Temporal Attention Mechanism with Conditional Inference for Large-Scale Multi-Label Video Classification

Eun-Sol Kim[1], Kyoung-Woon On[2], Jongseok Kim[1], Yu-Jung Heo[2], Seong-Ho Choi[2], Hyun-Dong Lee[2] and Byoung-Tak Zhang[2]

[1] Kakao Brain, 13494, Seongnam, South Korea
{epsilon, ozmig}@kakaobrain.com
[2] Seoul National University, 08826, Seoul, South Korea
{kwon, yjheo, shchoi, hdlee, btzhang}@bi.snu.ac.kr

Abstract. Here we show neural network based methods, which combine multimodal sequential inputs effectively and classify the inputs into multiple categories. Two key ideas are 1) to select informative frames among a sequence using attention mechanism and 2) to utilize correlation information between labels to solve multi-label classification problems. The attention mechanism is used in both modality (spatio) and sequential (temporal) dimensions to ignore noisy and meaningless frames. Furthermore, to tackle fundamental problems induced by independently predicting each label in conventional multi-label classification methods, the proposed method considers the dependencies among the labels by decomposing joint probability of labels into conditional terms. From the experimental results (5th in the Kaggle competition), we discuss how the suggested methods operate in the YouTube-8M Classification Task, what insights they have, and why they succeed or fail.

Keywords: Multimodal Sequential Learning, Attention, Multi-label classification, Video understanding

1 Introduction

We focus on finding neural network based methods capable of learning large-scale multimodal sequential data, which are videos collected from YouTube, and classifying the data into multiple categories. To tackle this challenging goal, we postulate three subproblems as follows: 1) combining multimodal inputs effectively, 2) modeling temporal inputs, and 3) using correlation information between labels to resolve multi-label classification problem. Specifically, only two modalities, e.g., image and audio, are considered as the multimodal inputs in this work.

In this work we make the following two contributions. First, we explore spatio-temporal aggregation of visual and auditory features by designing new gate modules. Compared to existing methods for learning spatio-temporal inputs, such as NetVLAD[2], GRU[6] and LSTM[10], the suggested method can

find different importance weight between the temporally neighboring frames. Second, we use correlation information between labels to resolve multi-label classification problem. While the simple binary relevance (BR) method approaches this problem by treating multiple targets independently, the suggested method focuses on exploiting the underlying label structure or inherent relationships.

We evaluate our method on the YoutTube-8M dataset containing about 6.1M videos and 3862 labels. The proposed method shows significant performance improvement over the baseline models, and finally our ensemble model is ranked 5th out of about 400 teams in the 2nd YouTube-8M Video Understanding Challenge [3].

The remainder of the paper is organized as follows. In the next section, we summarize previous research, including papers from the 1st YouTube-8M workshop related to multimodal, sequential learning and multi-label classification. Then, suggested methods and modules are shown in section 3. In section 4, YouTube-8M dataset is described and the experimental results are shown. In section 5, we show the ensemble model submitted to the Kaggle competition. Finally, we conclude with a discussion about why the methods are successful or not.

2 Related Work

We summarize previous research related to this work in terms of the following topics: multimodal learning, temporal aggregation and large-scale multi-label classification.

2.1 Multimodal Learning

Multimodal learning has been widely used to define representations of multimodal inputs to project unimodal features together into a multimodal space. The simplest method is concatenation of individual unimodal features (Figure 1(a)). As neural networks has become a popular method for learning unimodal features, it has been considered more popular to concatenate the unimodal features learned from each neural network(Figure 1(b)). Instead of naive concatenation, each unimodal feature from neural networks projects into a joint representation space with additional networks (Figure 1(c)).

For the Kaggle competition, preprocessed visual and audio features for each frame are distributed to participants. Visual features are extracted using Inception-V3 image annotation model[20] and audio features are extracted using a VGG-inspired acoustic model[9].

