Paper Review #5

Matching Networks for One Shot Learning

Dept. of Computer Science & Engineering

201502755 Meeyun Kim

2021. 07. 23.

Contents

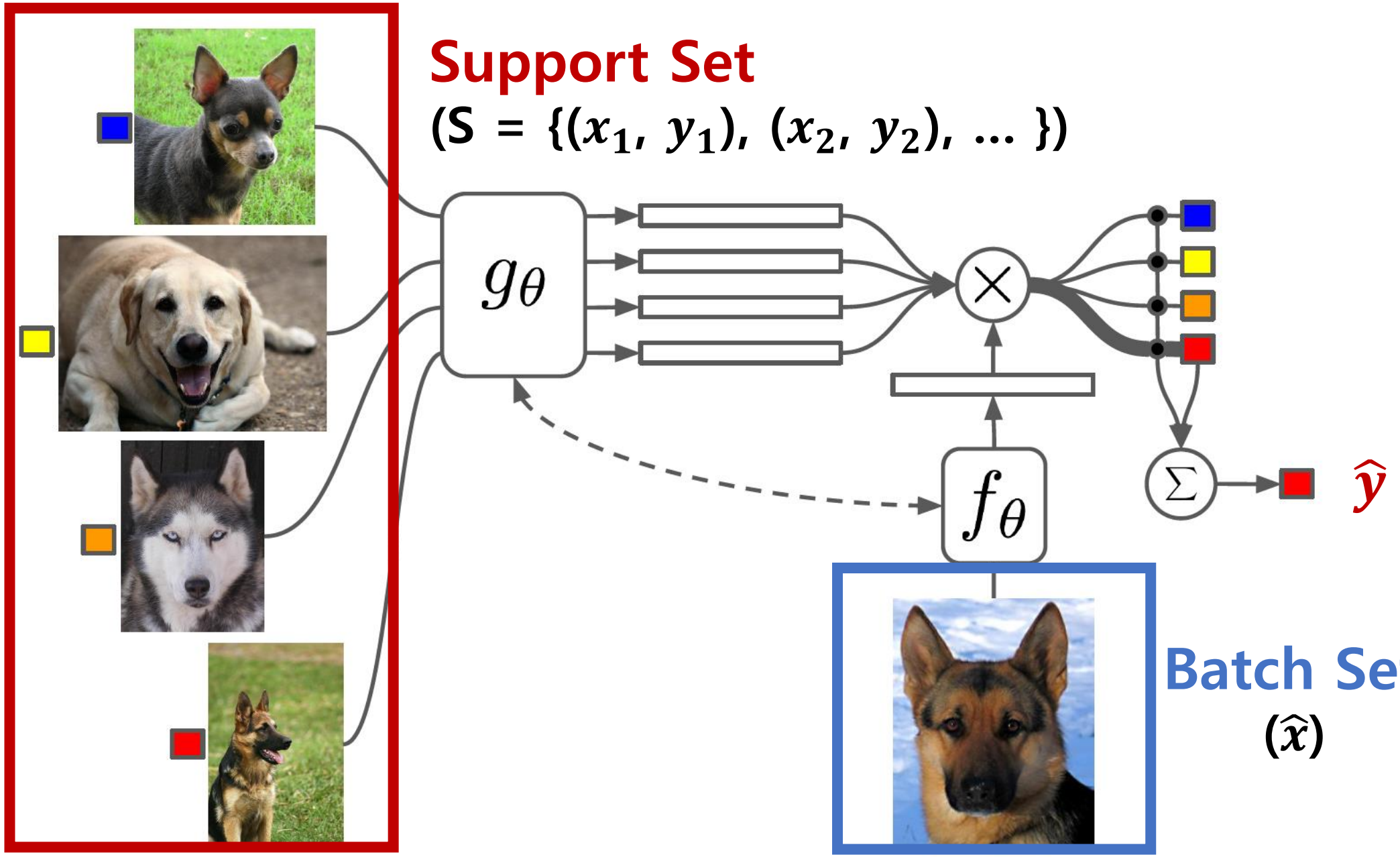
1. Introduction

Deep learning has made major advances in areas such as speech [7], vision [13] and language [16], but is notorious for requiring large datasets. Data augmentation and regularization techniques alleviate overfitting in low data regimes, but do not solve it. Furthermore, learning is still slow and based on large datasets, requiring many weight updates using stochastic gradient descent. This, in our view, is mostly due to the parametric aspect of the model, in which training examples need to be slowly learnt by the model into its parameters.

In contrast, many non-parametric models allow novel examples to be rapidly assimilated, whilst not suffering from catastrophic forgetting. Some models in this family (e.g., nearest neighbors) do not require any training but performance depends on the chosen metric [1]. Previous work on metric learning in non-parametric setups [18] has been influential on our model, and we aim to incorporate the best characteristics from both parametric and non-parametric models – namely, rapid acquisition of new examples while providing excellent generalisation from common examples.

2. Model Architecture

Matching Networks Architecture (One shot Learning)



Matching Network (Cont'd)

Support Set

$$\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{k}$$

Classifier

$$c_S(\hat{x}) \, , P(\hat{y}|\hat{x}, S)$$

Matching Network (Cont'd)

$$c_S(\hat{x}) \quad , \quad P(\hat{y}|\hat{x}, S) \Rightarrow \hat{y} = \sum_{i=1}^{k} a(\hat{x}, x_i) y_i$$

Attention Mechanism

$\hat{x}$: Batch sample

x_i : Support sample

The Attention Kernel

$$a(\hat{x}, x_i) = \frac{e^{c(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{k} e^{c(f(\hat{x}), g(x_j))}}$$

Embedding Function

Cosine function

Softmax Function

Full Context Embeddings

$$f(\hat{x}) \Rightarrow f(\hat{x}, S) \quad g(x_i) \Rightarrow g(x_i, S)$$

$$f(\hat{x}, S) = \text{attLSTM}(f'(\hat{x}), g(S), K)$$

3. Training Strategy

Episode Training

- **Training Set : Support Set + Batch Set**

$$\theta = \arg\max_{\theta} E_{L \sim T} \left[E_{S \sim L, B \sim L} \left[\sum_{(x,y) \in B} \log P_\theta \left(y | x, S \right) \right] \right]$$

L : Label set
T : distribution over possible label sets ***L***