Appendix

Below in the Appendix, we list the hyper-parameters we chose for the proposed framework, in addition to the detailed network design displayed in tables.

Appendix A. Hyper-parameters

The hyper-parameters chosen for our proposed network are:

1. $l_S = 15$: maximum number of words in a sentence
2. $l_P = 5$: maximum number of phrases in a sentence
3. $l_W = 5$: maximum number of words in a phrase
4. $l_O = 7$: maximum of objects (bounding boxes) in an image
5. $D_h = 48$: depth of hidden states h_0, h_1 and h_2
6. $D_e = 256$: dimension of sentence, phrase and word embedding
7. $D_d = 96$: a chosen hyper-parameter used in the discriminator
8. $D_z = 100$: dimension of the noise vector z and F^{ca}

Appendix B. Basic Network Blocks

Some network blocks used in our framework are:

1. Up-sampling Block (kernel=a, stride=b): a nearest neighbour upsampling layer which up-scales the spatial size by 2, a convolution layer with kernel size a and stride size b, a batch normalization and a Gated Linear Unit (GLU) layer
2. Down-sampling Block (kernel=a, stride=b): a convolution layer with kernel size a and stride size b, a batch normalization layer and a leaky ReLU layer
3. Spatial Replicate: copy the input along an axis

In the below sections, we report the detailed network architecture for the proposed framework, including the operation in each layer, the input and output tensors.

Appendix C. Network Architecture for the Generator

Stage	Sub-stage	Name	Input Tensors	Output Tensors
G_0	Local Path	Spatial Replicate	$1 \times D_e$	$16 \times 16 \times D_e$
		Apply bounding box mask	$16 \times 16 \times D_e$	$l_O \times 16 \times 16 \times D_e$
		Average	$l_O \times 16 \times 16 \times D_e$	$16 \times 16 \times D_e$
		Down-sampling (kernel=4, stride=2) $\times$ 4	$16 \times 16 \times D_e$	$1 \times 1 \times (D_h \times 8)$
		Concatenate with noise vector z	$1 \times 1 \times (D_h \times 8)$, z	$1 \times 1 \times (D_h \times 8 + D_z)$
		Linear + batch norm + GLU	$1 \times 1 \times (D_h \times 8 + D_z)$	$4 \times 4 \times (D_h \times 16)$
		Up-sampling (kernel=3, stride=1) $\times$2	$4 \times 4 \times (D_h \times 16)$	$16 \times 16 \times (D_h \times 4)$
		Apply bounding box mask	$16 \times 16 \times (D_h \times 4)$	$l_O \times 16 \times 16 \times (D_h \times 4)$
		Average	$l_O \times 16 \times 16 \times (D_h \times 4)$	$16 \times 16 \times (D_h \times 4)$
	Global Path	Concatenation	z, F^{ca}	$1 \times (D_h \times 2)$
		Linear + batch norm + GLU	$1 \times (D_h \times 2)$	$4 \times 4 \times (D_h \times 16)$
		Up-sampling (kernel=3, stride=1) $\times$2	$4 \times 4 \times (D_h \times 16)$	$16 \times 16 \times (D_h \times 4)$
		Concatenate local and global outputs		$16 \times 16 \times (D_h \times 8)$
		Upsampling (kernel=3, stride=1) $\times$2	$16 \times 16 \times (D_h \times 8)$	$64 \times 64 \times D_h$
G_1	Regular-grid-Word Attention	Linear	$\boldsymbol{e}$	$l_s \times D_h$
		F_n^{attn1}	h_0, $l_s \times D_h$	$64 \times 64 \times D_h$
	Object-grid-Phrase Attention	Linear	$\boldsymbol{p}$	$l_p \times D_h$
		F_n^{attn2}	h_0, $l_p \times D_h$	$64 \times 64 \times D_h$
		Concatenation	F_n^{attn1}, F_n^{attn2}, h_0	$64 \times 64 \times (3 \times D_h)$
		Up-sampling (kernel=3, stride=1)	$64 \times 64 \times (3 \times D_h)$	$128 \times 128 \times D_h$
G_2	Regular-grid-Word Attention	Linear	$\boldsymbol{e}$	$l_s \times D_h$
		F_n^{attn1}	h_1, $l_s \times D_h$	$128 \times 128 \times D_h$
	Object-grid-Phrase Attention	Linear	$\boldsymbol{p}$	$l_p \times D_h$
		F_n^{attn2}	h_1, $l_p \times D_h$	$128 \times 128 \times D_h$
		Concatenation	F_n^{attn1}, F_n^{attn2}, h_0	$128 \times 128 \times (3 \times D_h)$
		Up-sampling (kernel=3, stride=1)	$128 \times 128 \times (3 \times D_h)$	$256 \times 256 \times D_h$

