Visual Cues and Error Correction for Translation Robustness

Zhenhao Li, Marek Rei, Lucia Specia
Language and Multimodal AI (LAMA) Lab, Imperial College London
{zhenhao.li18, marek.rei, l.specia}@imperial.ac.uk

Abstract

Neural Machine Translation models are sensitive to noise in the input texts, such as misspelled words and ungrammatical constructions. Existing robustness techniques generally fail when faced with unseen types of noise and their performance degrades on clean texts. In this paper, we focus on three types of realistic noise that are commonly generated by humans and introduce the idea of *visual context* to improve translation robustness for noisy texts. In addition, we describe a novel *error correction training* regime that can be used as an auxiliary task to further improve translation robustness. Experiments on English-French and English-German translation show that both multimodal and error correction components improve model robustness to noisy texts, while still retaining translation quality on clean texts.

1 Introduction

Neural Machine Translation (NMT) has been shown to be very sensitive to noise (Belinkov and Bisk, 2018; Michel and Neubig, 2018; Ebrahimi et al., 2018), with even small perturbations in the inputs often leading to mistranslations. To improve the robustness of NMT models, current research mostly focuses on adapting the model to noisy texts via methods such as fine-tuning (Michel and Neubig, 2018; Alam and Anastasopoulos, 2020), noise-injection (Belinkov and Bisk, 2018; Cheng et al., 2018; Karpukhin et al., 2019), and data augmentation through back-translation (Berard et al., 2019; Vaibhav et al., 2019; Li and Specia, 2019), etc. In these approaches, the translation model is trained or fine-tuned on the noisy data so that it can learn from the noise. However, methods using extra context to help translate noisy texts have not been investigated.

Studies in Multimodal Machine Translation (MMT) have shown that visual information improves translation quality when the textual context is incomplete (Caglayan et al., 2019; Imankulova et al., 2020; Caglayan et al., 2020). However, as exemplified by Caglayan et al. (2019) (Figure 1), an MMT model trained on clean data was not able to handle noise. When the word "son" was misspelled as "song", the model disregarded the visual information and used the literal translation "chanson". The MMT model attended to the relevant region in the image and generated the intended translation "enfant" only when the noise was masked by a placeholder in the input, imitating an out-of-vocabulary (OOV) example.

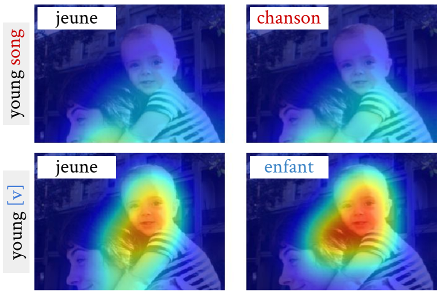


Figure 1: As showed by Caglayan et al. (2019), multimodality can help translate unknown words, but fail when there is noise in the input. The misspelled word "song" is correctly translated as "enfant" (child) when it is replaced with an unknown token, but translated literally as "chanson" (song) otherwise.

Given that the visual modality has been shown to help predict unknown words, we investigate whether adding multimodal information to adaption-based methods would further improve translation robustness. To answer this question, we build MMT models in conjunction with noise injection techniques and investigate their behaviour during training and inference on both noisy and clean data. To further improve robustness, we extend the current *adversarial training* method (i.e., training NMT models on noisy texts) and propose an *error correction training* method. In addition to training the model with noise-injected source

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3153–3168
November 7–11, 2021. ©2021 Association for Computational Linguistics

sentences and their clean translation counterparts, we introduce error correction as an auxiliary task and add a separate decoder to the model, which is used to denoise the source sentence.[1] Our **main contributions** can be summarized as:

- To the best of our knowledge, this is the first work combining adversarial training with *multimodal* NMT to improve translation robustness. We evaluate robustness on three types of noise that mimic errors commonly introduced by humans. Systematic experiments reveal that multimodality can improve model performance on both known and unseen noise.
- We propose an *error correction training* method for translation by introducing denoising as an auxiliary task. We show that the robustness of both NMT and MMT models is improved with this method.
- We demonstrate that the model using visual features also learns to correct grammatical errors more accurately, indicating the potential for multimodal monolingual error correction.

The paper is organised as follows: In Section 2, we present the background and related work. In Section 3, we introduce the types of noise injected and the error correction training method. In Section 4, we describe our experiment settings, with experiment results in Section 5, and further analysis in Section 6.

2 Background and Related Work

Robust NMT Although NMT models can achieve high performance on clean data, they are very brittle to non-standard inputs, such as noisy texts (Belinkov and Bisk, 2018). Different types of noisy data have been proposed to test translation robustness, e.g. synthetic word perturbations (Belinkov and Bisk, 2018), grammatical errors (Anastasopoulos et al., 2019), and user-generated texts from social platform (Michel and Neubig, 2018; Li et al., 2019; Specia et al., 2020).

The most common approach to improve translation robustness is to train the model on noisy data, which is referred to as adversarial training. Since parallel data with noisy source sentences and clean translations is difficult to obtain, the clean training data is often injected with different types of artificial noise, e.g. random word perturbations like character insertion/deletion/substitution (Belinkov and Bisk, 2018; Karpukhin et al., 2019; Passban et al., 2020; Xu et al., 2021), noise generated via back-translation (Berard et al., 2019; Vaibhav et al., 2019; Li and Specia, 2019), and adversarial examples generated by white-box generator model (Cheng et al., 2018, 2019, 2020). Even though this method has been shown to improve NMT performance on noisy data, the types of noise used thus far are not common in real data. For example, it would be highly unlikely for human authors to misspell the word “robust” as “zobust”, but such random transformations are used when synthesizing noisy training data for MT. In addition, back-translation paraphrases the texts to introduce noise, however such noise is less realistic as human-generated errors, which include mispellings and grammatical errors. In adversarial approaches for other NLP tasks, Ribeiro et al. (2020) and Ma (2019) introduce various methods to inject both artificial and realistic noise. Inspired by these work, we focus on three types of noise that are commonly generated by humans in real texts and experiment with these for the translation task.

MMT Multimodal machine translation extends the framework of NMT by incorporating extra modalities, e.g. image (Specia et al., 2016a) or audio (Sulubacak et al., 2020). In our case, the extra modality is given as visual features from an image network to complement the textual context. In standard MMT, these features can be fused with the textual representation by simple operations such as concatenation (Caglayan et al., 2016), hidden states initialization (Calixto and Liu, 2017), or via attention mechanisms (Libovický and Helcl, 2017; Calixto et al., 2016, 2017; Yao and Wan, 2020) and latent variables (Calixto et al., 2019).

