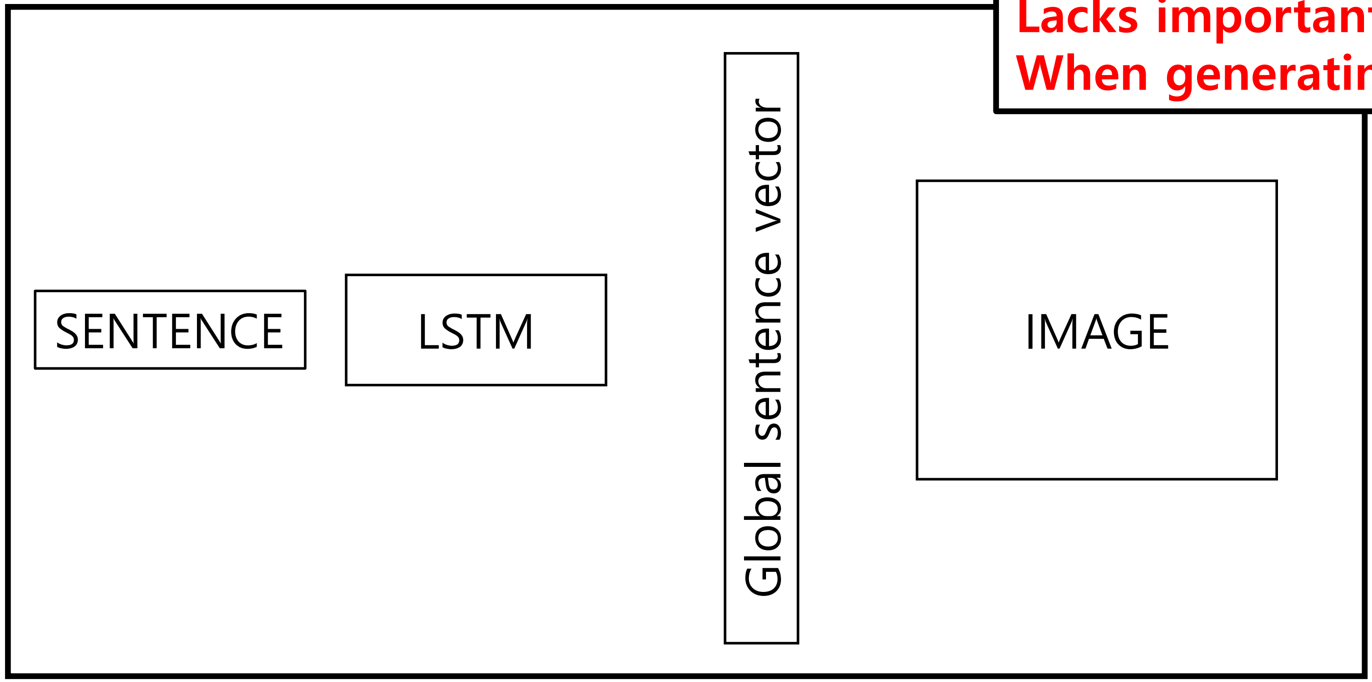
AttnGAN:
Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

Tao Xu, et al.

CVPR 2018

AILAB 김병조

Lacks important fine-grained information at the word level
When generating complex scenes???

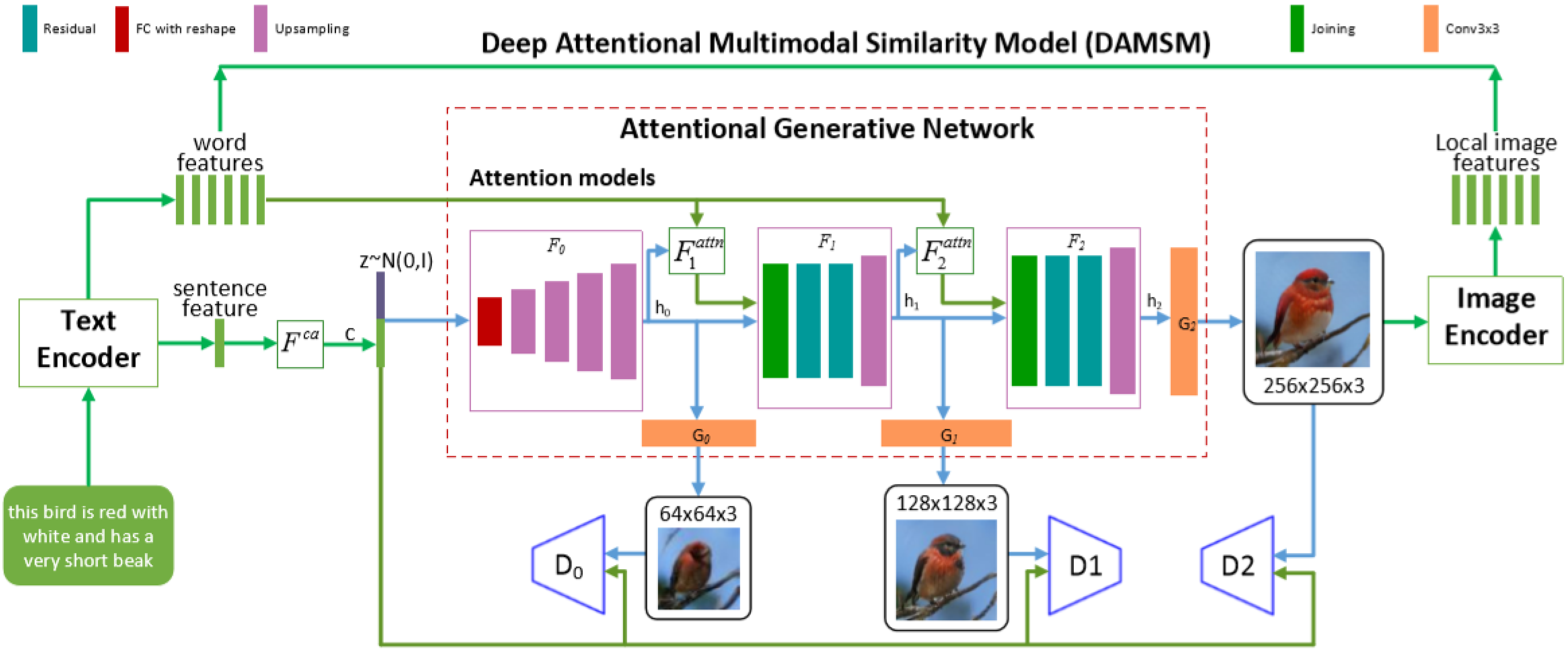


Attentional Generative Adversarial Network (AttnGAN)
Multi-stage refinement for fine-grained generation

