by describing the visual world around them in natural language. Increasing independence and inclusion for people with disabilities contribute to the United Nations Sustainable Development Goals *Good health and well-being* and *Reduce inequalities*, as well as the UN Sustainable Development Principle *Leave no one behind*.

---

* Work done at Heriot-Watt University.
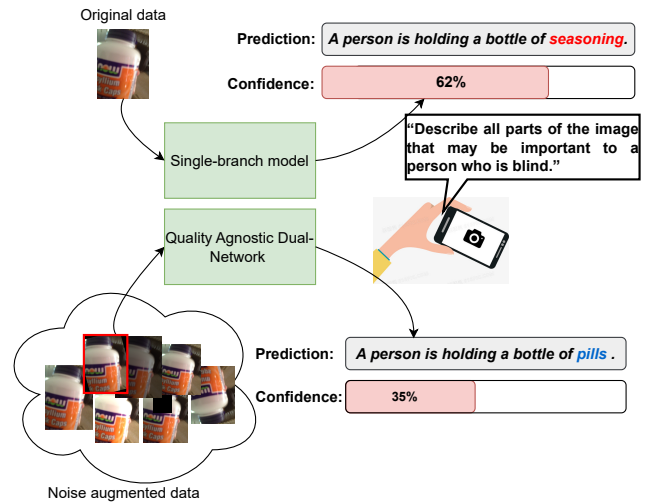† Now at Google DeepMind.

Figure 1: Example demonstrating the importance of robust and reliable models for people with vision impairments: the single-branch baseline model labels a medication container as a 'bottle of seasoning', whereas our proposed model, which extends the baseline to a dual-branch model with quality-agnostic data argumentation and consistency regularization, correctly describes it as a 'bottle of pills'. The baseline is over-confident in its prediction, making the mistake hard to detect.

Image captioning is an active area of research where deep neural networks are trained on large-scale datasets, such as MS-COCO [Lin *et al.*, 2014] and Flickr [Young *et al.*, 2014]. The vast majority of collected images in these datasets are clean and high-quality, which is a reasonable assumption for commercial applications such as image indexing or social media. However, people who suffer from visually impairment are one important stakeholder group which is not sufficiently represented by these datasets. Images taken by people with visual impairments often exhibit high noise levels introduced by the inability of the photographer to perceive the target object. This results in a shift in distribution, which makes standard image captioning approaches less robust [Gurari *et al.*, 2020]. This shortfall in performance combined with over-confident predictions can result in safety critical situations for vulnerable users, e.g. in the case of medication packaging [Davis *et al.*, 2020], cf. Example in Figure 1.

The VizWiz-Captions dataset [Chiu *et al.*, 2020b] aims to alleviate this lack of in-domain data by releasing the first image captioning dataset, containing images taken by people with vision impairments. However, compared to the standard datasets used for image captioning, VizWiz-Captions is relatively small – containing 39,000 images each paired with five captions. The moderate amount of data in combination with the high amounts of diverse noise make this datasets especially challenging for training image captioning models.

We address this challenge in a joint project with the Royal National Institute of Blind People. In the following, we introduce a new framework which is agnostic to image quality, i.e. the model predicts the same label for the same image content irrespective of noise. This framework is part of an assisted living application, which we will demonstrate at the conference.

In particular, we first extend the dataset by augmenting it with different types of distortion mimicking the noise observed in the original VisWiz data. Next, we take the top performing benchmark model from [Gurari *et al.*, 2020] – the Attention on Attention Network (AoANet) [Huang *et al.*, 2019] – and extend it to a dual-network architecture, where one branch explicitly models noise. We experiment with three types of consistency losses to coordinate the training signal between the originally labeled and the noise augmented branch.

Finally, we introduce a safety-focused evaluation framework for this task, which evaluates whether the model's confidence scores accurately reflect the likelihood of being correct. Accurate confidence scores are important to determine whether or not to issue an image caption, or at least indicate uncertainty in the prediction, which is essential in safety critical situations with vulnerable users.

Thus, our scientific contributions are along the full modelling pipeline of data, model and evaluation. All resources will be released with the final version of this paper.

2 Related Work

While most Vision and Language (V+L) technology is catered to the needs of the average population, there has been an increasing interest in using it in assistive settings, e.g. supporting people with vision impairments [Bennett *et al.*, 2018; MacLeod *et al.*, 2017; Pantazopoulos *et al.*, 2021; Tseng *et al.*, 2022; Chen *et al.*, 2022] and new datasets are created to capture the requirements of this population [Bigham *et al.*, 2010; Gurari *et al.*, 2020; Chiu *et al.*, 2020b]. However, there are still some main challenges to overcome with respect to performance, robustness, and reliability of these models, including dataset size, image quality and general safety concerns, which we will discuss in the following.