Recent research has shown that the extra modality helps translation, especially when the input is incomplete (Caglayan et al., 2019, 2020; Imankulova et al., 2020) or ambiguous (Ive et al., 2019; Wu et al., 2019b). Wu et al. (2019a) hinted at the possibility of multimodality helping NMT in dealing with natural noise stemming from the speech recognition system used as a first step in their pipeline approach to speech translations from videos. Their results, however, were inconclusive.

Salesky et al. (2021) investigate the robustness of open-vocabulary translation by representing texts

[1] Codes are available at `https://github.com/Nickeilf/Visual-Cues-Error-Correction`

clean	a pink flower is starting to bloom .
edit-distance	a pink flower is staring to loom .
homophone	a pink flour is starting to bloom .
keyboard	a pink flower is starring to bloom .

Table 1: An example of noise injected to the clean text. The noisy substitutes are marked in red.

as images followed by optical character recognition to cover some cases of noise such as misspellings. This is an interesting but orthogonal area of research since no external visual information is used.

Therefore, it remains an open question whether MMT can perform better than NMT on noisy texts, and whether multimodality can be complementary rather than redundant to previous text-based robustness techniques. The work by Caglayan et al. (2019) is the closest to our approach, however they focused mainly on identifying *when* the visual information is helpful. As such, they only performed experiments comparing NMT and MMT in the presence of unknown words consisting of placeholders used to mask out words in the source sentence. In contrast, we focus on multimodal models for realistic noise that includes in- and out-of-vocabulary words, such as misspellings or correctly-spelled words used in an incorrect context.

3 Methods

In this section, we introduce our methods to improve and evaluate the robustness of NMT and MMT models. In Section 3.1, we describe three techniques to inject realistic noise into training and test data. In Section 3.2, we introduce our error correction training method.

3.1 Noise Injection

In previous work on noise injection, the perturbations are often arbitrary, which would result in unrealistic noise. To simulate the natural noise in real situations, we add constraints to the random perturbations. We select three constrained noise injection methods that can be applied to both training and test data, with each method simulating one type of human-generated errors:

Edit distance A word is randomly replaced with another word in the vocabulary where the edit distance between the two words is less than two characters. The edit-distance noise simulates the occurrence of confusable spellings (e.g. sat vs seat) and also some grammatical errors (e.g. horse vs horses).

Homophones A word is randomly replaced with another word that shares the same pronunciation. We use the CMU Pronouncing Dictionary[2] to transform words into phonemes and find noisy substitutes with the same pronunciation. This simulates errors made by applications such as automatic speech recognition, or by non-native speakers.

Keyboard (Belinkov and Bisk, 2018) A character in a word is randomly replaced with an adjacent key on the standard QWERTY keyboard. The keyboard noise simulates the real-life typos when users accidentally press wrong keys while typing.

Table 1 shows examples of the three types of noise we experimented with. The edit distance and homophone noise types are applied on the word level, while the keyboard noise is on the character level. Word-level noise is more likely to break the sentence context even though the noisy substitutes are correctly spelled words. On the contrary, character-level noise is likely to introduce misspelled words and increase the out-of-vocabulary (OOV) rate.

When constructing the noisy training or test sets, we sample from the three types of noise following a uniform distribution, where to each sentence we apply only one type of noise. To avoid substituting words not carrying much contextual information (e.g. articles and punctuations) , we only perturb words with more than two characters. The noise level is controlled by the hyperparameter n, which defines the maximum number of words replaced with noisy counterparts per sentence. The noise injection procedure can be characterized as: given a source sentence $\mathbf{x} = [x_1, x_2, ..., x_M]$ and a target translation $\mathbf{y} = [y_1, y_2, ..., y_N]$, noise will be injected to the clean source sentence $\mathbf{x}$ to obtain its noisy variant $\mathbf{x}' = [x_1, ..., x'_{a_i}, ..., x_M]$, where a_i is the position of the noisy substitutes ($i = \{1, 2, ..., n\}$).

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict?in=C+M+U+Dictionary

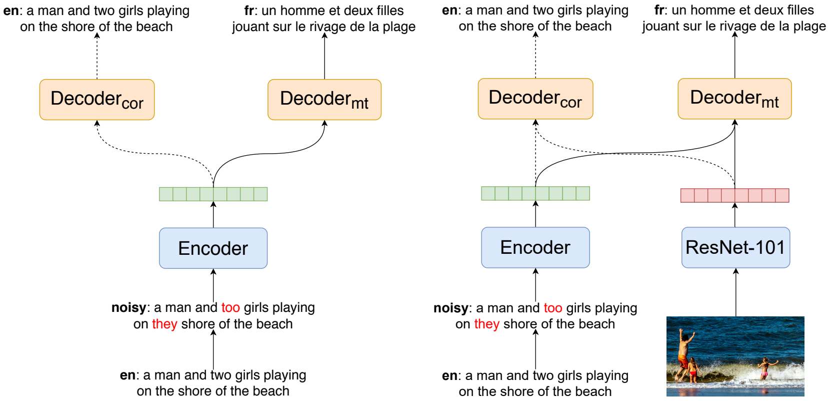

en: a man and two girls playing on the shore of the beach
fr: un homme et deux filles jouant sur le rivage de la plage
Decoder_cor
Decoder_mt
Encoder
noisy: a man and too girls playing on they shore of the beach
en: a man and two girls playing on the shore of the beach
en: a man and two girls playing on the shore of the beach
fr: un homme et deux filles jouant sur le rivage de la plage
Decoder_cor
Decoder_mt
Encoder
ResNet-101
noisy: a man and too girls playing on they shore of the beach
en: a man and two girls playing on the shore of the beach

Figure 2: Illustration of the joint training of machine translation and error correction for NMT and MMT models. Solid lines: translation flow. Dotted lines: error correction flow. Left: NMT with error correction training. Right: MMT with error correction training.

3.2 Error Correction Training

We introduce error correction (Ng et al., 2014; Yuan and Briscoe, 2016) as an auxiliary task to help improve the robustness against noisy inputs. For that, we add a second decoder to the MT architecture, which is only used for the error correction task. During training, the noisy sentence $\mathbf{x}'$ is encoded by the encoder, which is shared between the translation and correction tasks, into hidden states $\mathbf{h}'$. The hidden state representation is then fed to both decoders. The translation decoder aims to generate a correct translation $\mathbf{y}$ while the correction decoder aims to recover the original source sentence $\mathbf{x}$. This method is also compatible with the MMT model, where the error correction decoder will use both visual and textual hidden states to recover the clean source sentences. Figure 2 gives an illustration of the model architecture.