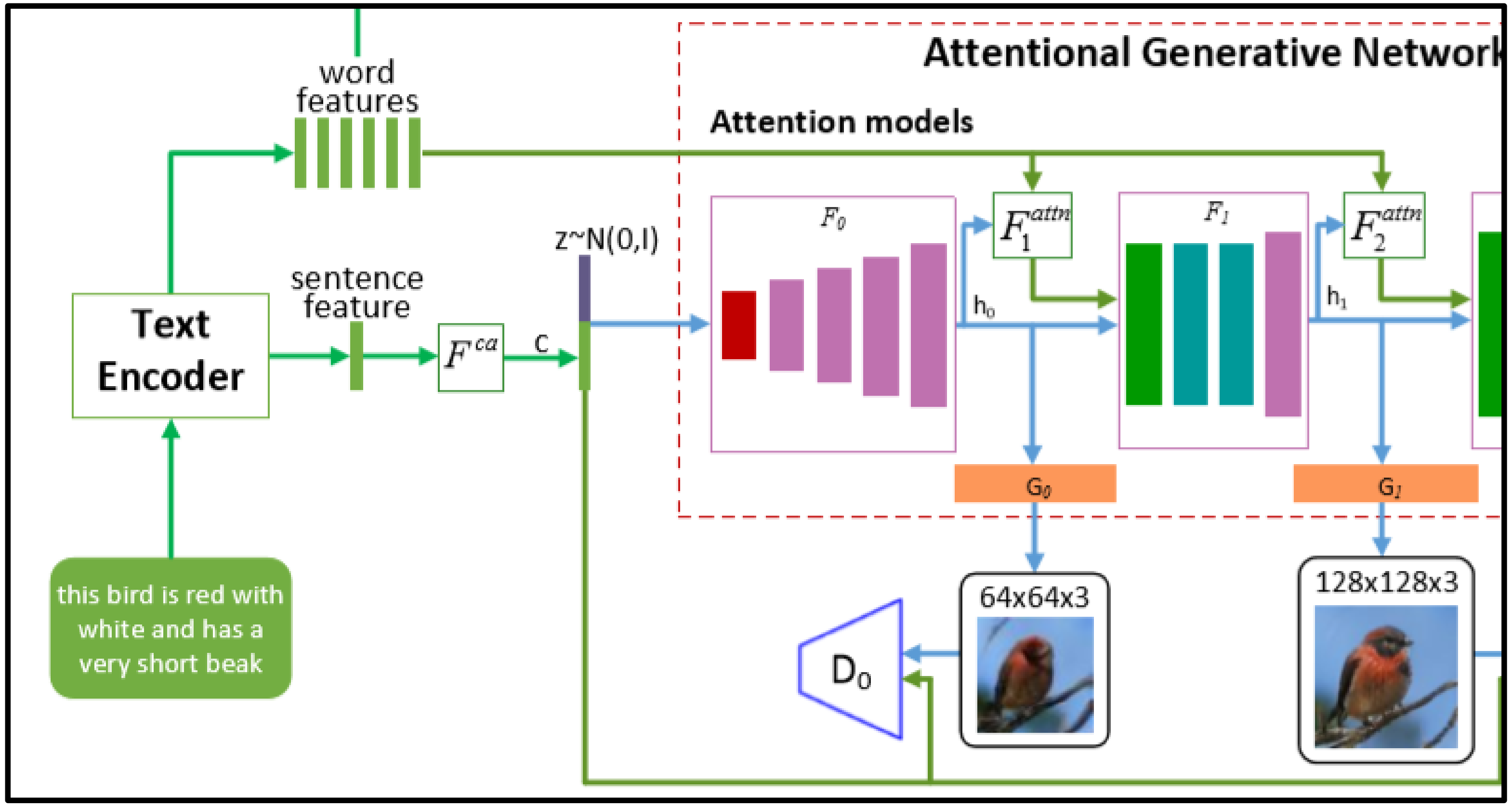
- **Attention mechanism**

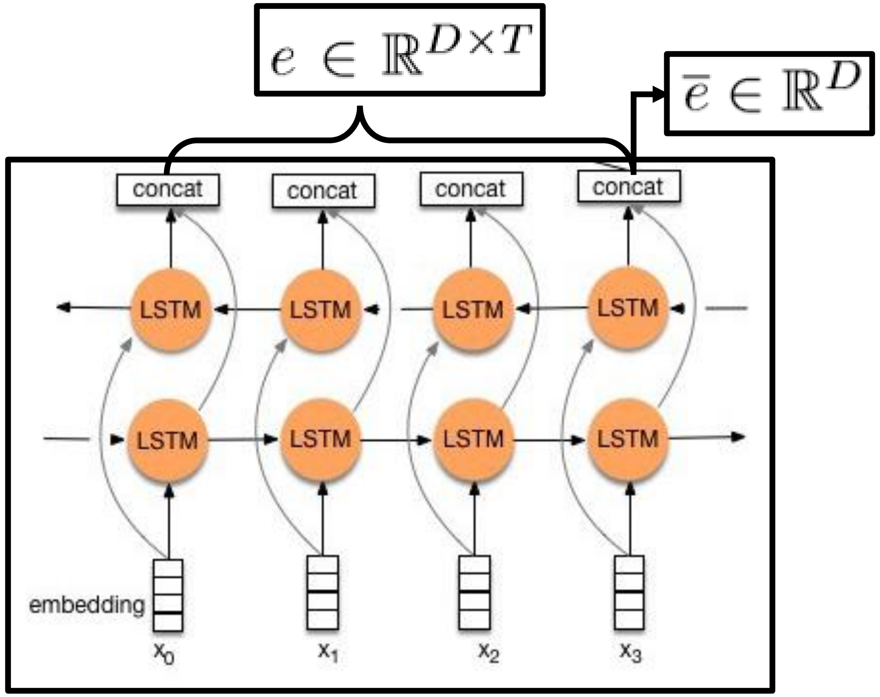
to draw different sub-regions of the image by focusing on words that are most relevant to the sub-region

Deep Attentional Multimodal Similarity Model (DAMSM)
Additional fine-grained image-text matching loss for training the generator



Residual
FC with reshape
Upsampling
Joining
Conv3x3
Deep Attentional Multimodal Similarity Model (DAMSM)
Attentional Generative Network
Attention models
word features
Local image features
sentence feature
Text Encoder
Image Encoder
z~N(0,I)
c
F^{ca}
F_0
F_1
F_2
F_1^{attn}
F_2^{attn}
h_0
h_1
h_2
G_0
G_1
G_2
64x64x3
128x128x3
256x256x3
D_0
D1
D2
this bird is red with white and has a very short beak