Alternatives to Pretraining. Datasets gathered with and for visually impaired people are typically small. The prevalent paradigm to deal with data sparsity in V+L tasks is to fine-tune large, pretrained models on the target domain. However, pretraining is not always effective. [He *et al.*, 2019], for example, show that ImageNet pretraining has limited impact on COCO object detection. Similarly, [Gurari *et al.*, 2020] show that pretraining on large vision datasets, includ-

ing COCO and ImageNet, has limited benefit for models finetuned on the VisWiz-Captions. [Zoph *et al.*, 2020] argue that self-training and data augmentation are powerful alternatives to pretraining, which we will further explore in this paper. In particular, we use data augmentation to increase robustness to noisy images and self-training via *consistency regularization*, which we investigate with and without the pretraining-finetuning paradigm. While previous work on consistency regularization primarily focuses on vision-only models – such as FixMatch [Sohn *et al.*, 2020], AlphaMatch [Gong *et al.*, 2021] and SimPLE [Hu *et al.*, 2021] – we will adapt this framework to V+L image captioning, where we we map synthetically designed noising techniques to real-world noise as introduced by images taken by visually impaired people.

Safety Concerns. Applying technology in the context of assistive technology requires an increased awareness of and concerns for the safety and well-being of vulnerable users – in our case visually impaired people seeking assistance in their daily lives. While applying generative deep neural networks has lead to a stark performance increase in many language generation tasks (based on measuring the similarity with one or more human-references), concerns have been raised about the sensitive situations and the lack of safety-centered evaluation techniques [Dinan *et al.*, 2021]. In particular, applying V+L technology with visually impaired people could cause severe physical harms, e.g. when applying VQA to medication packaging [Davis *et al.*, 2020]. One suggested solution is for the model to predict when an image is of insufficient quality to generate a caption [Chiu *et al.*, 2020a]. Here, we follow a different human-centred approach where we generate model confidence scores and place the decision of whether to 'trust' the model with a human operator/ stakeholder who can set an appropriate context-dependent threshold, e.g. by conducting a risk-based analysis using Value Sensitive Design methods [Friedman *et al.*, 2017]. However, neural models are often overconfident and their confidence scores do not reliably reflect the likelihood that their prediction is correct [Guo *et al.*, 2017; Wang *et al.*, 2020], and thus not providing the human decision-maker with reliable information. We address this issue via calibration analysis.

3 Method

Our modelling framework consists of three key ideas, see Figure 2. First, we augment the data by duplicating and distorting existing VizWiz images with synthetic noise that reflects real-world quality issues of images taken by people with vision impairment (Section 3.1); Second, we propose a dual image captioning network which benefits from the fact that the augmented distorted images share the same captions with the original image, which enhances the model's robustness to various types of noise (Section 3.2); Third, we explore three types of quality-agnostic losses to enforce the consistency between the original image and the augmented image, respectively for latent space, logits and label consistency (Section 3.3).
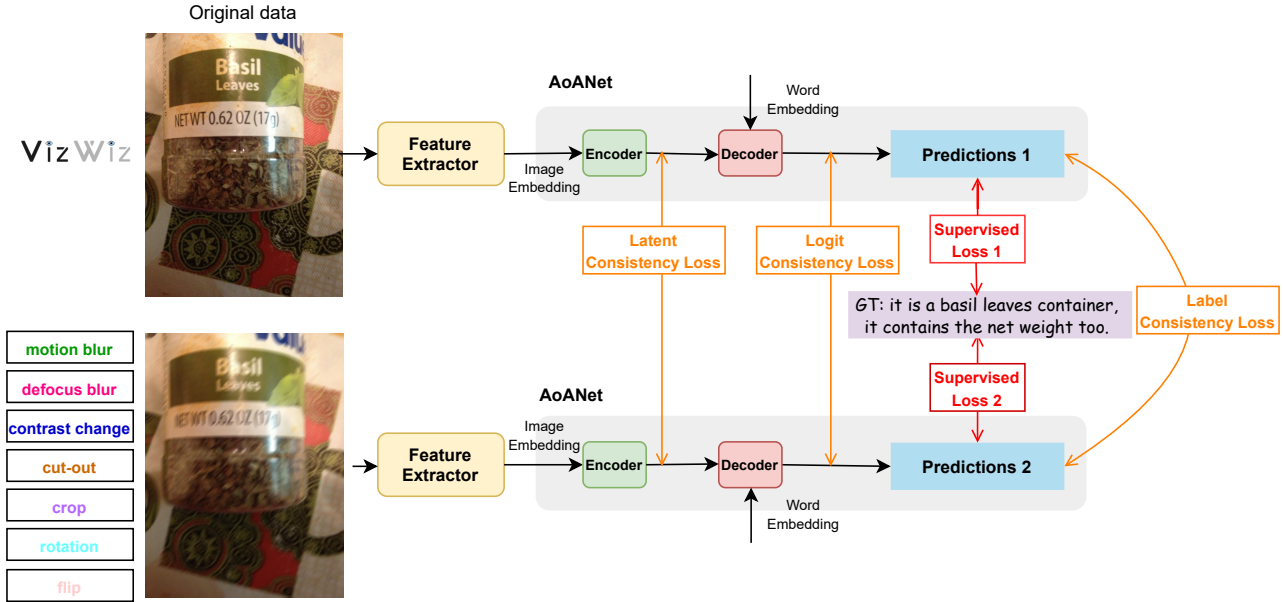
Figure 2: The proposed quality-agnostic image captioning framework for people with vision impairment.