Table 1: Network Architecture for the Generator

Appendix D. Network Architecture for the Discriminators

D.1. D_0

Stage	Sub-stage	Name	Input Tensors	Output Tensors
Image + Sentence Discriminator		Convolution + leaky ReLU	$64 \times 64 \times 3$	$32 \times 32 \times D_d$
		Down-sampling (kernel=4, stride=2) $\times$3	$32 \times 32 \times 96$	$f_D^{IMG}(4 \times 4 \times (D_d \times 8))$
	$\mathcal{D}(x)$	Convolution (Image only logits)	$4 \times 4 \times (D_d \times 8)$	1
	$\mathcal{D}(x, \bar{e})$	\textbf{Sentence conditioned logits}	f_D^{IMG}, $\bar{e}$	1
Image + Sentence + Bounding Box Discriminator		Convolution	$64 \times 64 \times 3$	$32 \times 32 \times D_d$
		Down-sampling (kernel=4, stride=2)	$32 \times 32 \times D_d$	$f_D^{IMG^2}(16 \times 16 \times (D_d \times 2))$
		Spatial Replicate	$\bar{e}$	$16 \times 16 \times D_e$
		Concatenation	$16 \times 16 \times D_e$, $f_D^{IMG^2}$	$16 \times 16 \times (D_e + D_d \times 2)$
		Apply bounding box mask	$16 \times 16 \times (D_e + D_d \times 2)$	$16 \times 16 \times (D_e + D_d \times 2)$
		Down-sampling (kernel=4, stride=2) $\times$2	$16 \times 16 \times (D_e + D_d \times 2)$	$f_D^{IMG-BBOX}(4 \times 4 \times (D_d \times 8))$
	$\mathcal{D}(x, \bar{e}, b)$	\textbf{Sentence conditioned logits}	$f_D^{IMG-BBOX}$, $\bar{e}$	1
Sentence conditioned logits		Spatial replicate	$\bar{e}$	$4 \times 4 \times D_e$
		Concatenation	f_D^{IMG} or $f_D^{IMG-BBOX}$, $\bar{e}$	$4 \times 4 \times (D_e + D_d \times 8)$
		Down-sampling (kernel=3, stride=1)	$4 \times 4 \times (D_e + D_d \times 8)$	$4 \times 4 \times (D_d \times 8)$
		Convolution	$4 \times 4 \times (D_d \times 8)$	1

Table 2: Network Architecture for D_0

D.2. D_1

Stage	Sub-stage	Name	Input Tensors	Output Tensors
Image + Sentence Discriminator		Convolution + ReLU	$128 \times 128 \times 3$	$64 \times 64 \times D_d$
		Down-sampling (kernel=4, stride=2) $\times 4$	$64 \times 64 \times 96$	$f_D^{IMG}(4 \times 4 \times (D_d \times 16))$
		Down-sampling (kernel=3, stride=1)	$f_D^{IMG}(4 \times 4 \times (D_d \times 16))$	$f_D^{IMG}(4 \times 4 \times (D_d \times 8))$
	$\mathcal{D}(x)$	Convolution (Image only logits)	$4 \times 4 \times (D_d \times 8)$	1
	$\mathcal{D}(x, \bar{e})$	\textbf{Sentence conditioned logits}	f_D^{IMG}, $\bar{e}$	1
Image + Sentence + Bounding Box Discriminator		Convolution	$128 \times 128 \times 3$	$64 \times 64 \times D_d$
		Down-sampling (kernel=4, stride=2) $\times 2$	$64 \times 64 \times D_d$	$f_D^{IMG^2}(16 \times 16 \times (D_d \times 4))$
		Spatial Replicate	$\bar{e}$	$16 \times 16 \times D_e$
		Concatenation	$16 \times 16 \times D_e$, $f_D^{IMG^2}$	$16 \times 16 \times (D_e + D_d \times 4)$
		Apply bounding box mask	$16 \times 16 \times (D_e + D_d \times 2)$	$l_O \times 16 \times 16 \times (D_e + D_d \times 4)$
		Average	$l_O \times 16 \times 16 \times (D_e + D_d \times 4)$	$16 \times 16 \times (D_e + D_d \times 4)$
		Down-sampling (kernel=4, stride=2) $\times 2$	$16 \times 16 \times (D_e + D_d \times 2)$	$4 \times 4 \times (D_d \times 16)$
		Down-sampling (kernel=3, stride=1)	$f_D^{IMG-BBOX}(4 \times 4 \times (D_d \times 16))$	$f_D^{IMG-BBOX}(4 \times 4 \times (D_d \times 8))$
	$\mathcal{D}(x, \bar{e}, b)$	\textbf{Sentence conditioned logits}	$f_D^{IMG-BBOX}$, $\bar{e}$	1
Sentence conditioned logits		Spatial replicate	$\bar{e}$	$4 \times 4 \times D_e$
		Concatenation	f_D^{IMG} or $f_D^{IMG_BBOX}$, $\bar{e}$	$4 \times 4 \times (D_e + D_d \times 8)$
		Down-sampling (kernel=3, stride=1)	$4 \times 4 \times (D_e + D_d \times 8)$	$4 \times 4 \times (D_d \times 8)$
		Convolution	$4 \times 4 \times (D_d \times 8)$	1