In the last YouTube-8M competition, almost all of the participants tried to concatenate these visual and audio feature vectors via either 1)early fusion or 2)late fusion. The early fusion method concatenates two feature vectors before being fed into a frame level model which deals with both modalities. On the

[3] https://www.kaggle.com/c/youtube8m-2018/leaderboard

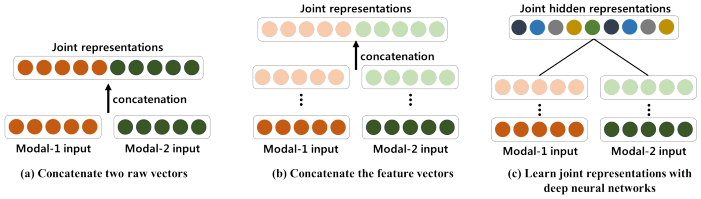

Fig. 1. Multimodal learning with joint representations

other hand, late fusion means that visual and audio features are concatenated after having been processed by two frame level models which deal with each modality. Na et al[14] tried to learn multimodal joint representation using multimodal compact bilinear pooling[8]. However, they reported that their newly joint features performed significantly worse than simple feature concatenation.

2.2 Temporal Aggregation

In terms of neural network architectures, many problems with sequential inputs are resolved by using Recurrent Neural Networks (RNNs) and their variants as it naturally takes sequential inputs frame by frame. However, as RNN-based methods take frames in (incremental) order, the parameters of the methods are trained to capture patterns in transitions between successive frames, making it hard to find long-term temporal dependencies through overall frames. For this reason, their variants, such as Long Short-Term Memory (LSTM, [10]) and Gated Recurrent Units (GRU, [6, 4]), have made the suggestion of ignoring noisy (unnecessary) frames and maintaining the semantic flow by turning switches on and off.

Recently, a number of researches shed new light on Bag-of-Visual-Words (BoVW) techniques[16, 19] in order to construct a set of visual descriptors from image data, such as VLAD[3] and DBoF methods[1]. BoVW-based methods have been expanded to the temporal domain, that is, the visual descriptors are extracted from not only an image, but from a sequence of images[2]. After constructing a set of spatio-temporal visual descriptors, a representative vector of a sequence is constructed by applying pooling methods over the set (averaging operations over the descriptors).

2.3 Multi-Label Classification

Multi-label classification is a supervised learning problem where each instance has two or more labels. It is more challenging than single-label classification since combinations of labels grow exponentially.

The most common approach to multi-label classification is Binary Relevance(BR), which decomposes the multi-label learning task into a number of independent binary learning tasks. This approach can reduce the search space

from $O(2^n)$ as combinations of labels to $O(n)$ as the number of labels n. However, this decomposition makes BR models incapable of exploiting dependencies and correlations between labels.

Classifier Chain (CC) overcomes such disadvantages of basic BR models by passing label information between each BR classifier along a chain[17]. CC treats multi-label classification as a sequential prediction problem, which resembles following a single path in a binary tree in a greedy manner. Probabilistic Classifier Chains (PCC) is an extension of CC and probability theory. PCC estimates the entire joint distribution of the labels and constructs a perfect binary tree required to find the optimal path[7]. Nam et al[15] applied Recurrent Neural Networks (RNNs) to model the sequential prediction problem. The key idea of the approach is to model the joint probability of positive labels, not the entire joint distribution.

3 The Model

In this section, several methods used for the YouTube-8M competition are introduced. Basically, we tried to find better representations of the multimodal inputs using attention mechanisms, which can capture the correlations between modalities. Furthermore, we suggest a new multi-label classification method that reflects our investigation of the statistics of the label set.

3.1 Multimodal Representation Learning with Attention

Here, we show three multimodal representation learning methods. Before feeding visual vectors $\mathbf{x}_v$ and audio vectors $\mathbf{x}_a$ into temporal aggregation methods, a new vector $\mathbf{x}_f$ is learned using the following methods.