Real-word issue	blur		bright	dark	framing	obscured	rotation	
Synthetic noise	motion blur	defocus blur	contrast change		crop	cut-out	rotation	flip

Table 1: Real-word issues in Vizwiz-Captions dataset and the corresponding synthetic noise type.

3.1 Augmentation Strategy

The VizWiz-Captions dataset contains six main types of quality flaws, as annotated via crowdsourcing [Chiu *et al.*, 2020b], where an image can have more than one type of flaw: 41.0% of *blur*, 5.3% of *overexposure (bright)*, 5.6% of *underexposure (dark)*, 55.6% of *improper framing*, 3.6% of *obstructions*, 17.5% of *rotated views* and 0.8% of *other reason*. To alleviate data sparsity, we augmented the existing data by adding synthetic noise coherent to the above noise types using the *imgaug* library,[1] which provides a range of image distortion functions. Table 1 provides an overview how we map the synthetic noise functions to the real-word issues described in [Chiu *et al.*, 2020b]. Each original image is augmented with a randomly selected noise type following the distribution of their appearance in the dataset.[2] We then assign the ground truth (GT) caption from the original image to the augmented synthetic data since they share the same content but with a different quality level.

3.2 Dual Network

Next, we extend AoANet [Huang *et al.*, 2019], which is the top-performing model on VizWiz Captions according to [Gurari *et al.*, 2020], to a dual network architecture in order to enhance its robustness to noise. Note that the suggested dual-network extension is model independent and the choice of

---

[1]https://github.com/aleju/imgaug *last accessed 21 Feb 2023.*

[2]We experimented with different noise distributions, which all lead to very similar performance.

baseline model is somewhat arbitrary. In particular, we extend the model with two identical branches as summarised in Figure 2: the original data I from VizWiz-Captions is fed into one branch, the augmented data $\widehat{I}$ representing one of the real-world noise types in Section 3.1 is fed into the other. For each image, we apply the feature extractor VinVL [Zhang *et al.*, 2021] to extract the visual embeddings, which are then combined with the word embeddings from the caption (i.e. the same caption assigned to the noisy augmented and original image). These multimodal embeddings are then used to train AoANet for generating two image captions – one from each branch. We train this dual network by optimizing two cross entropy (XE) losses as following:

$$L_{XE}^{orig}(I, \theta) = -\sum_{t=1}^{T} log(p_\theta(y_t^*|y_{1:t-1}^*, I)) \qquad (1)$$

$$L_{XE}^{aug}(\widehat{I}, \theta) = -\sum_{t=1}^{T} log(p_\theta(y_t^*|y_{1:t-1}^*, \widehat{I})) \qquad (2)$$

where

$$p(y_t|y_{1:t-1}, x) = softmax(F(x)) \qquad (3)$$

where $y_{1:T}^*$ denotes the target ground truth sequence and $F(x)$ is the logits of image x ($x \in \{I, \widehat{I}\}$).

3.3 Quality-agnostic Consistency

In order to increase model robustness, we aim for both branches of the model to make the same prediction – no mat-

ter the image quality. We thus adapt consistency regularization to our framework in order to increase the generalization and stability for the model on the synthetic noisy data. We explore pairwise similarity losses as previously used in unsupervised [Wu *et al.*, 2019; Isobe *et al.*, 2021] and semi-supervised learning [Abuduweili *et al.*, 2021; Hu *et al.*, 2021; Lai *et al.*, 2021; Gong *et al.*, 2021]. In contrast to these previous works we consider multi-modal inputs for our network. We explore three types of consistency losses as shown in Figure 2: one for image embeddings only (denoted as *latent consistency*) and two for cross-modal embeddings (denoted as *logit consistency* and *label consistency* respectively). The latent consistency (LAC) loss is maintained between the original refining image embeddings (output of encoder module) and the augmented image embeddings in latent space. The logit consistency (LOC) and label consistency (LBC) are applied between two branches before and after softmax modules respectively.