Compared to the standard MT model, the version with error correction training (which we refer to as *NMT-cor* and *MMT-cor* hereinafter) maximizes both the probability of generating correct translations $P(\mathbf{y}|\mathbf{x}';\boldsymbol{\theta_{mt}})$ and the probability of recovering the clean source sentences $P(\mathbf{x}|\mathbf{x}';\boldsymbol{\theta_{cor}})$.

$$P(\mathbf{y}|\mathbf{x}';\boldsymbol{\theta_{mt}}) = \prod_{t=1}^{N} P(y_t|\mathbf{y}_{1:t-1},\mathbf{x}';\boldsymbol{\theta_{mt}})$$
$$P(\mathbf{x}|\mathbf{x}';\boldsymbol{\theta_{cor}}) = \prod_{t=1}^{M} P(x_t|\mathbf{x}_{1:t-1},\mathbf{x}';\boldsymbol{\theta_{cor}}) \tag{1}$$

The $\boldsymbol{\theta_{mt}}$ represents parameters for the translation component and the $\boldsymbol{\theta_{cor}}$ represents parameters for the error correction component, with $\boldsymbol{\theta_{mt}} = \{\boldsymbol{\theta_{enc}}, \boldsymbol{\theta_{mt\_dec}}\}, \boldsymbol{\theta_{cor}} = \{\boldsymbol{\theta_{enc}}, \boldsymbol{\theta_{cor\_dec}}\}$. Our hypothesis is that the auxiliary task of error correction may help the encoder with a noise-invariant representation, which would indirectly improve the translation of noisy sentences. During training, we jointly optimize the sum of the translation loss and the error correction loss, as is shown in Equation 2:

$$\mathcal{L}_{mt}(\boldsymbol{\theta}_{mt}) = \frac{1}{|\mathbf{D}|}\sum_{(\mathbf{x}',\mathbf{y})\in\mathbf{D}} -\log P(\mathbf{y}|\mathbf{x}';\boldsymbol{\theta}_{mt})$$
$$\mathcal{L}_{cor}(\boldsymbol{\theta}_{cor}) = \frac{1}{|\mathbf{D}|}\sum_{(\mathbf{x}',\mathbf{x})\in\mathbf{D}} -\log P(\mathbf{x}|\mathbf{x}';\boldsymbol{\theta}_{cor})$$
$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{mt}(\boldsymbol{\theta}_{mt}) + \boldsymbol{\lambda}\mathcal{L}_{cor}(\boldsymbol{\theta}_{cor}) \tag{2}$$

where $\boldsymbol{\lambda} \geq 0$ is the factor that controls the weight of the error correction loss, and $\mathbf{D}$ represents the noise-injected data consisting of triples in the form of $(\mathbf{x}, \mathbf{x}', \mathbf{y})$.

4 Experiments

4.1 Datasets

We experiment with the Multi30K dataset (Elliott et al., 2016), using both the En-Fr and En-De language pairs. This is the standard dataset for MMT and has been used in all open challenges on the topic (Specia et al., 2016b; Elliott et al., 2017a; Barrault et al., 2018). Following Caglayan et al. (2019), we use both the *train* and *valid* splits as our training set. The *test2016-flickr* set is used as our development set for checkpoint selection. For evaluation, we test the models on both *test2017-flickr* and *test2017-mscoco* sets (Elliott et al., 2017b). We use a word-level vocabulary and build vocabularies for the original source and target languages, as well as the vocabulary on noisy source texts.[3] We use the pre-processed data in Multi30K, which is lowercased, normalized, and tokenized with Moses (Koehn et al., 2007). We also performed experiments using a subword-level vocabulary (BPE), which led to further improvements, but the trend in the results is the same (see Appendix A).

Following Caglayan et al. (2020), we use the "bottom-up-top-down" (BUTD) features (Anderson et al., 2018) extracted from a pre-trained Faster R-CNN ResNet-101 object detector. Each image is represented as 36 pooled feature vectors $V \in \mathbb{R}^{36\times 2048}$, with each vector representing a local object region.

4.2 Models

NMT and MMT Models Our baseline NMT model is the standard Transformer model (Vaswani et al., 2017), with 6 layers for both the encoder and the decoder. The hidden state size is 512 while the feed-forward dimension is 1024. The number of attention heads is set to 4. Dropout (0.3) is applied to both self/cross-attention and the position-wise feed-forward layer, and Pre-norm (Nguyen and Salazar, 2019) is applied to boost convergence. Our baseline MMT model follows the same architecture and hyperparameters as the baseline NMT model, except for the multimodal components. We use the serial multimodal cross-attention (Libovický et al., 2018), where an extra cross-attention sublayer is appended in the decoder layer to perform attention over the visual features. We also experiment with GRU models (Cho et al., 2014), following the hyperparameter settings of Caglayan et al. (2019). Due to space restrictions, we include the detailed results with GRU models in Appendix C. The GRU results display the same trend as the experimental results using Transformer models.

Error Correction Models The error correction NMT/MMT models adopt the same encoder and decoder as the baseline NMT/MMT models, except for a second decoder added for error correction training. During training, we compute the cross-entropy loss for translation, as well as for error correction in the correction-based models. In these models, the two losses are summed and optimized jointly on the same batch. We found the best λ value ($\lambda \in \{0.2, 0.2, 0.4, 0.4, 0.8\}$) for different levels of noise (number of noisy words $n \in \{1, 2, 4, 6, 10\}$) during hyperparameter tuning. See Appendix B for more details.

Training and Evaluation We use ADAM (Kingma and Ba, 2015) as the optimizer and adopt the *noam* learning rate scheduler (Vaswani et al., 2017) with a warm-up of 8000 steps. The training batch size is 64. Models are evaluated using the METEOR score (Denkowski and Lavie, 2014), which is the main metric for multimodal machine translation (Barrault et al., 2018). For the evaluation of error correction, we use ERRANT (Bryant et al., 2017) to compute the $F_{0.5}$ score. During evaluation, we select the checkpoint with the best performance on the development set and generate the translation and correction using beam search of size 12. All models are implemented using *nmtpytorch*[4] and *pysimt*[5]. Each model is run with three random seeds and the average results are reported. Each run takes approximately 2 hours to train on an RTX 2080 Ti GPU.