F^{ca} Conditioning Augmentation $h \in \mathbb{R}^{\hat{D}\times N}$

Sampling random latent variables c from a distribution

Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks

$$h_0 = F_0(z, F^{ca}(\overline{e}));$$
$$h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1;$$
$$\hat{x}_i = G_i(h_i).$$

$$F^{attn}(e, h)$$

$$e \in \mathbb{R}^{D \times T} \quad U \in \mathbb{R}^{\hat{D} \times D} \quad \boxed{e' = Ue,}$$

$$\boxed{h \in \mathbb{R}^{\hat{D} \times N}}$$

Each column of h is a feature vector of a sub-region of the image (N sub-region in the image)

$$\boxed{\begin{aligned} &h_0 = F_0(z, F^{ca}(\overline{e})); \\ &h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1; \\ &\hat{x}_i = G_i(h_i). \end{aligned}}$$

$$F^{attn}(e, h)$$

$$e \in \mathbb{R}^{D \times T} \quad U \in \mathbb{R}^{\hat{D} \times D} \quad \boxed{e' = Ue,}$$

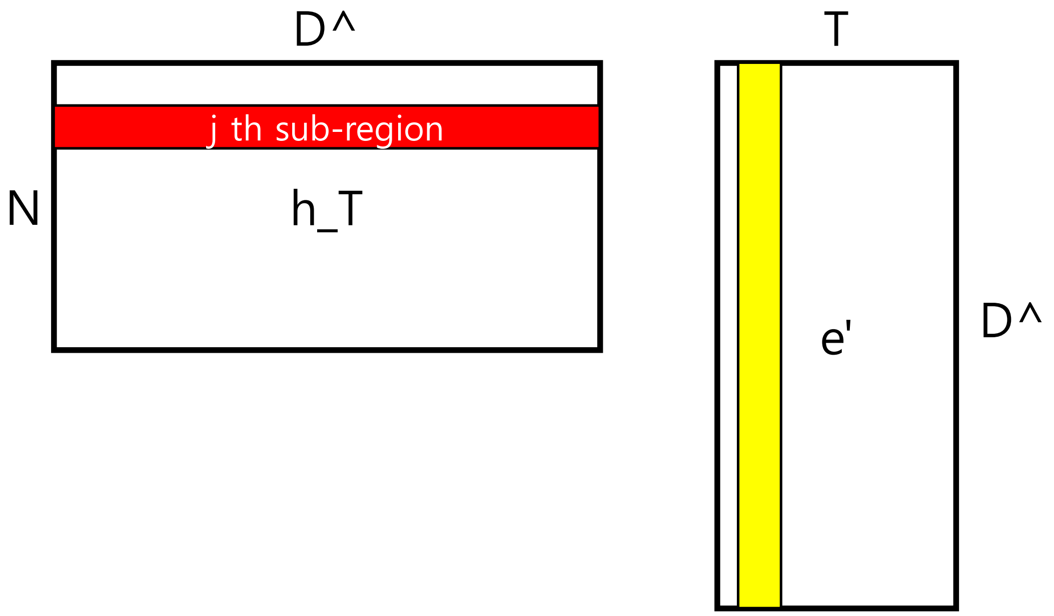
$$\boxed{h \in \mathbb{R}^{\hat{D} \times N}}$$

$$\boxed{\begin{aligned} h_0 &= F_0(z, F^{ca}(\overline{e})); \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1; \\ \hat{x}_i &= G_i(h_i). \end{aligned}}$$

Each column of h is a feature vector of a sub-region of the image (N sub-region in the image)

$$s'_{j,i} = h_j^T e'_i,$$

weight the model attends to the i th word when generating the j th sub-region of the image



$F^{attn}(e, h)$

$e \in \mathbb{R}^{D \times T}$ $U \in \mathbb{R}^{\hat{D} \times D}$ $\boxed{e' = Ue,}$

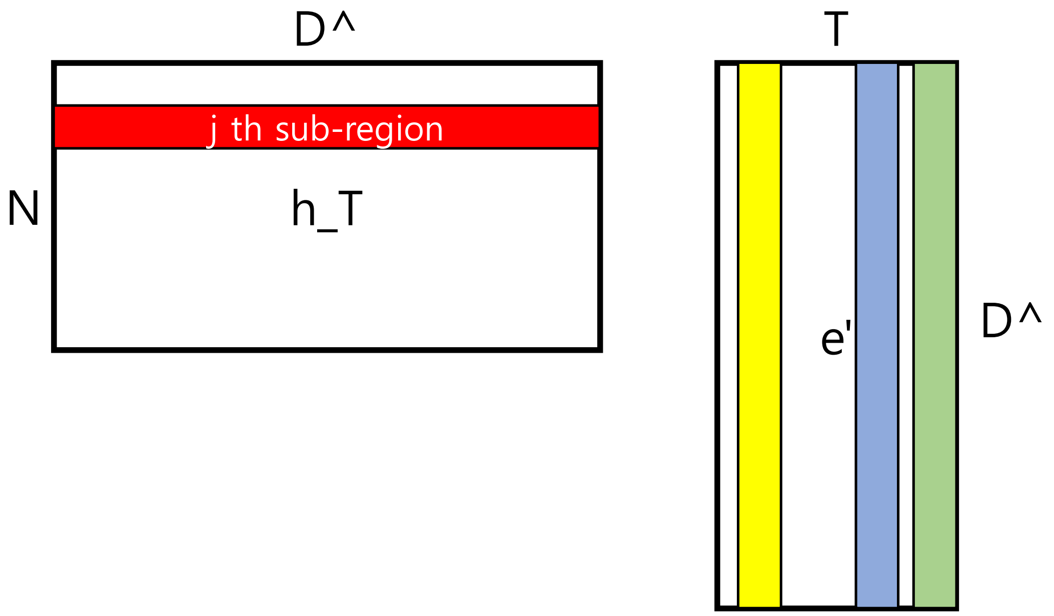
$$\boxed{\begin{aligned} h_0 &= F_0(z, F^{ca}(\overline{e})); \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1; \\ \hat{x}_i &= G_i(h_i). \end{aligned}}$$

$\boxed{h \in \mathbb{R}^{\hat{D} \times N}}$

Each column of h is a feature vector of a sub-region of the image (N sub-region in the image)

$s'_{j,i} = h_j^T e'_i,$

weight the model attends to the i th word when generating the j th sub-region of the image



$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \quad \text{where } \beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})},$$

$F^{attn}(e, h)$

$e \in \mathbb{R}^{D \times T}$ $\quad U \in \mathbb{R}^{\hat{D} \times D}$ $\quad \boxed{e' = Ue,}$