The latent consistency loss minimizes the distance between image embeddings in the latent space to enforce the image features aligned.

$$L_{LAC} = \left\| F_{latent}(I) - F_{latent}(\widehat{I}) \right\| \quad (4)$$

where $\|.\|$ refers to the Frobenius norm and $F_{latent}(x)$ (where $x \in \{I, \widehat{I}\}$) is the latent image embedding.

The logit consistency loss constrains the cross-modality output embeddings between these two branches by minimizing the following loss:

$$L_{LOC} = \left\| F(I) - F(\widehat{I}) \right\| \quad (5)$$

The label consistency loss minimizes the distance between the predictions of the original images and the augmented images:

$$L_{LBC} = \left\| p(y_t|y_{1:t-1}, I) - p(\widehat{y}_t|\widehat{y}_{1:t-1}, \widehat{I}) \right\| \quad (6)$$

where $y_{1:T}$ and $\widehat{y}_{1:T}$ denote the predictions from image I and $\widehat{I}$ respectively and refer Equation 3 for $p(y_t|y_{1:t-1}, x)$ (where $x \in \{I, \widehat{I}\}$).

3.4 Algorithm

Figure 2 summarises our framework so far, where we introduced both supervised (in red) and unsupervised (in orange) losses in the previous section: supervised losses are based on the fact that images with different quality level share the same captions; and unsupervised losses represent mutual information in latent space and final output space. Our final loss thus contains two terms: the supervised loss L_{XE} and the consistency loss L_{cons}:

$$L = L_{XE} + \lambda \times L_{cons} \quad (7)$$

where $cons \in \{LAC, LOC, LBC\}$ and the hyper-parameter λ represents the trade-off between the two terms..

During inference, we disregard the branch trained on the augmented data, and perform caption generation using only the branch trained on the original data.

4 Implementation Details

Model Training and Testing. We follow the same protocol as first established by [Gurari *et al.*, 2020] on VizWiz-Captions dataset. We evaluate on the test set using the EvalAI evaluation server[3]. We use VinVL [Han *et al.*, 2021] to extract the image features for both original images and the images with synthetic noise. All models are implemented with Pytorch [Paszke *et al.*, 2017]. We train our models with the initial learning rate $2e^{-4}$ and $2e^{-5}$ with 40 epochs for two phases of training from scratch. We use Euclidean distance to compute the consistency loss. Following [Gurari *et al.*, 2020], We evaluate all methods with eight similarity metrics that are frequently used for image caption generation: BLEU-1-4 [Papineni *et al.*, 2002], METEOR [Denkowski and Lavie, 2014], ROUGE-L [Lin, 2004], CIDEr-D [Vedantam *et al.*, 2015], and SPICE [Anderson *et al.*, 2016].

Synthetic Noise Generation. As described in Section 3.1, we augment the original data with synthetic noise using the imgaug python library. We use the following parameters: *Crop*: crop each side by up to 20 percent relative to its original size. *Rotate*: create an augmenter that rotates images by a random value between -45 and 45 degree. *Flip*: flip images vertically. *Motion blur*: Apply motion blur with a kernel size of randomly from 15x15 to 50x50 pixels to images, a blur angle of either -45 or 45. *Defocus Blur*: the image below visualizes severity 1 to 5. *Contrast*: Modify the contrast of images according to 255*((v/255)**γ), where v is a pixel value and γ is sampled uniformly from the interval [0.5, 2.0] (once per image). *Cutout*: fill per image one area, each having between 10% to 50% of the corresponding size of the height and width (for non-square images this results in non-square areas to be filled). The augmented dataset will be released with this paper.

5 Experimental Results

5.1 Comparison of State-of-the-Art Methods

Table 2 summarises the results on the VizWiz-Captions test set. Following [Gurari *et al.*, 2020], we consider the following baselines based on the original, single-branch AoANet model [Huang *et al.*, 2019]:

Pretrained: AoANet pretrained on the MS-COCO dataset;

AoANet: AoANet trained from scratch on VizWiz-Captions;

AoANet+finetuned: AoANet pretrained on MS-COCO and finetuned on VizWiz-Captions;

We also experiment with the following versions of our regularized, quality-agnostic dual-network model.

DualNet: Dual Network model without any consistency losses and trained from scratch on VizWiz-Captions.

DualNet+finetuned: Dual Network model without any consistency losses and pretrained on the MS-COCO and finetuned on VizWiz-Captions;

DualNet+cons: Dual Network model trained from scratch with label consistency loss (LBC);

DualNet+cons+finetuned: Dual Network model pretrained on the MS-COCO and finetuned with LBC.