5 Results

5.1 Testing for Robustness to Noise

We first evaluate the robustness of standard NMT and MMT models **trained on clean data** by **testing on the noise-injected data**. This setting represents regular models that are not specifically adapted to noise. Figure 3 presents the change in METEOR (ΔMETEOR) between standard MMT and NMT models tested on data with different noise levels. The ΔMETEOR is consistently above 0 for both test sets in the two language pairs. As the noise level increases, the difference between NMT and MMT models is larger, showing that the visual information in the MMT model leads to predictions that are more robust to noise.

5.2 Training for Robustness to Noise

To test models for their ability to adapt to noisy data, we **train models on data with added noise**, sampling from the three types of noise in Section 3.1 and **test them on noisy test data**, with noise added in the same fashion. METEOR score results are shown in Table 2.

The training on noisy data is equivalent to the "adversarial training" experiments in previous studies (Belinkov and Bisk, 2018; Karpukhin et al.,

[3] Therefore there is no OOV word in the noisy training data, but the test data might still contain OOV words – noisy or not – with respect to the training.

[4] https://github.com/lium-lst/nmtpytorch

[5] https://github.com/ImperialNLP/pysimt

	flickr2017						**mscoco2017**					
	clean	n=1	n=2	n=4	n=6	n=10	clean	n=1	n=2	n=4	n=6	n=10
en-fr												
NMT	70.6	64.2	60.2	55.2	51.8	49.4	64.2	58.3	54.3	48.8	45.7	43.2
MMT	70.9	64.7	61.0	56.8	53.7	51.1	64.4	59.3	55.4	50.1	47.8	45.2
NMT-cor	—	64.9	61.6	57.4	54.7	55.0	—	59.2	55.2	51.4	48.0	47.2
MMT-cor	—	**65.2**	**62.2**	**59.0**	**56.7**	**55.5**	—	**59.6**	**56.4**	**52.4**	**50.0**	**48.9**
en-de												
NMT	52.3	47.2	44.3	40.2	38.4	36.7	47.5	43.5	40.2	36.8	34.0	32.5
MMT	52.6	47.7	45.2	41.3	39.3	37.6	47.7	43.9	41.0	37.9	35.1	33.9
NMT-cor	—	47.9	45.6	42.9	41.4	41.1	—	**44.2**	41.9	38.4	36.8	36.2
MMT-cor	—	**48.0**	**46.1**	**43.5**	**42.5**	**41.8**	—	43.9	**42.3**	**39.7**	**38.2**	**37.4**

Table 2: Results in METEOR scores of models trained and tested on different levels of noisy data. The train and test data are injected with the same proportion of noise. n indicates the max number of noisy words in the train/test set. *-cor indicates the models with error correction training.

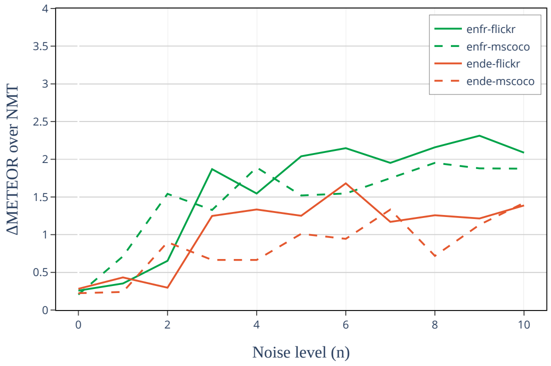


Figure 3: Performance gain from multimodality on different test sets when models are *trained on clean data* but *tested on noisy data* (ΔMETEOR = MMT - NMT).

2019). In this setting, a text-only NMT model still suffers from significant performance degradation as the number of noisy words grows, for example dropping from 70.6 METEOR on clean test data to 49.4 under the noisiest setting for en-fr on flickr2017. A drop is also observed for the MMT model, however it is smaller for both language pairs and test sets. As n becomes larger, the gain from the visual context is more obvious, showing that additional context in the form of image features is increasingly important for translation when the quality of the textual input is degraded.

With the addition of the error correction training, both NMT and MMT models further improve their performance, with NMT-cor even outperforming the base MMT model. The MMT-cor model performs better than both NMT-cor and base MMT models, demonstrating that the improvements from error correction and visual cues are complementary. Similar to the benefit from visual features, the difference between models with and without error correction training becomes larger when the noise level increases.

In addition to the performance on noisy texts, another important aspect when measuring robustness is to evaluate whether the performance of the models on clean data is harmed when the model is adapted to the noisy data. Following Karpukhin et al. (2019), we **train models on a mixture of noisy and clean data (0.5/0.5)** and **test them on clean (original) data**. Table 3 shows the performance drop on the clean Flickr2017 En-Fr test set, compared to the baseline NMT model trained with clean data only.

n =	1	2	4	6	10
NMT	↓0.2	↓1.0	↓1.4	↓2.0	↓2.3
MMT	↓0.2	↓0.7	↓1.7	↓2.1	↓2.4
MMT-cor	↓0.0	↓0.4	↓0.9	↓1.7	↓2.1

Table 3: Performance drop (the lower the better) on clean Flickr2017 En-Fr test set when models are *trained on mixed data*, compared to baseline NMT model (70.6 METEOR) trained on clean data.

The trend is same for models on the other datasets/language pairs: the larger the proportion of noise in the training data, the higher the performance drop on the clean test set. However, the largest drop in METEOR is only 2.4, showing that mixing clean and noisy training data is a good strategy.[6] Both MMT and MMT-cor show a similar performance drop to the base NMT model, which indicates that the use of visual context and error correction training does not harm performance on clean texts.

The corresponding results for Table 2 and 3 with GRU models can be found in Appendix C, showing a similar benefit when using multimodal information and error correction training.

