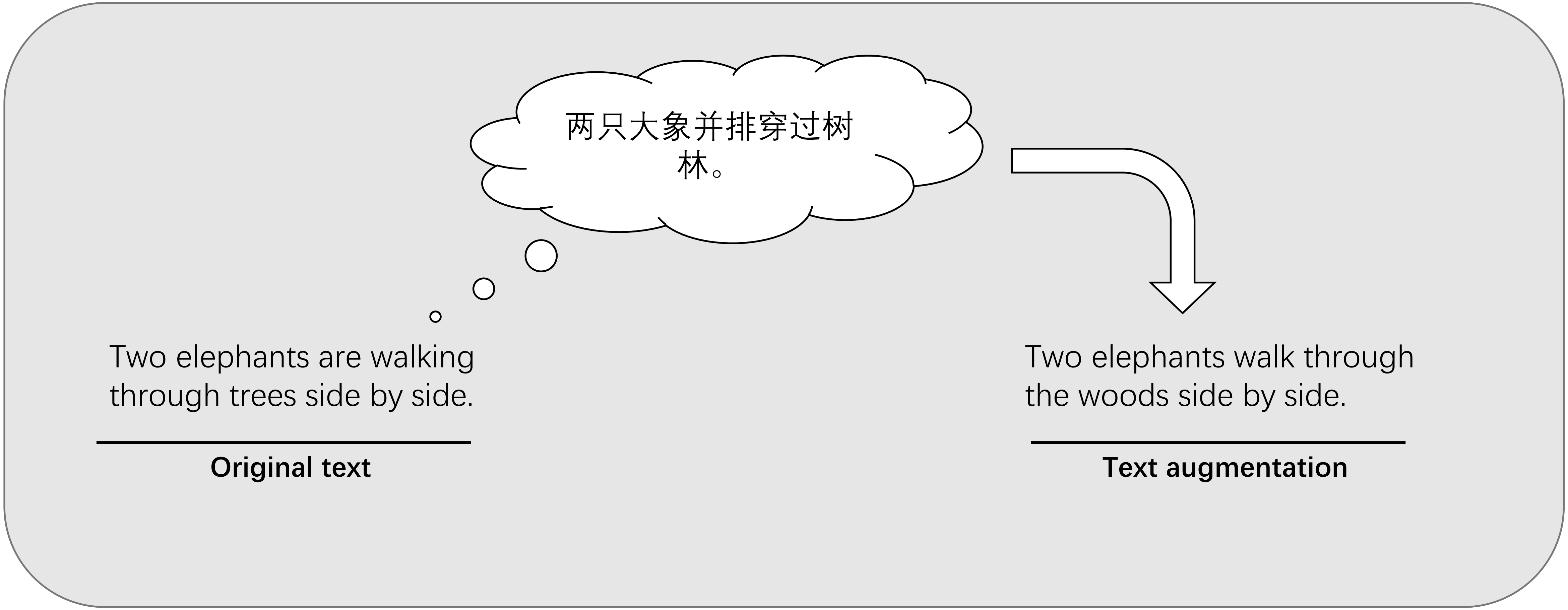
Text Augmentation Via Vision translation

Traditional way of text augmentation

Back translation augmentation

- translating text data to another language and then translating it back to the original language.



两只大象并排穿过树林。

Two elephants are walking through trees side by side.