---

[3]https://eval.ai/web/challenges/challenge-page/739/submission

	CIDEr	B@1	B@2	B@3	B@4	METEOR	ROUGE	SPICE
Pretrained	19.40	54.90	34.70	21.00	13.20	13.40	37.60	6.20
AoANet	60.47	66.52	47.98	33.74	23.42	20.10	46.81	15.58
AoANet+finetuned	58.08	66.12	46.82	32.40	22.30	19.50	46.28	14.97
DualNet	62.30	66.92	48.45	34.04	23.63	**20.43**	47.28	**16.03**
DualNet+finetuned	55.02	67.01	48.31	34.08	23.36	19.79	46.86	14.42
DualNet+cons	**62.62**	**67.08**	**48.60**	**34.21**	**23.75**	20.34	**47.32**	15.91
DualNet+cons+finetuned	52.43	65.58	46.7	32.55	22.06	19.14	45.84	14.26

Table 2: Performance on the VizWiz-Captions test set with respect to eight NLG metrics (B@ = BLEU-).

		CIDEr	B@1	B@2	B@3	B@4	METEOR	ROUGE	SPICE
	Easy	62.10	69.59	51.11	36.31	**25.53**	21.17	49.05	13.84
AoANet	Medium	53.39	62.91	43.27	28.91	19.43	18.80	43.81	**13.40**
	Hard	37.11	**34.49**	**19.37**	11.12	6.81	**11.38**	**28.04**	9.81
	Easy	62.92	**70.49**	**51.70**	**36.62**	25.52	21.15	**49.19**	14.00
DualNet	Medium	52.80	**63.45**	**43.52**	**29.18**	**19.67**	18.68	43.76	13.28
	Hard	**40.00**	34.38	19.29	**11.38**	**7.32**	11.15	27.98	**10.71**
	Easy	**63.09**	69.82	51.15	36.16	25.13	**21.18**	48.91	**14.05**
DualNet+cons	Medium	**53.98**	62.81	43.09	28.61	19.34	**18.94**	**43.83**	13.37
	Hard	37.65	34.47	19.00	10.84	6.62	11.21	27.76	9.77

Table 3: Comparison of the baseline algorithms and our proposed model on images with different 'difficulty' levels (easy / medium / hard) on the validation set with respect to eight metrics.

We find that AoANet pretrained on the MS-COCO dataset shows inferior performance with only 19.40% on CIDEr and 54.90% on B@1 which we attribute to domain shift. When training AoANet from scratch we observe an absolute improvement of 41.07 on CIDEr and 12.43 on B@1 over the pretrained baseline, whereas finetuning a pretrained model on VizWiz-Captions results in a small performance drop. These results are consistent with [Gurari *et al.*, 2020] and confirm our hypothesis that pretraining on mostly "clean" images is not effective for this domain. Our dual model with augmented noisy data (denoted as 'DualNet'), performs better than AoANet on all metrics. It achieves a 1.83 gain on CIDEr, which measures the similarity of a generated caption to the consensus/majority of GT captions. After applying the label consistency loss to our dual network (denoted as 'DualNet+cons'), we observe the best performance on the majority of metrics, including absolute improvement of 0.32 on CIDEr, 0.16 on B@1 over vanilla DualNet. Again, using pretraining and finetuning in this context leads to a slight drop in performance. These results are consistent with [Gurari *et al.*, 2020]. This highlights the importance of in-domain data for this task.

5.2 Robustness to Noise

Following [Gurari *et al.*, 2020], we evaluate method performance on images of different levels of difficulty/ noise, where we estimate difficulty by the number of crowdworkers who deemed the image of insufficient quality to generate a meaningful caption. In particular, images are considered 'easy' if all five human annotators were able to provide captions, 'medium' if captioned by 3-4 annotators, and 'hard' if only 1-

2 captions were collected. We apply the same categorisation on the validation set, since GT captions are not available for the test set.[4] This results in 4918 easy images, 9773 medium images and 1406 hard images. From the results in Table 3, we observe that the difficulty of the images is reflected by the performance metrics, with a gap of over 20 CIDEr points between easy and difficult images for all models. Compared to AoANet, the vanilla dual network without consistency loss obtains an improvement of 2.89 on CIDEr and 0.9 on SPICE for 'hard' images and up to 1 point gain on B@1-4 for 'easy' and 'medium' images. When adding the consistency loss, we observe improvements on CIDEr and METEOR for 'easy' and 'medium' images, but not for 'hard' images. We assume this is due to the fact 'hard' images has poor quality and fewer captions (only 1-2) which is more challenging to improve with the consistency loss.

5.3 Ablation Studies

We now evaluate the effectiveness of different losses used to train our dual network. All ablation studies are performed on the dev set. As explained in Section 3.3, we propose a total of five losses in our framework: two softmax-based supervised losses (original-data loss L_{XE}^{orig} and augmented-data loss L_{XE}^{aug}) and three types of consistency-based unsupervised losses (latent consistency L_{LAC}, logit consistency L_{LOC} and label consistency L_{LBC}). We show the effect of combinations of these losses in Table 4. The baseline represents the supervised loss(es) of single-branch network and the extended dual network, i.e. L_{XE}^{orig} and $L_{XE}^{orig} + L_{XE}^{aug}$.