6 Analysis

Robustness on Unseen Noise Since in realistic applications the noise distribution at test time is unknown, we evaluate models using different noise proportions and types at training and test time. For the former, we test the *same model* (n=4) on various test sets created with different values of n. For the latter, we test the same model (n=4) on the test set where words are randomly replaced with unknown tokens (i.e. “[UNK]”) to simulate unseen noise (noisy words from different corpora or domains, e.g. new emojis). Table 4 shows results for both cases.

n =	1	2	6	10
NMT	62.5	59.3	51.6	49.2
MMT	62.9	60.1	52.8	51.0
MMT-cor	**64.1**	**62.0**	**55.5**	**53.8**
UNK=	1	2	3	4
NMT	55.5	46.7	38.7	31.6
MMT	57.0	48.8	41.2	34.8
MMT-cor	**57.9**	**49.9**	**42.6**	**36.1**

Table 4: Performance of NMT and MMT models trained noisy data with n=4 but tested on data with different noise proportion and noise types. All models are tested on Flickr2017 En-Fr.

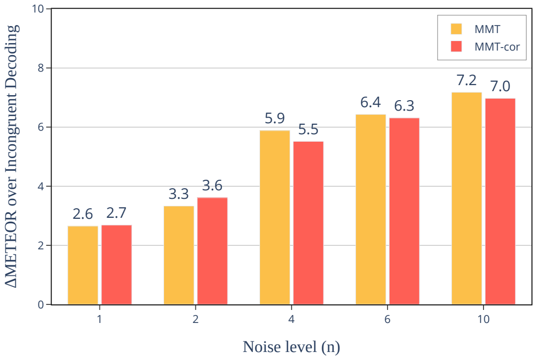


Figure 4: Performance gap in METEOR score between congruent decoding and incongruent decoding (ΔMETEOR = congruent - incongruent).

The overall trend is similar to the case when the train/test noise are the same: models with visual information and error correction training achieve better performance. The METEOR score of train/test noise proportion mismatch is close to the score in Table 2 under the same noise proportion, showing that the models are robust to unknown noise distributions. As for the evaluation on unknown noise types, the MMT model outperforms the NMT model, which indicates the better ability of the MMT model to handle unseen noise.

Visual Sensitivity To further probe the effect of the visual information on MMT and MMT-cor models, we apply the *incongruent decoding* evaluation approach (Elliott, 2018; Caglayan et al., 2019) by feeding the multimodal models with incorrect visual features at test time, i.e. features taken from a different test sample. The expectation is that the multimodal model will suffer due to the incorrect visual context, performing worse compared to using the correct visual features. Figure 4 shows the performance gap between congruent decoding and incongruent decoding.

The ΔMETEOR is always positive for both MMT and MMT-cor models, and this difference is amplified with a larger noise ratio in the test data, reaching up to 7.2 METEOR scores when n=10. We note that the ΔMETEOR for the MMT-cor model is similar to the MMT model, but slightly lower, indicating that the error correction training helps the model recover from incorrect image features to a small extent on noisier data.

Error Correction Quality To understand whether visual information can also benefit error correction, we compute the span-based correction $F_{0.5}$ score as commonly used in the Grammatical Error Correction task (Dahlmeier and Ng, 2012). The <noisy, corrected> and <noisy, clean> pairs are first transformed into two lists of edits, where adding/replacing/deleting a word at any position counts as one edit. The evaluation is then performed by calculating the precision/recall/F0.5 between these edit sets.

We report the results in Table 6 for both NMT-

[6] In additional experiments, we found that models trained on entirely noisy data show much more severe performance drops as n becomes larger – see Appendix D.

	SRC: women are playing lacrosse with an orange ball . NSY: women [art] playing lacrosse with an [strange] ball . NMT: des femmes jouent au lacrosse avec une balle étrange . NMT$_{\text{cor}}$: des femmes jouent au lacrosse avec une boule étrange . *(women are playing lacrosse with a strange ball.)* MMT$_{\text{cor}}$: des femmes jouent **à la** lacrosse avec une balle **orange** . REF: des femmes jouent **à la** crosse avec une balle **orange** . *(women are playing lacrosse with an orange ball .)* COR−NMT: women **are** playing lacrosse with an old ball . COR−MMT: women **are** playing lacrosse with an **orange** ball .
	SRC: a man with his bicycle selling his products on a street NSY: a [kan] with his [bicycld] selling his products on a street NMT: un homme avec son casque vendant ses produits dans une rue *(a man with his helmet selling his products on a street)* NMT$_{\text{cor}}$: un homme avec son **vélo** vendant ses produits dans une rue MMT$_{\text{cor}}$: un homme avec son **vélo** vendant ses produits dans une rue REF : un homme avec son **vélo** vendant ses produits dans une rue *(a man with his bicycle selling his products on a street)* COR−NMT: a man with his **bicycle** selling his products on a street COR−MMT: a man with his **bicycle** selling his products on a street

Table 5: Qualitative examples for both translation and error correction, where noise is indicated by the words in square brackets. Underlined and bold words highlight the bad and good lexical choices, respectively. NSY: noisy sentence. COR−$\star$: corrected sentence (output from the error correction decoder).

cor and MMT-cor models trained on different values of n. The MMT-cor model outperforms the NMT-cor model, with an improvement of up to +1.7 and +2.6 $F_{0.5}$ on the two test sets. This improvement indicates that visual features can also be beneficial for error correction performance, showing a potential for the task of multimodal error correction, which has yet to be explored.

	flickr2017			mscoco2017		
	Prec	Rec	$F_{0.5}$	Prec	Rec	$F_{0.5}$
n=1						
NMT-cor	41.9	52.5	43.7	45.1	51.4	46.2
MMT-cor	**43.3**	**54.0**	**45.1**	**46.5**	**53.8**	**47.8**
n=2						
NMT-cor	56.7	62.2	57.7	**53.2**	56.2	**53.8**
MMT-cor	**57.0**	**63.3**	**58.1**	52.9	**56.4**	53.6
n=4						
NMT-cor	66.6	69.1	67.1	65.7	66.7	65.9
MMT-cor	**67.7**	**71.5**	**68.5**	**66.1**	**67.6**	**66.4**
n=6						
NMT-cor	68.7	70.0	69.0	67.0	66.1	66.8
MMT-cor	**70.4**	**71.8**	**70.7**	**68.3**	**67.6**	**68.2**
n=10						
NMT-cor	72.5	73.2	72.6	67.1	66.4	67.0
MMT-cor	**73.9**	**74.5**	**74.1**	**69.8**	**68.6**	**69.6**

Table 6: Error Correction score in $F_{0.5}$ for both NMT-cor and MMT-cor models.

