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

Lu Yu\textsuperscript{1,}\.CollectionsAuthorMalvina Nikandrou\textsuperscript{2}, Jiali Jin\textsuperscript{1}, Verena Rieser\textsuperscript{2,3}\CollectionsAuthor†
\textsuperscript{1}Tianjin University of Technology, Tianjin, China
\textsuperscript{2}Heriot-Watt University, Edinburgh, United Kingdom
\textsuperscript{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 \textit{Good health and well-being} and \textit{Reduce inequalities}, as well as the UN Sustainable Development Principle \textit{Leave no one behind}.

---
\textsuperscript{*} Work done at Heriot-Watt University.
\textsuperscript{†} 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-url)

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 visual 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, including COCO and ImageNet, has limited benefit for models fine-tuned 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 pretrained-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 led 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).
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 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 $\hat{I}$ 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:

$$L^{orig}_{XE}(I, \theta) = -\sum_{t=1}^{T} \log(p_0(y^*_t | y_{1:t-1}^*, I))$$ (1)

$$L^{aug}_{XE}(\hat{I}, \theta) = -\sum_{t=1}^{T} \log(p_0(y^*_t | y_{1:t-1}^{*\hat{I}}, \hat{I}))$$ (2)

where

$$p_0(y_t | y_{1:t-1}, x) = \text{softmax}(F(x))$$ (3)

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

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-

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

$$L_{LAC} = \| F_{latent}(I) - F_{latent}(\hat{I}) \|$$ (4)

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

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

$$L_{LOC} = \| F(I) - F(\hat{I}) \|$$ (5)

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

$$L_{LBC} = \| p(y_t|y_{t-1}, I) - p(y_t|y_{t-1}, \hat{I}) \|$$ (6)

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

### 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 $L_{cons}$:

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

where $cons \in \{ LAC, LOC, LBC \}$ and the hyper-parameter $\lambda$ 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. 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 $255^\gamma(v/255)^\gamma$, where $v$ is a pixel value and $\gamma$ 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;
- **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 ($L_{BC}$);
- **DualNet+cons+finetuned**: Dual Network model pretrained on the MS-COCO and finetuned with $L_{BC}$.

3https://eval.ai/web/challenges/challenge-page/739/submission
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. 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 $L_{XE}^{\text{orig}}$ and augmented-data loss $L_{XE}^{\text{aug}}$) 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. $L_{XE}^{\text{aug}}$ and $L_{XE}^{\text{orig}} + L_{XE}^{\text{aug}}$.

We will release the indexed validation set with the final version.
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):

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

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
model provides more details. Both systems make mistakes directly describe Example 1 labelled as ‘easy’, however our model is able to provide more descriptive and accurate captions 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 dishwashing 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

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.
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 practice 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

Lu Yu and Verena Rieser were supported by the EPSRC project ‘AISEC: AI Secure and Explainable by Construction’ (EP/T026952/1), Lu Yu was also supported by NSFC project ‘AISEC: AI Secure and Explainable by Construction’ (NO.62202331), Verena Rieser was also supported by ‘Gender Bias in Conversational AI’ (EP/T023767/1) and ‘Equally Safe Online’ (EP/W025493/1), as well as by a Leverhulme Trust Senior Research Fellowship (SRF/R1/201100).

References


[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.