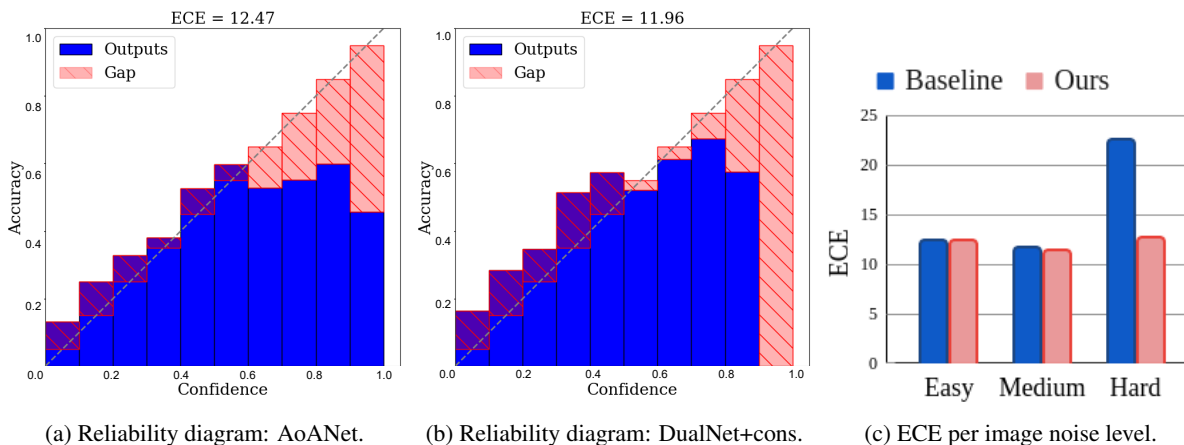[4]We will release the indexed validation set with the final version.

(a) Reliability diagram: AoANet.   (b) Reliability diagram: DualNet+cons.   (c) ECE per image noise level.

Figure 3: Results from the calibration analysis on the validation set.

Losses	CIDEr	B@1	B@2	B@3	B@4	METEOR	ROUGE	SPICE
L_{XE}^{orig}	60.50	66.40	47.90	33.40	23.20	20.30	47.10	14.00
$L_{XE}^{orig} + L_{XE}^{aug}$	62.20	66.83	48.35	33.92	23.49	20.42	47.37	**16.11**
$L_{XE}^{orig} + L_{XE}^{aug} + L_{LAC}$	53.26	66.41	46.30	31.16	20.71	19.65	45.94	14.89
$L_{XE}^{orig} + L_{XE}^{aug} + L_{LOC}$	60.45	**67.09**	47.96	33.31	22.99	20.00	46.80	15.37
$L_{XE}^{orig} + L_{XE}^{aug} + L_{LBC}$	**63.00**	**67.09**	**48.53**	**34.22**	**23.75**	**20.48**	**47.46**	16.06

Table 4: Performance of our algorithm with different combinations of losses on the VizWiz-Captions dev set, which includes two softmax-based supervised losses (original-data loss L_{XE}^{orig} and augmented-data loss L_{XE}^{aug}) and three types of consistency-based unsupervised losses (latent consistency L_{LAC}, logit consistency L_{LOC} and label consistency L_{LBC}).

Adding latent consistency loss L_{LAC} lowers performance compared to the baseline and logit consistency losses L_{LOC} obtains similar results as baseline. We interpret this result as follows: L_{LAC} is positioned right after the encoder (cf. Fig 2) and thus only affects consistency between image embeddings but doesn't constraint the text prediction. L_{LOC}, in contrast, is based after the decoder and thus also considers multimodal embeddings, but is placed before softmax. L_{LBC} places the consistency constraints after softmax which obtained gains on all evaluation metrics except for SPICE. In general, the results show that adding label consistency loss L_{LBC} achieves the highest performance gain across all evaluation metrics except for SPICE. Note that SPICE is based on the scene graph rather than the surface form and does not take text fluency into account. While the consistency losses constrain the text embedding change to some extent.