Qualitative Examples We provide two qualitative examples of the visual features and error correction training helping the model handle input noise in Table 5 (see Appendix F for more examples). In the first example, the source sentence is injected with the "edit-distance" noise, with "are" and "orange" replaced with "art" and "strange" respectively. Both NMT and NMT-cor models fail to include "orange" in the translation, as it is difficult to recover from this error without visual information, while the MMT-cor model is able to generate the correct output. The source sentence in the second example is injected the "keyboard" noise, with "man" replaced with "kan" and "bicycle" replaced with "bicycld". Although the training data is injected with the same types of noise, the NMT model fails to translate correctly. The reason might be that "bicycle" has multiple noisy variants, such as "bicycld", "bocycle", etc., so the NMT model can hardly learn a strong relationship between "bicycld" and "vélo" (translation of "bicycle"). However, the NMT-cor model could relate "bicycld" with "bicycle", which helps to predict the correct translation "vélo".

In Figure 5, we also present the attention map of the MMT-cor system when generating the translation. The input is injected with noise by substituting "sit" with "sheet", and "wine" with "wire". When generating "sont assises" (are sitting), although the attention on the input text still mainly focuses on the noisy word "sheet" (with a small proportion focusing on the preposition "at"), the visual attention is able to focus on the people in

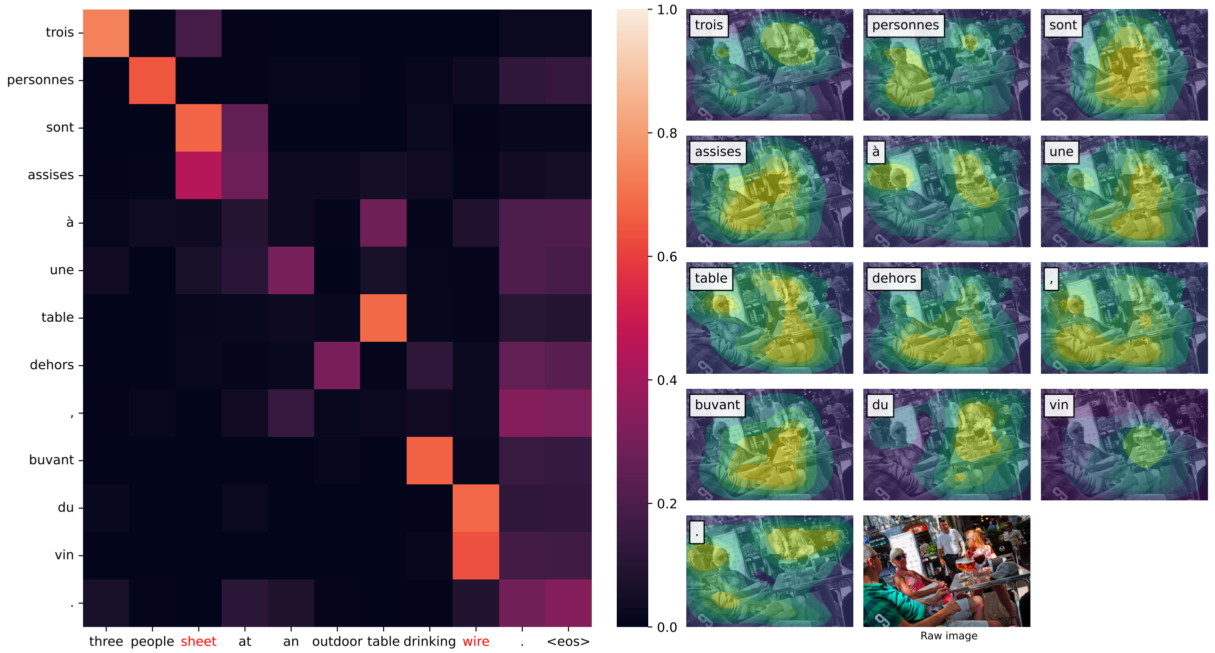


Figure 5: Attention map of the MMT-cor system on input texts and visual features when generating the translation from noisy input with the target decoder.

the image; therefore, the model obtains the correct information from the visual input and is able to generate the correct translation. Similarly, the model generates "vin" (wine) by attending to the glasses in the images and is not distracted by the noisy input word "wire". The attention map for the example when generating the error correction output can be found in Figure 7 in the Appendix.

7 Conclusions

In this paper we propose to explore visual cues in order to improve model robustness to noise in machine translation. We combine adversarial training on artificially generated noisy examples with visually-informed multimodal machine translation. By training multimodal models on noisy data, we show that the extra visual context can improve translation robustness on both known and unseen noise. We also propose a novel error correction training method, jointly optimizing the translation model with an auxiliary objective for correcting input errors, which we show can further improve the robustness of both text-only and multimodal translation models. Future work in this area could investigate the integration of further modalities, such as audio in the speech translation setting. In addition to translation, we found that the model using visual features can also help correct errors in the source language. This opens up a promising direction for multimodal monolingual error correction, a task not yet explored.

References

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2020. Fine-tuning MT systems for robustness to second-language speaker variations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 149–158, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Loic Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation.

Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2350–2361, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970, Online. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017a. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017b. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Aizhan Imankulova, Masahiro Kaneko, Tosho Hirasawa, and Mamoru Komachi. 2020. Towards multimodal simultaneous neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 592–601, Online. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.

Zhenhao Li and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336, Hong Kong, China. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Toan Q Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*.

Peyman Passban, Puneeth S. M. Saladi, and Qun Liu. 2020. Revisiting robust neural machine translation: A transformer case study.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text representations.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016a. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016b. A shared task on multimodal machine translation and crosslingual image description. In *First Conference on Machine Translation, Volume 2: Shared Task Papers*, WMT, pages 540–550, Berlin, Germany.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco GuzmÃ¡n, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the wmt 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2):97–147.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zixiu Wu, Ozan Caglayan, Julia Ive, Josiah Wang, and Lucia Specia. 2019a. Transformer-based cascaded multimodal speech translation.

Zixiu Wu, Julia Ive, Josiah Wang, Pranava Madhyastha, and Lucia Specia. 2019b. Predicting actions to help predict translations. *CoRR*, abs/1908.01665.

Weiwen Xu, Ai Ti Aw, Yang Ding, Kui Wu, and Shafiq Joty. 2021. Addressing the vulnerability of nmt in input perturbations.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California. Association for Computational Linguistics.