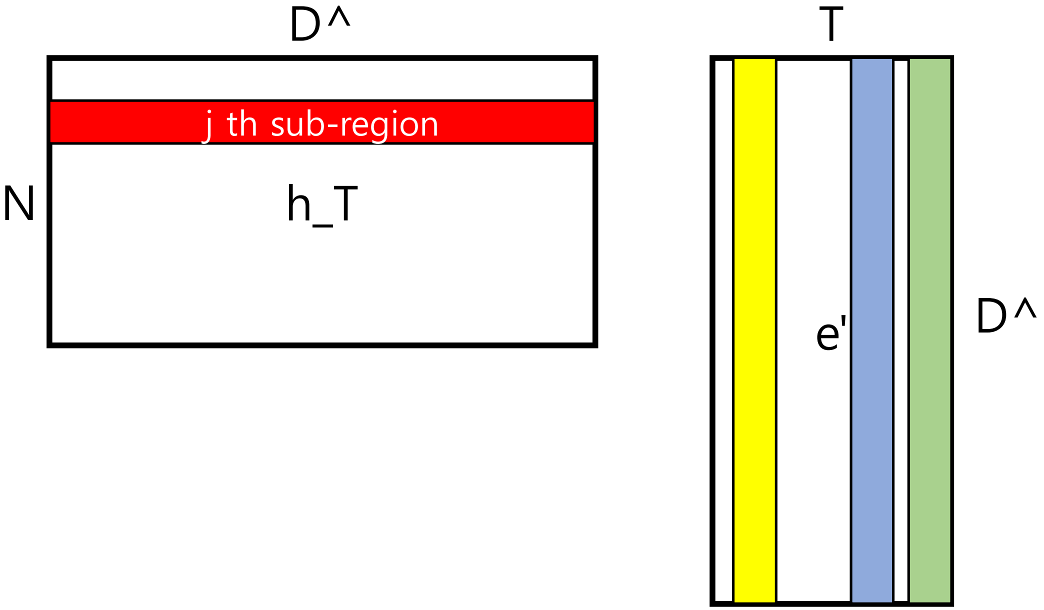
$$\boxed{\begin{aligned} h_0 &= F_0(z, F^{ca}(\overline{e})); \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1; \\ \hat{x}_i &= G_i(h_i). \end{aligned}}$$

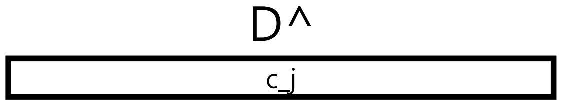
$\boxed{h \in \mathbb{R}^{\hat{D} \times N}}$

Each column of h is a feature vector of a sub-region of the image (N sub-region in the image)

$s'_{j,i} = h_j^T e'_i,$

weight the model attends to the i th word when generating the j th sub-region of the image





word-context vector

dynamic representation of word vectors relevant to h_j

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \quad \text{where } \beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})},$$

$F^{attn}(e, h)$

$e \in \mathbb{R}^{D \times T}$ $\quad U \in \mathbb{R}^{\hat{D} \times D}$ $\quad \boxed{e' = Ue,}$

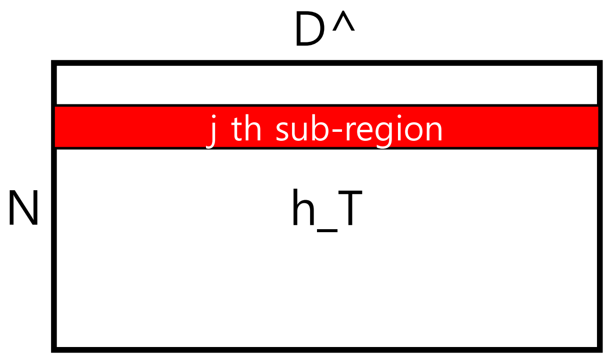
$$\begin{aligned} h_0 &= F_0(z, F^{ca}(\overline{e})); \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1; \\ \hat{x}_i &= G_i(h_i). \end{aligned}$$

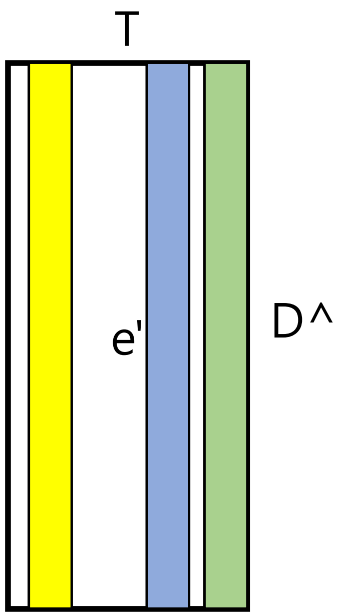
$\boxed{h \in \mathbb{R}^{\hat{D} \times N}}$

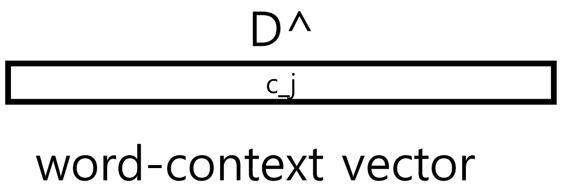
Each column of h is a feature vector of a sub-region of the image (N sub-region in the image)

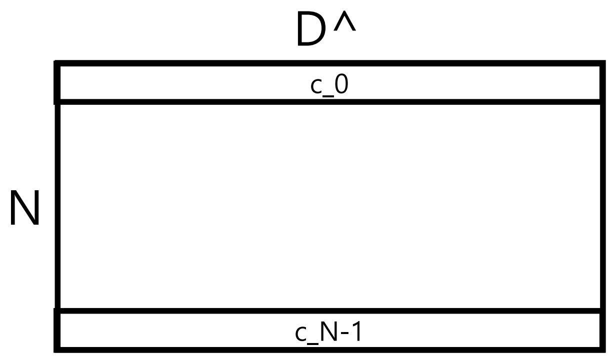
$s'_{j,i} = h_j^T e'_i,$

weight the model attends to the i th word when generating the j th sub-region of the image









$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \quad \text{where } \beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})},$$

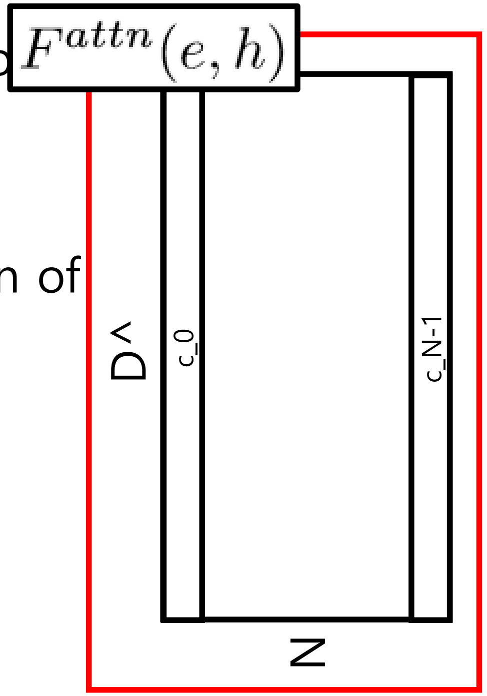
$F^{attn}(e, h)$

$e \in \mathbb{R}^{D \times T}$ $\quad U \in \mathbb{R}^{\hat{D} \times D}$ $\quad \boxed{e' = Ue,}$

$$\boxed{\begin{aligned} h_0 &= F_0(z, F^{ca}(\overline{e})); \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1; \\ \hat{x}_i &= G_i(h_i). \end{aligned}}$$

