# Automatic Discovery, Association Estimation and Learning of Semantic Attributes for a Thousand Categories Supplementary Material

Ziad Al-Halah Rainer Stiefelhagen
Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
{ziad.al-halah, rainer.stiefelhagen}@kit.edu

This document provides additional information about the proposed approach.

**Network Architecture** Fig. 1 shows the detailed architecture of the deep neural networks used for the joint and attribute models. We use a deep network model similar to the one proposed by Krizhevsky *et al*. [2]. All the convolutional and hidden fully connected layers are followed with a batch normalization layer and a ReLU activation function. Note, this is not a Siamese architecture, the branched representation is just to highlight the common layers between the two models.

**Submodular Functions** A set function $\mathcal{F}$ is submodular if it satisfies the decreasing marginal gain condition [1]:

$$\mathcal{F}(A \cup \{s\}) - \mathcal{F}(A) \geq \mathcal{F}(B \cup \{s\}) - \mathcal{F}(B) \text{ for } A \subseteq B. \quad (1)$$

Our proposed objective function $\mathcal{F}(\cdot)$ is submodular if all terms $\mathcal{F}_{dis}$, $\mathcal{F}_{div}$ and $\mathcal{C}$ are submodular.

*Discrimination*: A formal proof of the submodularity of the entropy rate of a random walk on $\mathcal{G}$ can be found in [3]. Informally, according to the defined edge weights $g_{ij}(\cdot)$ in the constructed graph $\mathcal{G}$, adding a word $w$ to $S$ will result in increasing the edge weights between some node pairs in $\mathcal{G}$. However this increase will result in a lower increase in the uncertainty of the walk if it happens in a later stage, since then it will be shared with the contributions of other words in $S$ discriminating between the same pairs.

*Diversity*: Note that $s(w_i, T_k) \geq 0 \ \forall w_i$. Adding a new word $w$ that belongs to a new topic (the outer sum) will result in a higher gain than adding a word to an already existing topic in $S$ (the inner sum). This is due to the fact that the square root function is a monotonically increasing concave function and thus submodular [1]. $\mathcal{F}_{div}$ is sum of submodular functions and, therefore, it is submodular.

*Saliency*: Note that $p(T_k|w_i) \geq 0 \ \forall w_i$, therefore $\mathcal{C}(S) > 0$ and is monotonically increasing. Adding a new word $w$ to $S$ in $\mathcal{C}$ will always increase the cost by the same margin, hence $\mathcal{C}$ is submodular.

**Data Setup** To collect the encyclopedia articles, we coded a simple tool to automatically retrieve the articles from Wikipedia. For each synset in the ImageNet data set, the tool queries the Wikipedia API using the different words in the synset to get articles with the same title. We were able to collect about 89% of the articles automatically using the previous method. The rest of the synsets either did not have an exact matching article title in Wikipedia or there were ambiguities in the retrieved articles since multiple ones matched the query. Articles for these synsets were then acquired interactively.

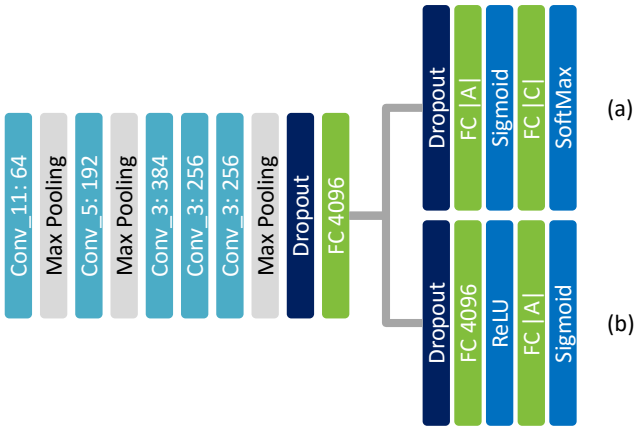


Figure 1: The detailed network architecture used for (a) the joint and (b) attribute models.

## References

[1] S. Fujishige. *Submodular functions and optimization*, volume 58. Elsevier Science, 2005. 1

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 1

[3] M.-y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy-Rate Clustering: Cluster Analysis via Maximizing a Submodular Function Subject to a Matroid Constraint. *PAMI*, 36, 2014. 1