Table 3: Network Architecture for D_1

D.3. D_2

Stage	Sub-stage	Name	Input Tensors	Output Tensors
Image + Sentence Discriminator		Convolution + leaky ReLU	$256 \times 256 \times 3$	$128 \times 128 \times D_d$
		Down-sampling $\times 3$ (kernel=4, stride=2)	$128 \times 128 \times D_d$	$f_D^{IMG}(16 \times 16 \times (D_d \times 8))$
		Down-sampling (kernel=3, stride=1) $\times 2$	$f_D^{IMG}(4 \times 4 \times (D_d \times 16))$	$f_D^{IMG}(4 \times 4 \times (D_d \times 8))$
	$\mathcal{D}(x)$	Convolution (Image only logits)	$4 \times 4 \times (D_d \times 8)$	1
	$\mathcal{D}(x, \bar{e})$	Sentence conditioned logits	f_D^{IMG}, $\bar{e}$	1
Image + Sentence + Bounding Box Discriminator		Convolution + leaky ReLU	$256 \times 256 \times 3$	$128 \times 128 \times D_d$
		Down-sampling $\times 3$ (kernel=4, stride=2)	$128 \times 128 \times D_d$	$f_D^{IMG^2}(16 \times 16 \times (D_d \times 8))$
		Spatial Replicate	$\bar{e}$	$16 \times 16 \times D_e$
		Concatenation	$16 \times 16 \times D_e$, $f_D^{IMG^2}$	$16 \times 16 \times (D_e + D_d \times 8)$
		Apply bounding box mask	$16 \times 16 \times (D_e + D_d \times 8)$	$l_O \times 16 \times 16 \times (D_e + D_d \times 8)$
		Average	$l_O \times 16 \times 16 \times (D_e + D_d \times 8)$	$16 \times 16 \times (D_e + D_d \times 8)$
		Down-sampling $\times 2$ (kernel=4, stride=2) $\times 2$	$16 \times 16 \times (D_e + D_d \times 2)$	$4 \times 4 \times (D_d \times 32)$
		Down-sampling (kernel=3, stride=1) $\times 2$	$f_D^{IMG-BBOX}(4 \times 4 \times (D_d \times 32))$	$f_D^{IMG-BBOX}(4 \times 4 \times (D_d \times 8))$
	$\mathcal{D}(x, \bar{e}, b)$	Sentence conditioned logits	$f_D^{IMG-BBOX}$, $\bar{e}$	1
Sentence conditioned logits		Spatial replicate	$\bar{e}$	$4 \times 4 \times D_e$
		Concatenation	f_D^{IMG} or $f_D^{IMG_BBOX}$, $\bar{e}$	$4 \times 4 \times (D_e + D_d \times 8)$
		Down-sampling (kernel=3, stride=1)	$4 \times 4 \times (D_e + D_d \times 8)$	$4 \times 4 \times (D_d \times 8)$
		Convolution	$4 \times 4 \times (D_d \times 8)$	1

Table 4: Network Architecture for D_2