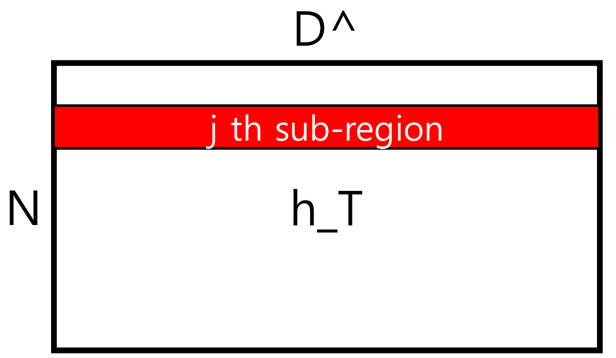
$\boxed{h \in \mathbb{R}^{\hat{D} \times N}}$

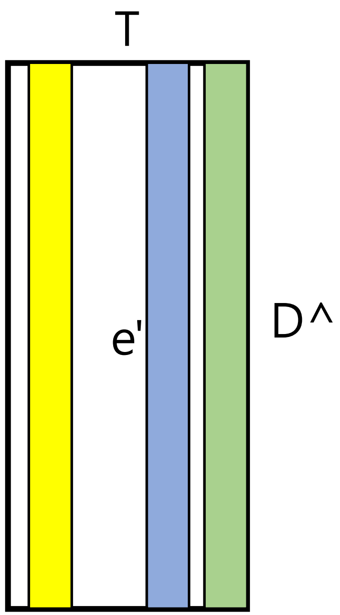
Each column of h is a feature vector of a sub-region of the image (N sub-regio