6 Calibration Analysis

The previous sections have focused on evaluating performance (Section 5.1) and robustness to noise levels (Section 5.2), we now evaluate the reliability. In particular, we evaluate whether our model is well calibrated, meaning whether the word-based confidence scores reflect the likelihood of the generated caption being correct. Accurate confidence scores can empower users with a measure of how much the system's output is to be trusted. Previous studies have found

that deep neural networks tend to be miscalibrated skewing towards over-confidence on average [Guo *et al.*, 2017]. In order to evaluate the calibration of our proposed model, we use the Expected Calibration Error (ECE) [Naeini *et al.*, 2015], which is computed by binning the total of n predictions into M equally sized confidence bins and averaging the difference between model's accuracy and confidence within each bin B_m (the lower the better):

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \qquad (8)$$

Following previous work on the calibration of neural machine translation systems [Wang *et al.*, 2020], we use Translation Error Rate (TER) in order to compute word-level accuracy between the generated caption and the GT. Given that we have multiple references, we keep the accuracy scores that lead to lowest TER.

Figures 3a-3b present the reliability diagrams for the baseline AoANet and the proposed model. Each bar shows the average accuracy score for the corresponding confidence bin. Our DualNet+cons model has 0.51 lower overall ECE which indicates better-calibrated outputs. Moreover, we notice that for confidence scores higher than 0.6, AoANet tends to output more over-confident predictions. These results are in line with previous research, e.g. [Thulasidasan *et al.*, 2019], showing that models that use data augmentation methods tend to
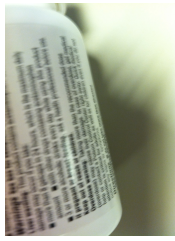
Example 1 (Easy)

GT: The back side of a white pill bottle and black font.

AoANet: The back of a white bottle with black text (C = 0.47).

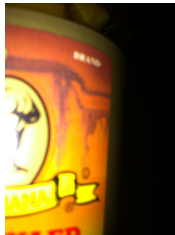**Ours:** The back of a bottle of **medicine** with a white **label** (C = 0.35).

Example 2 (Medium)

GT: A crumpled and folded up one dollar bill.

AoANet: A **twenty** dollar bill laying on a red carpet (C = 0.40).

Ours: A **one** dollar bill laying on a red and **white** rug (C = 0.35).

Example 3 (Hard)

GT: Corner of a label from a round container that is orange, yellow and red.

AoANet: A yellow and orange **container** of some sort of food product (C = 0.46).

Ours: A **bottle** of a red and yellow label (C = 0.25).

Example 4 (Easy)

GT: A bottle of Fairy brand original dishwashing liquid.

AoANet: A green bottle of **green tea** with a green label (C = 0.48).

Ours: A green can of **mountain dew soda** on a table (C = 0.31).

Figure 4: Examples of generated captions by AoANet, our model and the corresponding ground truths. C: Aggregated confidence score for the generated caption. Mistakes made in the captions are highlighted in red and correspondingly correct ones in blue.

be better calibrated. Figure 3c compares the calibration error of our proposed model to the baseline for each image 'difficulty' level (cf. Section 3). We observe that both models have similar ECE on easy and medium images. However, Dual-Net+cons has 44% lower ECE on hard images. This suggests that our confidence scores are more reliable for cases when the model is likely to fail, which is the most important scenario to prevent safety issues.

7 Qualitative Examples

In order to further illustrate our results Figure 4 shows examples from the AoANet baseline and the proposed Dual-Net+cons model. Examples 1-3 represent images with increasing difficulty levels. In general, we observe that our model is able to provide more descriptive and accurate captions for images labelled as 'easy' or 'medium', but underperforms for images labelled as 'hard'. Both systems correctly describe Example 1 labelled as 'easy', however our model provides more details. Both systems make mistakes when describing Example 2 labelled as 'medium': our model wrongly assigns the colour 'white' to the rug in the background, whereas the baseline mislabels the dollar bill as 'twenty' with high confidence. Arguably, the latter mistake could lead to worse consequences.

In Example 3, labelled as 'hard', our model wrongly describes an object labeled as 'container' as a 'bottle', however with low confidence. This is also a hard task for humans to solve: Arguably, it is not clear from the image whether the 'container' might also be of type 'bottle'. Note that human annotators also tend to disagree more on harder examples [Bhattacharya *et al.*, 2019], and thus it is harder for the model to agree with a single GT.

Example 4 shows another 'easy' image, where both models describe a bottle of dishwasher liquid as a drink ('green tea' vs. 'mountain dew soda'). The main advantage of our model is here that it assigns lower confidence to its prediction, whereas the baseline model produces errors with higher confidence. As argued earlier, this can lead to safety critical situations for blind or otherwise vulnerable users. Possible solutions include setting a threshold to determine when not to issue an image caption, or to explicitly indicate uncertainty in the prediction, e.g. by generating hedge phrases [Gkatzia *et al.*, 2016].

