Using the Forest to See the Trees Summary

Finn Hittson – fxh157

CSDS 464

Due: March 10th, 2023

Introduction: An issue with object recognition systems is that they do not use the context of the scene they are analyzing to help aid in the identification and recognition process. Object recognition algorithms commonly use object models compared with selected regions of the image to identity the contents of an image, but this approach ignores the silent dialogue that takes place between the entirety of the objects in the image and what they say about the entire scene. The paper "Using the Forest to See the Trees: Exploiting Context for Visual Object Detection and Localization" written by A. Torralba, K.P. Murphy, and W.T. Freeman proposes a new method that integrates more contextual information present in the scene of the image to help aid in the identification process.

Summary: The authors describe a three step model that uses the context of the scene to aid in object identification. The first step is computing the "gist" of the image. This way it is possible to develop a global representation of the entire scene without knowing the individual objects identities, that are present in the scene. The gist descriptor performs this computation, capturing the features of an image into individual vectors using the following equation:

$$g_k = \sum_{x,y} w_k(x,y) \times |I(x,y) \otimes h_k(x,y)|^2$$

$I(x,y)$ is the luminance of the image, $h_k(x,y)$ is a bank of multiscale-orientated Gabor filters, and $w_k(x,y)$ is the spatial window that computes the average output energy of each filter at different locations in the image. In simple terms, the gist descriptor essentially cuts down the image to its barebones content, only keeping the necessary details in order to identify the overarching idea of the image. It is similar, in a theoretical sense, to Bluetooth music compression; only the main, most prominent frequencies/features of the song are transmitted to the headphones to save power, knowing that the human ear will struggle to hear the difference between the compressed and uncompressed sound.

The next step is joint scene classification and object detection. For object detection there are two tasks that need to be solved: object presence detection and object localization. Object detection is done by estimating the probability of an object being present, $P^t = 1$, in an image I, $p(P^t = 1|I)$. Object localization is done by using a log-likelihood ratio. However, these techniques do not incorporate the global identity of the scene. The authors solve this problem in a two step approach. They first estimate the category or type of the scene, $p(S = s|g)$, and then they use this information to predict the number of objects present, $p(N^t|S = s)$. Classifying a scene is done using the Parzen-window based density estimator which is a non-parametric approach that estimates a probability density function at a specific point in a sample without any prior knowledge of the true underlying probability distribution. With the scene estimated the number of objects can be estimated using the following equation:

$$p(N^t = n|g) = \sum_s p(N^t = n|S = s)p(S = s|g)$$

$p(N^t = n|S = s)$ is estimated by counted.

Next is object localization using global image features. With the spatial layout captured from the gist, location priming predicts the vertical location of object classes before running any detectors. This is easily achieved using any nonlinear regression model. The authors specifically define the following equation to perform this operation.

$$p(Y^t|g) = \sum_{k=1}^{K} w_k(g)\mathcal{N}\left(Y^t\middle|\beta_k^t g, \sigma_k^2\right)$$

This model determines object vertical locations in the image. Y^t is the vertical location of the class t, K is the number of mixture components and $\mathcal{N}$ is the Gaussian distribution where β_k are the regression weights for mixture component k and σ_k^2 are the residual variances. $w_k(g)$ is the weight of k and is defined as the SoftMax.

The authors combine these steps first by initially ignoring the location information. First the confidence score from the logistic regressor used to determine localization is used as a local likelihood term to help build and fit a model that predicts the parameters of the Gaussian distribution. This is done by computing empirical means and variances of the logistic regressor scores when it's applied to areas in the image that contain and do not contain the object. Using this information with the gist then tells us which objects are present and which are not. Then using the object location method described above with the gist determines the number of object classes present. The authors close their methods briefly describing how to combine multiple types of objects in scene analysis (i.e., cars and pedestrians). They introduce latent variables to represent any number of object types and, assuming that they are conditionally independent, they perform inference for multiple object types in parallel to determine which are most highly prominent in the image.

They found that local priming based on the gist produced improbable object locations. They also found that the detector is able to produce low confidence scores even when the object is in the predicted region. Also, if the detector produces errors that are contextually correct, the integrated model will not be able to get rid of them. But even with these errors, this system is still able to outperform car detectors acting in isolation.

Conclusion: The system is not perfect. It sometimes produces objects out of context but there are methods to mediate this issue. The model also sometimes makes false positive errors in contextually plausible locations which can be seen as an equivalent to human's hallucinations. Ultimately, the proposed method is not perfect. This model is probabilistic information fusion which has potential other applications in speech recognition other language models, and is a steppingstone for developing the ideal model to perfectly perform scene analysis.