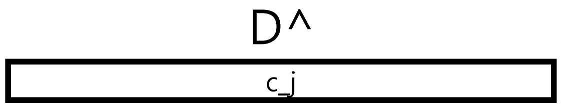


$s'_{j,i} = h_j^T e'_i,$

weight the model attends to the i th word when generating the j th sub-region of

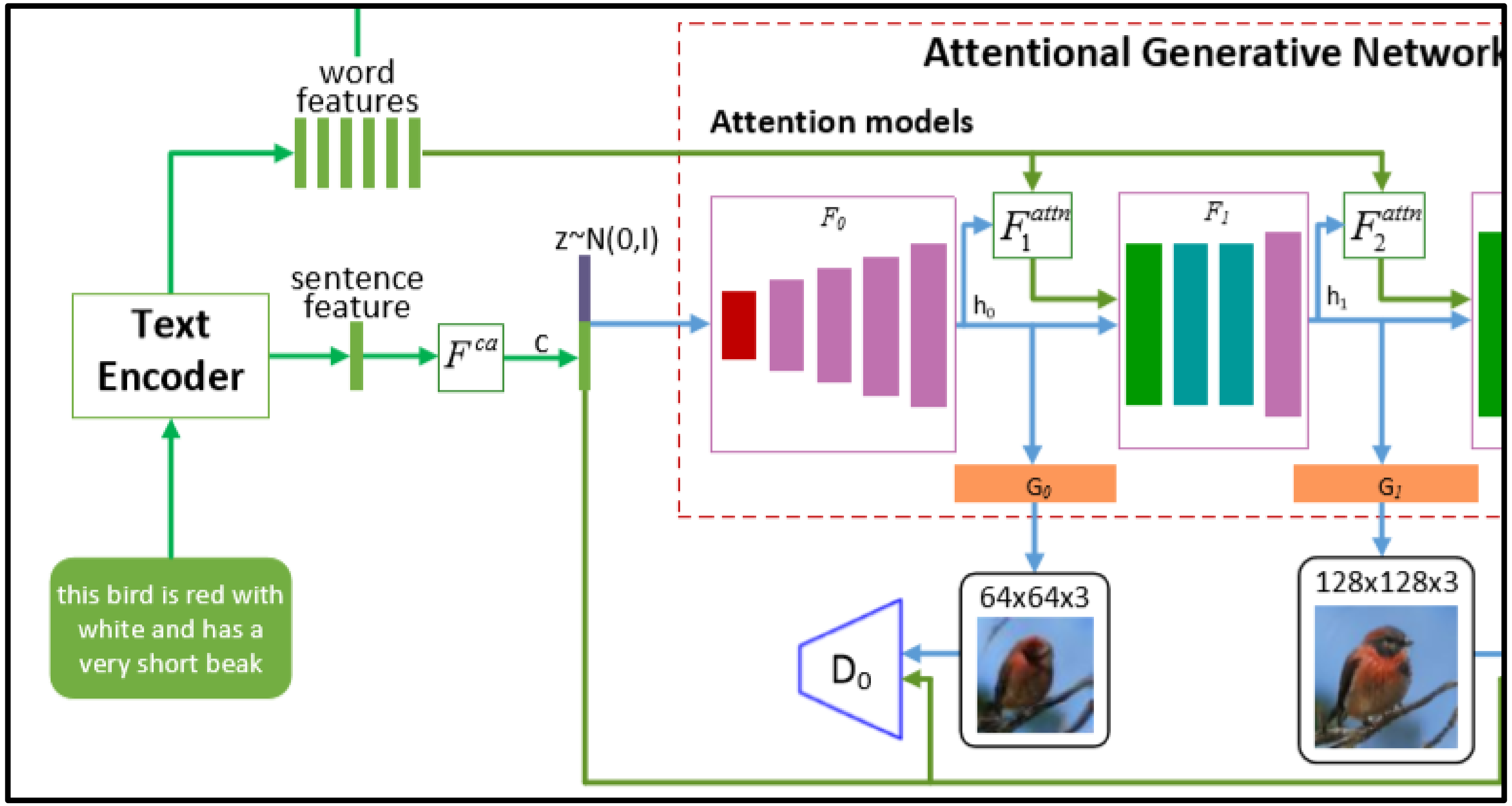


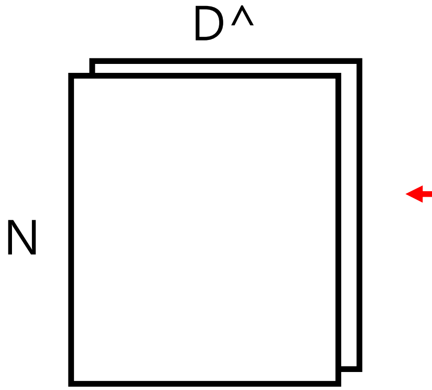




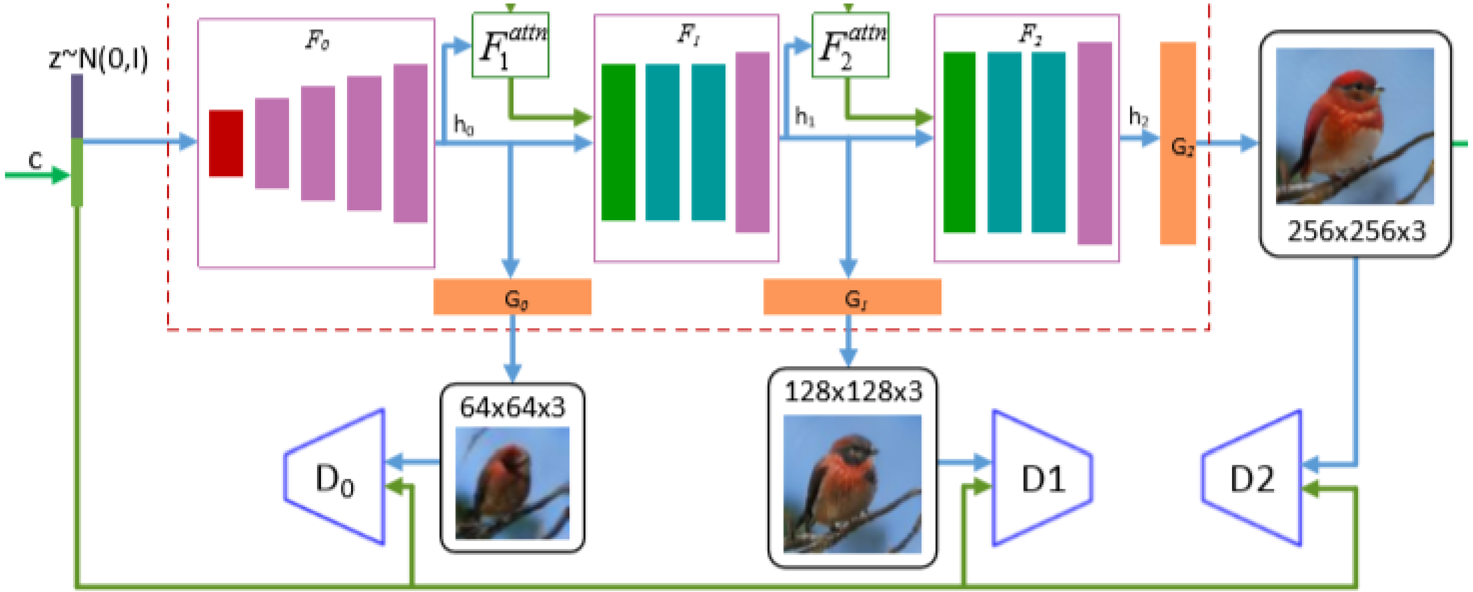
word-context vector

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \quad \text{where } \beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})},$$





$$h_0 = F_0(z, F^{ca}(\overline{e}));$$
$$h_i = F_i(\boxed{h_{i-1}, F_i^{attn}(e, h_{i-1})}) \text{ for } i = 1, 2, ..., m-1;$$
$$\hat{x}_i = G_i(h_i).$$



$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(D_i(\hat{x}_i)]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(D_i(\hat{x}_i, \overline{e})]}_{\text{conditional loss}},$$

$$\mathcal{L}_{D_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i)] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i)]}_{\text{unconditional loss}} + \underbrace{-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i, \overline{e})] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i, \overline{e})]}_{\text{conditional loss}},$$

구현 된 코드:
3 -> 512 -> 1

3 -> 512 ; D (repeat) -> 512+D -> 512 -> 1

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \quad \text{where } \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}.$$

Text Encoder

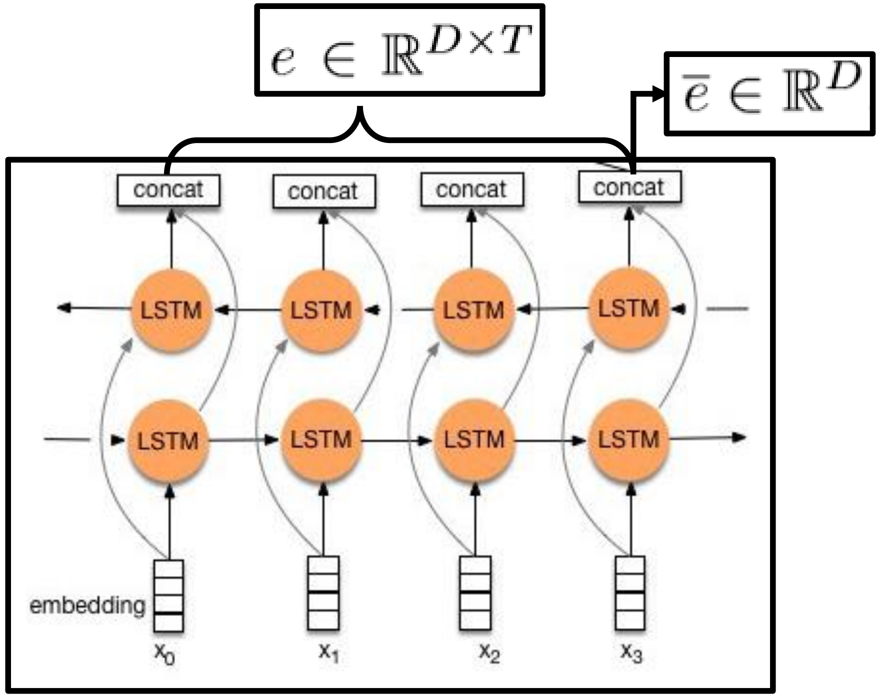
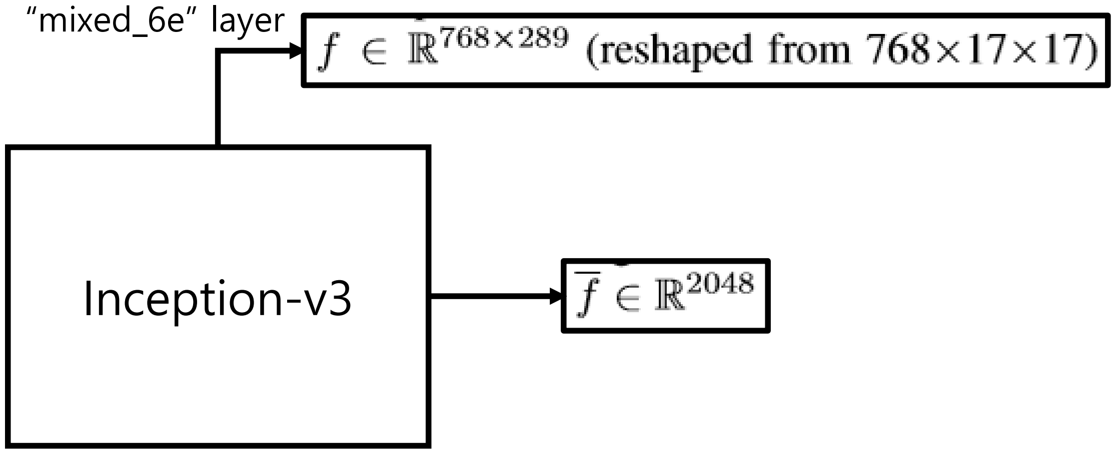