8 Conclusion and Limitations

We present a quality agnostic framework for generating text captions for images taken by people with visual impairments. This is a challenging task due to the low quality and quantity of data, which we address by synthetic data generation and consistency regulation. Our results show consistent and considerable improvements over state-of-the-art baseline systems. We also show that our model produces more reliable confidence scores, especially for hard cases where the model is likely to make an error. "Knowing when you don't know" is especially important in the context of assistive technology for vulnerable users, since wrong but confident predictions can cause severe harm. In these cases, the model prediction should either be discharged and replaced by a human operator, or the uncertainty in the model predictions needs to be communicated to the human decision maker. The advances presented in this paper now enable us to experimentally explore these two scenarios as part of an assisted living application, developed in partnership with the Royal National Institute of Blind People.

Ethical Statement

VizWiz-Captions is a publicly-available image captioning dataset [Gurari *et al.*, 2020]. It builds on top of the original VizWiz data [Bigham *et al.*, 2010], which provides images taken by people with visual impairments. The original images were filtered with respect to privacy concerns, e.g. images showing people's faces were removed. The images were collected by 11 blind iPhone users aged 22 to 55.

The image captions were then collected by [Gurari *et al.*, 2020]. They assigned five workers to each image using the crowdsourcing platform Amazon Mechanical Turk (AMT). The instructions specify to include at least eight words as

well as what not to do when creating the caption (e.g., do not speculate what people in the image might be saying/thinking or what may have happened in the future/past). We were not able to find information regarding the demographics of the crowdworkers or consent forms, as this is not common practise on AMT.

The overall goal of this research is to develop an assisted living application in partnership with Royal National Institute of Blind People. While this study marks an important step towards making image captioning technology more robust and safe for real-world applications, there are additional ethical challenges which need to be solved, such as the privacy of the user when uploading images. We hope to address this in future work.

Acknowledgments

References

[Abuduweili *et al.*, 2021] Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6923–6932, 2021.

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[Bennett *et al.*, 2018] Cynthia L. Bennett, Jane E, Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. *How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2018.

[Bhattacharya *et al.*, 2019] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4271–4280, 2019.

[Bigham *et al.*, 2010] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[Chen *et al.*, 2022] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the*

[Chiu *et al.*, 2020a] T. Chiu, Y. Zhao, and D. Gurari. Assessing image quality issues for real-world problems. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3643–3653, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[Chiu *et al.*, 2020b] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3646–3656, 2020.

[Davis *et al.*, 2020] Nathan Davis, Bo Xie, and Danna Gurari. Quality of images showing medication packaging from individuals with vision impairments: Implications for the design of visual question answering applications. *Proceedings of the Association for Information Science and Technology*, 57, 10 2020.

[Denkowski and Lavie, 2014] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[Dinan *et al.*, 2021] Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*, 2021.

[Friedman *et al.*, 2017] Batya Friedman, David G Hendry, and Alan Borning. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(2):63–125, 2017.

[Gkatzia *et al.*, 2016] Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. Natural language generation enhances human decision-making with uncertain information. In *54th Annual Meeting of the Association for Computational Linguistics 2016*, pages 264–268. Association for Computational Linguistics, 2016.

[Gong *et al.*, 2021] Chengyue Gong, Dilin Wang, and Qiang Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13683–13692, 2021.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[Gurari *et al.*, 2020] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer, 2020.

[Han *et al.*, 2021] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Image scene graph generation (sgg) benchmark. *arXiv preprint arXiv:2107.12604*, 2021.

[He *et al.*, 2019] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[Hu *et al.*, 2021] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021.

[Huang *et al.*, 2019] 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.

[Isobe *et al.*, 2021] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8187–8196, 2021.

[Lai *et al.*, 2021] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021.

[Lin *et al.*, 2014] 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*, pages 740–755. Springer, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[MacLeod *et al.*, 2017] 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*, CHI '17, page 5988–5999, New York, NY, USA, 2017. Association for Computing Machinery.

[Naeini *et al.*, 2015] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Pantazopoulos *et al.*, 2021] George Pantazopoulos, Jeremy Bruyere, Malvina Nikandrou, Thibaud Boissier, Supun Hemanthage, Binha Kumar Sachish, Vidyul Shah, Christian Dondrup, and Oliver Lemon. Vica: Combining visual, social, and task-oriented conversational ai in a healthcare setting. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 71–79, New York, NY, USA, 2021. Association for Computing Machinery.

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

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Thirty-first Conference on Neural Information Processing Systems Workshop*, 2017.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020.

[Thulasidasan *et al.*, 2019] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[Tseng *et al.*, 2022] Yu-Yun Tseng, Alexander Bell, and Danna Gurari. Vizwiz-fewshot: Locating objects in images taken by people with visual impairments. In *European Conference on Computer Vision*, pages 575–591. Springer, 2022.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[Wang *et al.*, 2020] Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*, 2020.

[Wu *et al.*, 2019] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6922–6931, 2019.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[Zhang *et al.*, 2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[Zoph *et al.*, 2020] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In *Advances in Neural Information Processing Systems*, volume 33, pages 3833–3845, 2020.