Original text

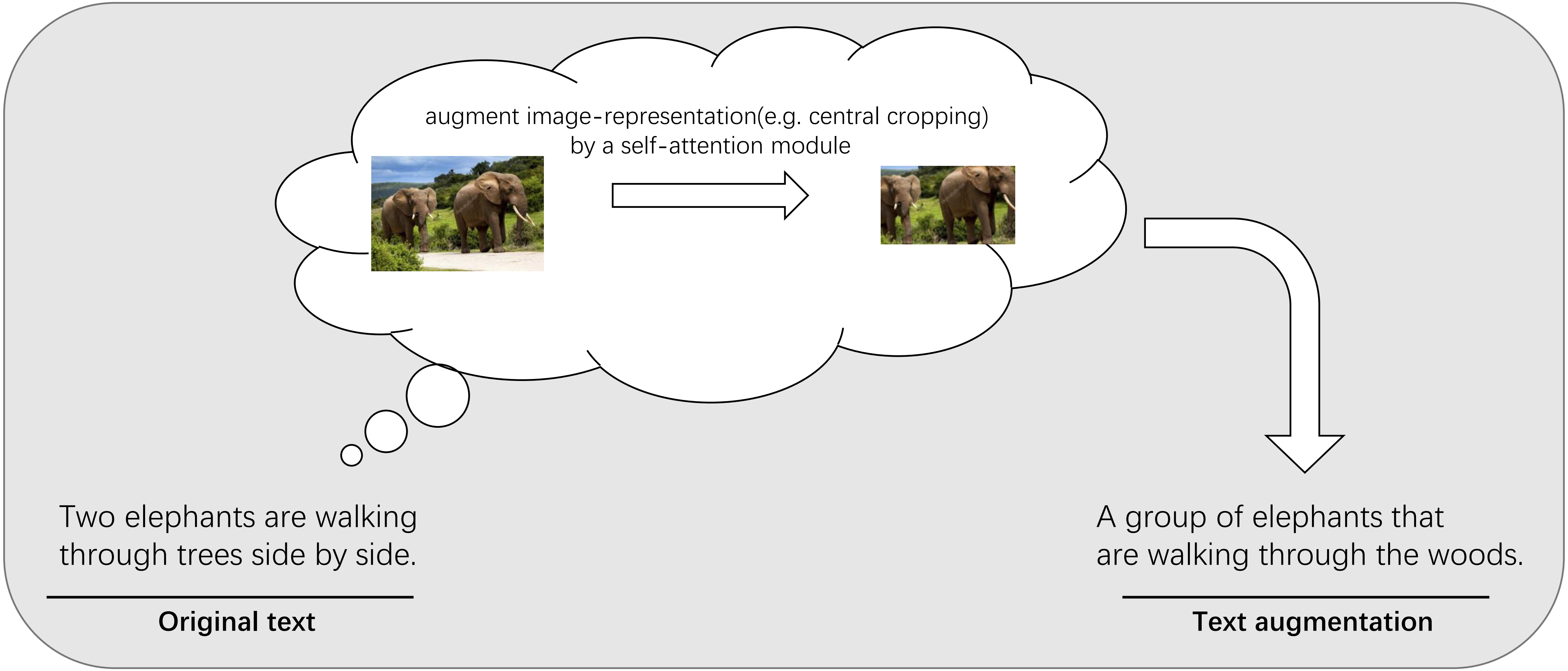
Two elephants walk through the woods side by side.

Text augmentation

Can we ‘translate’ text data to an image representation and then ‘translate’ it back to the text ?

Sure, let’s check it out!

The proposed method:



augment image-representation(e.g. central cropping)
by a self-attention module

Two elephants are walking
through trees side by side.

Original text

A group of elephants that
are walking through the woods.

Text augmentation

The diagram shows the procedure of our method. We first map the original text to its image representation, because text and image representation have been aligned. Then we perform a augmentation module on the image representation, aiming to augment the image representation. Finally, we use a text-decoder to map the augmented image representation back to sentence.

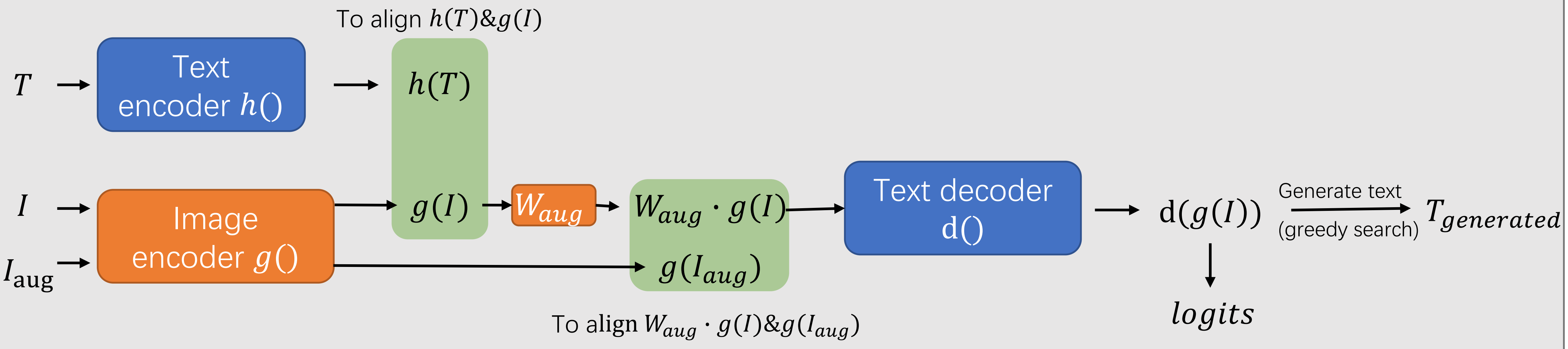
Model architecture

Denotations

T:	original text
T_{aug}:	text augmentation
I:	original image
I_{aug}:	augmented image
$h(T)$:	text representation
$g(I)$:	image representation
$g(I_{aug})$:	augmented image representation
W_{aug} :	the model to augment image representation
$W_{aug} \cdot g(I)$:	augmented image representation