Image Encoder



$$v = Wf, \qquad \overline{v} = \overline{W}\,\overline{f},$$
$$v \in \mathbb{R}^{D \times 289}$$
$$\overline{v} \in \mathbb{R}^{D}$$

$s = e^T v,$

$s \in \mathbb{R}^{T \times 289}$

T개의 단어 순서 != 289개의 이미지 지역 순서

s_i,j is dot-product similarity between i th word of the sentence and the j th sub-region of the image

$$\bar{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1} \exp(s_{k,j})}.$$

다른 단어들에 비해, 해당 구역 (j)과 특정 단어 (i)의 similarity

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \quad \text{where } \alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{288} \exp(\gamma_1 \bar{s}_{i,k})}.$$

region-context vector: c_i
dynamic representation of the image’s sub-regions related to the i th word of the sentence

gamma1:
how much attention is paid to features of its relevant sub-regions
when computing region-context vector for a word

$R(c_i, e_i) = (c_i^T e_i)/(||c_i|| ||e_i||)$

the relevance between the i th word and the image using the cosine similarity

$$R(Q, D) = \log \Big(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i)) \Big)^{\frac{1}{\gamma_2}},$$

attention-driven image-text matching score between the entire image (Q) and the whole text description (D)

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma_3 R(Q_i, D_j))},$$

for a batch of image-sentence pairs $\{(Q_i, D_i)\}_{i=1}^{M}$
In this batch of sentences, only Di matches the image Qi ,
and treat all other M − 1 sentences as mismatching descriptions.

$$\mathcal{L}_1^w = -\sum_{i=1}^{M} \log P(D_i|Q_i),$$

the negative log posterior probability
that the images are matched with their corresponding text descriptions

$$\mathcal{L}_2^w = -\sum_{i=1}^{M} \log P(Q_i|D_i),$$

the negative log posterior probability
that the text descriptions are matched with their corresponding images

Text Encoder

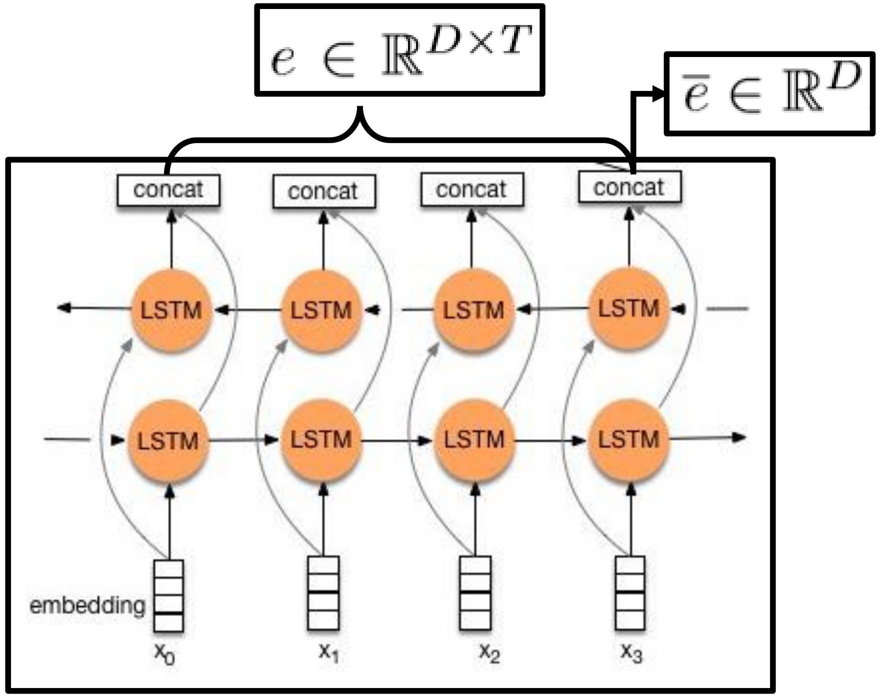
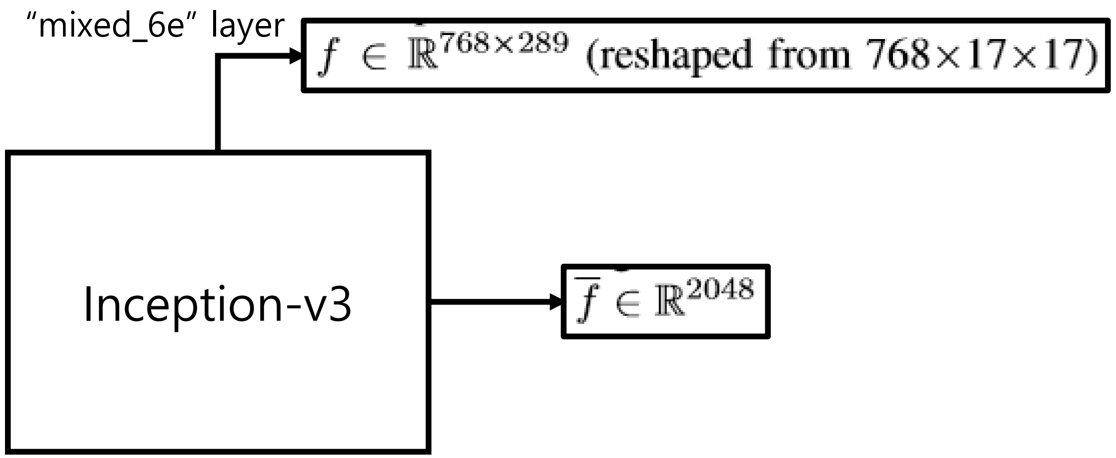


Image Encoder



$$v = Wf\,, \qquad \overline{v} = \overline{W}\,\overline{f},$$
$$v \in \mathbb{R}^{D\times 289}$$
$$\overline{v} \in \mathbb{R}^{D}$$