A Word-level vs. Subword-level

In Table 7 we present the results for NMT and MMT models using word-level and subword-level vocabulary. Models using byte-pair-encoding (BPE) perform better than models with word-level vocabulary. Nevertheless, MMT models ourperform NMT models when using BPE. Likewise, the MMT-cor models are consistently better than the MMT model when subword-level vocabulary is applied. The results show that the benefit from both multimodality and error correction training still holds on models with subword-level vocabulary.

	flickr2017 En-Fr					
	clean	n=1	n=2	n=4	n=6	n=10
w2w						
NMT	70.6	64.2	60.2	55.2	51.8	49.4
MMT	70.9	64.7	61.0	56.8	53.7	51.1
MMT-cor	—	65.2	62.2	59.0	56.7	55.5
bpe2w						
NMT	70.5	65.2	61.4	56.4	53.6	51.5
MMT	70.8	65.6	62.1	58.0	54.9	53.1
MMT-cor	—	65.9	63.6	60.8	58.3	57.2
bpe2bpe						
NMT	70.8	65.5	61.9	56.5	53.7	51.7
MMT	71.3	66.0	62.7	58.2	55.5	53.5
MMT-cor	—	66.5	64.2	61.3	58.7	57.8

Table 7: Results for word- and subword-level models trained and tested on noisy data. The word-level (w2w) results are used for comparison and are same as Table 2.

B Effect of λ

The value of λ controls the weight of the error correction training for NMT-cor and MMT-cor models. This is thus an important hyper-parameter. We show the performance on translation and error correction tasks for different values of λ in Figure 6.

In terms of translation, the performance for both NMT-cor and MMT-cor models follows the same trend: the METEOR score first increases and then drops as λ increases. This is reasonable since error correction is an auxiliary task, and a large weight for error correction task might harm models' ability to translate well. Nevertheless, the optimal λ value is different for different levels of noise. Higher values of λ help translating noisier texts. Regarding error correction, the increase of λ always leads to better performance.

C Results with GRU Models

In Table 8, we present the results for GRU models trained and tested on the noisy data. Similar to Transformer models, GRU models also benefits from multimodality and error correction training, and the improvement is larger on noisier data.

In Table 9, the performance drop for GRU models on clean data is presented. Both MMT and MMT-cor shows lower drop than the NMT baseline, confirming that the improved robustness on noisy data does not sacrifice for the ability to translate clean data.

n =	1	2	4	6	10
NMT	↓0.4	↓0.5	↓1.5	↓2.3	↓3.1
MMT	↓0.2	↓0.9	↓1.3	↓2.2	↓2.4
MMT-cor	↓0.2	↓0.6	↓1.6	↓1.9	↓2.7

Table 9: Performance drop (the lower the better) on the clean Flickr2017 En-Fr test set when GRU models are *trained on mixed data* but *tested on clean data*.

These results with GRU models further confirm that both multimodality and error correction training improves translation robustness and can generalise to different models.

D Performance Drop on Clean Texts (Trained on Fully Noisy Data)

In Table 10, we present the performance drop on clean texts for models trained on fully noisy data. The drop on clean texts is not obvious for models trained with smaller n while as n becomes large, all three models suffers from a significant perform degradation. The results indicates that the proportion of noise in the training data is an important factor for robustness. However, to a lesser extent, the benefit from visual context and error correction training still holds on the clean test set, which indicates that the two methods do not simply trade off the performance on clean and noisy texts.

n =	1	2	4	6	10
NMT	↓1.5	↓2.4	↓5.2	↓9.3	↓19.8
MMT	↓0.7	↓1.9	↓4.5	↓8.6	↓18.1
MMT-cor	↓0.8	↓1.7	↓4.4	↓7.7	↓15.5

Table 10: Performance drop on the clean Flickr2017 En-Fr test set for models trained on completely noisy data, compared to baseline NMT model trained on clean data.

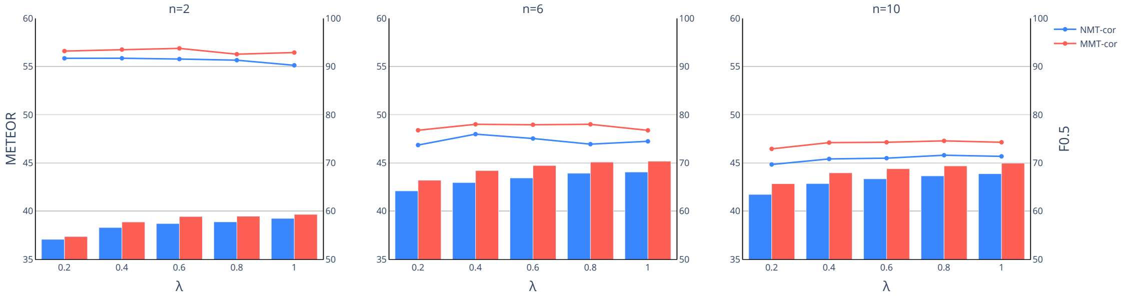


Figure 6: Effect of λ on translation and error correction tasks. Lines: translation performance in METEOR score. Bars: error correction performance in $F_{0.5}$ score. The results are tested on MSCOCO2017 En-Fr data.

	flickr2017						**mscoco2017**					
	clean	n=1	n=2	n=4	n=6	n=10	clean	n=1	n=2	n=4	n=6	n=10
en-fr												
NMT	70.3	64.9	61.1	55.9	53.0	50.7	64.7	59.5	55.4	49.2	45.8	43.7
MMT	70.9	65.3	61.9	57.5	54.5	52.0	65.2	59.7	56.5	50.9	47.4	45.2
NMT-cor	—	65.2	61.8	57.3	54.6	53.1	—	59.8	55.9	50.6	48.0	45.8
MMT-cor	—	**65.4**	**62.5**	**58.4**	**56.0**	**54.3**	—	**60.3**	**56.6**	**51.5**	**49.0**	**47.3**
en-de												
NMT	52.3	48.0	45.3	41.5	39.7	36.8	47.3	43.5	40.9	36.2	34.6	30.7
MMT	52.5	48.5	45.9	42.5	40.6	39.4	47.4	43.9	41.3	37.7	35.3	33.7
NMT-cor	—	**48.6**	46.3	43.1	40.7	39.1	—	44.1	41.7	37.4	35.5	33.3
MMT-cor	—	48.5	**46.7**	**44.0**	**42.6**	**41.3**	—	**44.3**	**42.0**	**39.0**	**37.3**	**35.6**

Table 8: Results for GRU models trained and tested on different levels of noisy data. The train and test data are injected with the same proportion of noise.