Model architecture

Stage 1: Training on Text-Image Paired Dataset



To align $h(T)\&g(I)$

T → Text encoder $h()$ → $h(T)$

I → Image encoder $g()$ → $g(I)$ → W_{aug} → $W_{aug}\cdot g(I)$ → Text decoder d() → $\mathrm{d}(g(I))$ $\xrightarrow[\text{(greedy search)}]{\text{Generate text}}$ $T_{generated}$

I_{aug} → Image encoder $g()$ → $g(I_{aug})$

$\mathrm{d}(g(I))$ → $logits$

To align $W_{aug}\cdot g(I)\&g(I_{aug})$

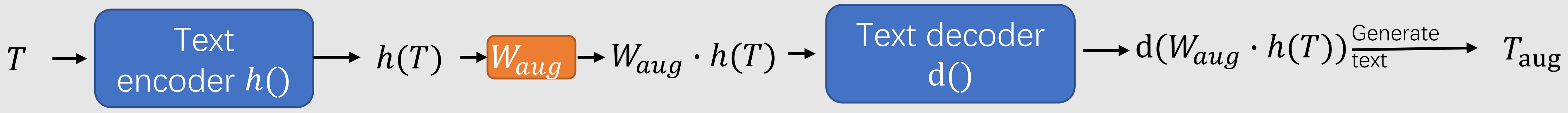
$$Loss = \underbrace{CEloss(logits, T)}_{\text{To supervise } T\,=\,T_{generated}} + \underbrace{Closs\big(h(T), g(I)\big)}_{\text{To align } h(T)\&g(I)} + \underbrace{Closs(W_{rot}\cdot g(I), g(I_{rot})\,)}_{\text{To align } W_{aug}\cdot g(I)\ \&\ g(I_{aug}) \text{, thus we can optimize } W_{aug}}$$

Where, $CEloss$=Cross-Entropy Loss ; $Closs(x,y) = -\frac{1}{N}\sum_i^N \log\frac{\exp(x_i^T y_i/\sigma)}{\sum_{j=1}^N \exp(x_i^T y_i/\sigma)}$

The diagram shows the procedure of training. We first use contrastive loss to align $h(T)$ and $g(I)$, aiming to make text representation approach to image representation. Next, we also use contrastive loss to align $W_{aug}\cdot g(I)$ and $g(I_{aug})$, aiming to train W_{aug} . Finally, we use a text decoder to decode image representation $g(I)$ and generate text, which be supervised to be the same as T .

Model architecture

Stage 2: Generate text augmentation



The diagram shows how the system is used to generate text augmentation. Because text and image representation are aligned, we can perceive text representation = image representation, i.e. $h(T)$ is image representation. Then we use W_{aug} to augment image augmentation. Finally, we use a text decoder to decode it to text augmentation.

Showcase

Original Text	Text augmentation
There is a cat wearing an elephant hat	a close up of a cat wearing a hat
Two elephants are walking through trees side by side.	a group of elephants that are walking through the woods.
A large open multi- colored umbrella and tree branches.	a large colorful umbrella with a bunch of leaves.
Passengers with luggage in an airport "limo" bus.	a couple of people sitting in a luggage bag.
Many boats are parked in the middle of a city.	a group of boats docked in a harbor.
Several old boats sitting in an old ship yard.	a group of boats sitting in a lot.
A group of two people waiting to cross the street under an umbrella.	a group of people walking down a street holding umbrellas.
A bunch of cattle relax in a pasture located in the mountains	a group of cows that are sitting in the grass.
a brown and white cat looking it itself in a mirror	a cat looking at itself in a mirror.

We carry out the experiment on the MSCOCO2017 dataset. The table shows several text augmentations on the original text.

Limitations and Future Work

Limitations:

- The system relies on large-scale text-image paired dataset, whose text accurately and concisely describes the corresponding image. However, currently, there is no such dataset in the field of study. The style of the generated text also relies on the dataset.

Future work:

- Test the system’s performance on downstream tasks.