# Learning Deep Representations of Fine-Grained Visual Descriptions

## Task

- **zero-shot classification:** Train classification model on images and its corresponding descriptions. But test on unseen images as well as its corresponding descriptions.
- **zero-shot text-based image retrieval:** Given a piece of description, retrieve an image that is best described by the given text.

## Framework

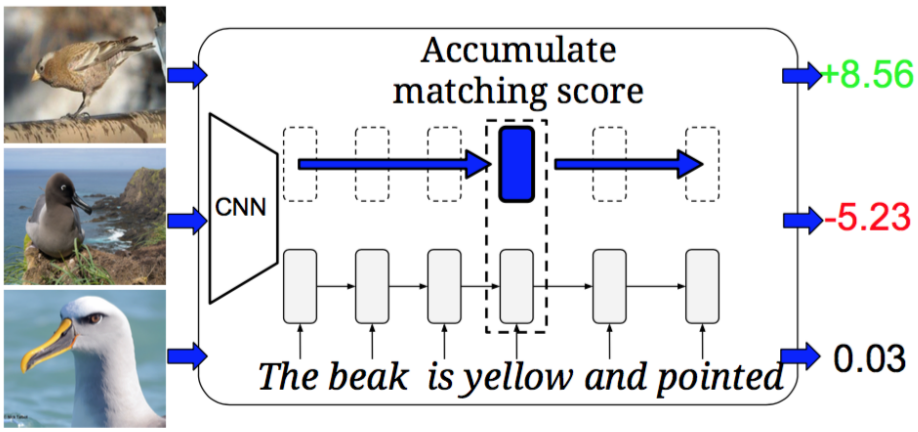


Figure 1: Our model learns a scoring function between images and text descriptions. A word-based LSTM is shown here, but we also evaluate several alternative models.

> **Main Idea:** Learn a compatibility function through zero-shot classification by optimizing a symmetric objective. For image retrieval, retrieve the image that has the largest compatibility score with the given text.

## Compatibility Function

- **Purpose:** Given a pair of image and a piece of description, the compatibility function outputs a score that measures the closeness between the image and description. If the description well described this image, their compatibility score should be high. While if the description and the image are unrelated, their compatibility score should be low.
- **Definition:** $F : V \times T \rightarrow R$ uses features from learnable encoder functions $\theta(v)$ for images and $\phi(t)$ for text: $F(v, t) = \theta(v)^t \phi(t)$. That means the compatibility score of an image and a piece of text is just the multiplication of the features of the image and the text.

# Image/Text Encoder

- **Image Encoder:** Pre-trained and fixed GoogleNet.
- **Text Encoder:** Stacked CNN and RNN as shown below

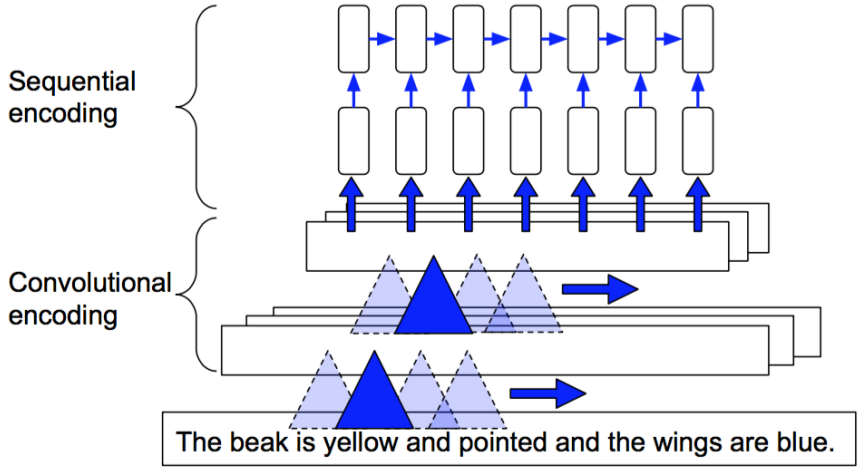


Figure 2: Our proposed convolutional-recurrent net.

# Objective

- Definition

  $$\frac{1}{N}\sum_{n=1}^{N}\Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n))$$

  In which $\Delta$ is the 0-1 loss, $v_n$ and $t_n$ are a pair of image and its corresponding description. $f_v(v_n)$ and $f_t(t_n)$ are image classifier and text classifier respectively. $y_n$ is the label of the given image.

  $f_v(v_n)$ and $f_t(t_n)$ are defined as:

  $$f_v(v) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}_{t\sim\mathcal{T}(y)}[F(v,t)]$$

  $$f_t(t) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}_{v\sim\mathcal{V}(y)}[F(v,t)]$$

  That means the image classifier is trying to classify the image by finding a label that text of this label has maximum compatibility score with the given image. And the text classifier is similar.

- Surrogation
  The objective function above is not continuous because of the 0-1 loss. Thus the author proposed a surrogate objective

$$\frac{1}{N}\sum_{n=1}^{N}\ell_v(v_n,t_n,y_n)+\ell_t(v_n,t_n,y_n)$$

where the misclassification losses are written as:

$$\ell_v(v_n,t_n,y_n)=\max_{y\in\mathcal{Y}}(0,\Delta(y_n,y)+\mathbb{E}_{t\sim\mathcal{T}(y)}[F(v_n,t)-F(v_n,t_n)])$$

$$\ell_t(v_n,t_n,y_n)=\max_{y\in\mathcal{Y}}(0,\Delta(y_n,y)+\mathbb{E}_{v\sim\mathcal{V}(y)}[F(v,t_n)-F(v_n,t_n)])$$

When the model correctly classifies an image, the loss is 0; otherwise the loss is `1 + (difference between target compatibility and predicted compatibility)`

# Reference

Reed, Scott, et al. “Learning deep representations of fine-grained visual descriptions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.