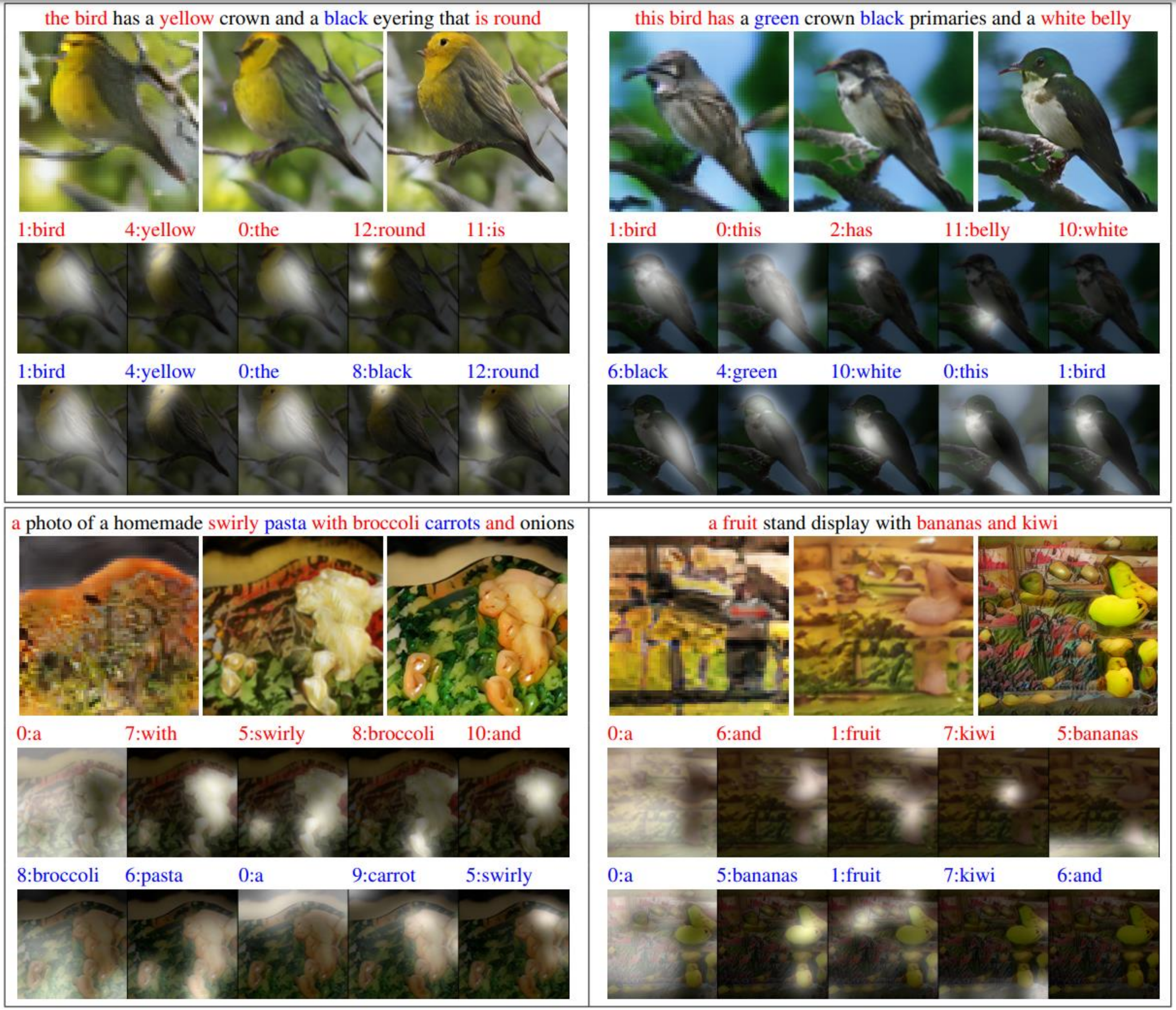
$$R(Q, D) = (\overline{v}^T \overline{e}) / (||\overline{v}|| ||\overline{e}||)$$ for sentence and entire image

$$\mathcal{L}_{DAMSM} = \mathcal{L}_1^w + \mathcal{L}_2^w + \mathcal{L}_1^s + \mathcal{L}_2^s.$$

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \quad \text{where } \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}.$$

$$\hat{\beta}_{j,i} = \begin{cases} \beta_{j,i}, & \text{if } \beta_{j,i} > 1/T, \\ 0, & \text{otherwise.} \end{cases}$$

Method	inception score	R-precision(%)
AttnGAN1, no DAMSM	3.98 ± .04	10.37± 5.88
AttnGAN1, $\lambda = 0.1$	4.19 ± .06	16.55± 4.83
AttnGAN1, $\lambda = 1$	4.35 ± .05	34.96± 4.02
AttnGAN1, $\lambda = 5$	4.35 ± .04	58.65± 5.41
AttnGAN1, $\lambda = 10$	4.29 ± .05	63.87± 4.85
AttnGAN2, $\lambda = 5$	**4.36 ± .03**	**67.82 ± 4.43**
AttnGAN2, $\lambda = 50$ (COCO)	**25.89 ± .47**	**85.47 ± 3.69**


the bird has a yellow crown and a black eyering that is round
1:bird 4:yellow 0:the 12:round 11:is
1:bird 4:yellow 0:the 8:black 12:round
this bird has a green crown black primaries and a white belly
1:bird 0:this 2:has 11:belly 10:white
6:black 4:green 10:white 0:this 1:bird
a photo of a homemade swirly pasta with broccoli carrots and onions
0:a 7:with 5:swirly 8:broccoli 10:and
8:broccoli 6:pasta 0:a 9:carrot 5:swirly
a fruit stand display with bananas and kiwi
0:a 6:and 1:fruit 7:kiwi 5:bananas
0:a 5:bananas 1:fruit 7:kiwi 6:and

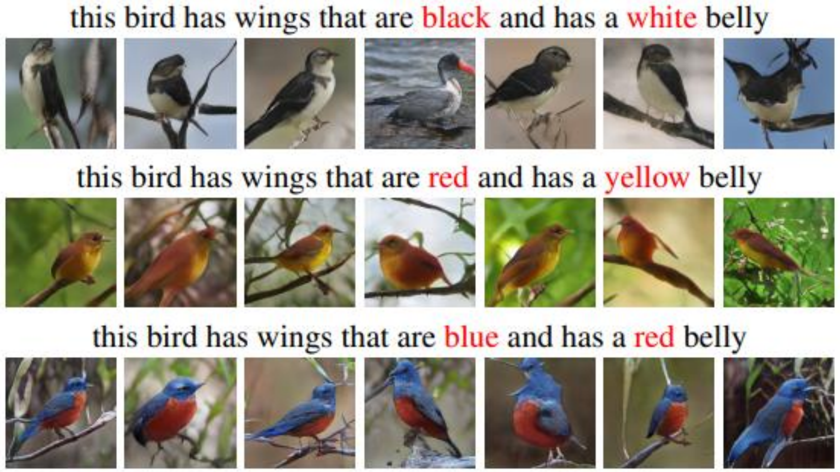


Figure 5. Example results of our AttnGAN model trained on CUB while changing some most attended words in the text descriptions.

Figure 6. 256×256 images generated from descriptions of novel scenarios using the AttnGAN model trained on COCO. (Intermediate results are given in the supplementary material.)

Figure 7. Novel images by our AttnGAN on the CUB test set.