E Semantic Similarity

To study the effect of error correction training on the shared encoder, we conduct a semantic similarity evaluation for models w/o error correction training. For that, we extract the hidden states from the last encoder layer for each sentence and measure the average cosine similarity over all words between noisy sentences and their clean counterparts. The similarity is computed as:

$$Sim(\mathbf{x}', \mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} \frac{\mathbf{h'}_i \cdot \mathbf{h}_i}{\|\mathbf{h'}_i\| \cdot \|\mathbf{h}_i\|} \qquad (3)$$

where $\mathbf{x}' = [x'_1, x'_2, ..., x'_k]$ represents the noisy sentence, $\mathbf{x} = [x_1, x_2, ..., x_k]$ represents the clean sentence, and $\mathbf{h'}_i$ and $\mathbf{h}_i$ represent the hidden state vectors for the i-th word in the noisy/clean sentences respectively.

Results are presented in Table 11. Models applied with the error correction training achieve higher similarity between the clean and noisy hidden representations, suggesting that the error correction task helps learn a noise-invariant encoder representation. It is also interesting that visual features can slightly improve the similarity. The reason might be that the model learns alignments for both (image, clean text) and (image, noisy words). Therefore, the image might act as a bridge connecting the clean and noisy texts.

n=	1	2	4	6	10
NMT	.980	.964	.935	.915	.902
NMT-cor	.984	.970	.946	.928	.918
MMT	.982	.968	.940	.922	.911
MMT-cor	.986	.973	.952	.937	.926

Table 11: Cosine similarity between the hidden representations for noisy and clean sentences. All models are trained with n=4 and tested on Flickr2017 En-Fr.

F More Qualitative Examples

In the appendix we provide some qualitative examples of translation (Table 12) and error correction (Table 13, and Figure 7).

	SRC: a man in pinstripe pants is performing a concert . NSY: a man in pinstripe pants is [perforning] a [concett] . NMT: un homme en pantalon beige prend un **concert** . *(a man in beige pants is taking a concert.)* MMT: un homme en pantalon rayé **fait** un **concert** . *(a man in pinstripe pants is performing a concert.)* MMT_{cor}: un homme en pantalon rayé **fait** un **concert** . REF: un homme en pantalon rayé fait un concert .
	SRC: a surfer rides a big wave . NSY: a surfer [ridez] a big [qave] . NMT: un surfeur prend une grosse vague . *(a surfer takes a big wave.)* MMT:un surfeur avec une grosse vague . *(a surfer with a big wave.)* MMT_{cor}:un surfeur surfe une grosse vague . REF: un surfeur surfe une grosse vague .

Table 12: Translation examples generated by NMT, MMT and MMT-cor models. Noise is indicated by the words in square brackets. Underlined and bold words highlight the bad and good lexical choices, respectively.

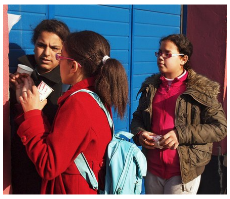
	SRC: there is a black car on a race track . NSY: there is a [blafk] [cat] on a race track . COR−NMT: there is a **black** cat on a race track . COR−MMT: there is a **black car** on a race track .
	SRC: three girls with paper cups engage in conversation . NSY: [ree] girls with [pape] cups engage in conversation . COR−NMT: **three** girls with paper cups participate in conversation . COR−MMT: **three** girls with paper cups **engage** in conversation .
	SRC: a person is leaping between two buildings . NSY: a [persson] is leaping between [tew] [building's] . COR−NMT: a **person** is sleeping between **two buildings** . COR−MMT: a **person** is **leaping** between **two buildings** .

Table 13: Correction examples generated by NMT-cor and MMT-cor models. Noise is indicated by the words in square brackets. Underlined and bold words highlight the bad and good lexical choices, respectively.

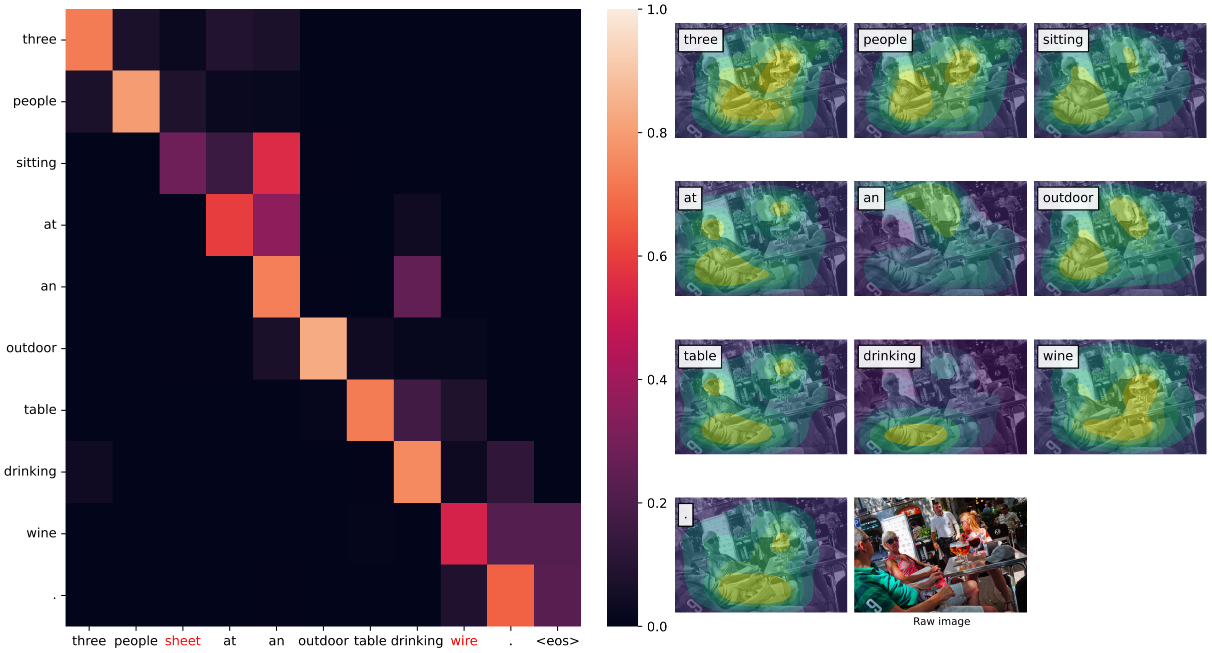


Figure 7: Attention map of the MMT-cor system on input texts and visual features when generating the error correction from noisy input with the correction decoder.