1. Element-wise summation after a linear transformation

$$\mathbf{x}_{a_{exp}} = \mathbf{W}_{va}\mathbf{x}_a + \mathbf{b}_{va} \tag{1}$$

$$\mathbf{x}_f = \mathbf{x}_v + \mathbf{x}_{a_{exp}} \tag{2}$$

2. Temporal attention on $\mathbf{x}_a$ guided by $\mathbf{x}_v$

$$\mathbf{x}_f = \mathbf{x}_v + softmax(\mathbf{x}_v^\top \mathbf{W}_a^{att} \mathbf{X}_{a_{exp}}^{t-\frac{w}{2}:t+\frac{w}{2}})\mathbf{X}_{a_{exp}}^{t-\frac{w}{2}:t+\frac{w}{2}} \tag{3}$$

3. Temporal attention on $\mathbf{x}_v$ guided by $\mathbf{x}_a$

$$\mathbf{x}_f = \mathbf{x}_{a_{exp}} + softmax(\mathbf{x}_{a_{exp}}^\top \mathbf{W}_v^{att} \mathbf{X}_v^{t-\frac{w}{2}:t+\frac{w}{2}})\mathbf{X}_v^{t-\frac{w}{2}:t+\frac{w}{2}} \tag{4}$$

Method 1 is a simple element-wise summation. Since $\mathbf{x}_v$ and $\mathbf{x}_a$ have different feature vector sizes, a linear transformation is applied to $\mathbf{x}_a$ to match the size.

With method 2, temporal correlations between a visual vector $\mathbf{x}_v$ and neighboring w audio inputs $\mathbf{X}_a^{t-\frac{w}{2}:t+\frac{w}{2}}$ are trained by learning an attention matrix W_a^{att}. By using the temporal attention methods, the latter aggregation methods

can focus on a subset of sequential inputs which are relevant to each other and ignore irrelevant and noisy parts of the input sequence. Furthermore, the temporal attention method, which gives different importance weights to the temporally neighboring audio inputs, summarizes the audio inputs based on the weights and assigns a new vector to the corresponding vector, can be interpreted as an aligning method. Although the distributed dataset is already aligned, the sequences of each modality may involve different semantic streams. Applying temporal attention to those sequences can be helpful in resolving the disentanglement in the semantic flows, as it could give a chance to be matched with the neighboring frames.

Similarly, temporal correlations between an audio vector and neighboring visual vectors are trained with method 3.

These three methods are summarized in Figure 2.

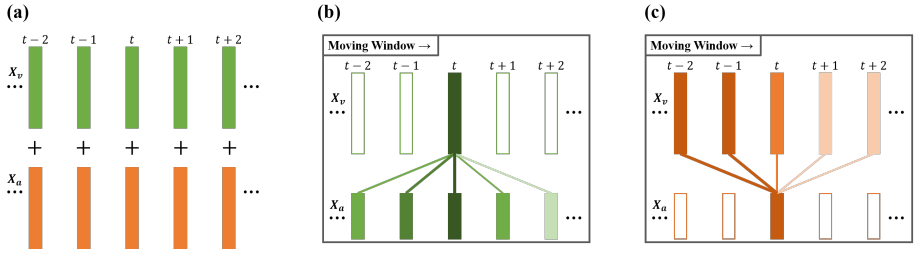

Fig. 2. (a): Element-wize summation after a linear transformation (b): Image guided attention mechanism (c): Audio guided attention mechanism

3.2 Conditional Inference using Label Dependency for Multi-Label Classification

The objective of the multi-label classification is to maximize likelihood of conditional probability $p(\mathbf{y}|x)$ where $x \in \mathbf{X}$ and $\mathbf{y} \in \{y_1, y_2, \ldots, y_q\}$ with $y_i \in \{0, 1\}$:

$$\mathcal{L}(\theta; \mathbf{y}|x) = \prod_{x \in \mathbf{X}} p(y_1, y_2, ..., y_q|x; \theta) \tag{5}$$

As discussed in section 2.3, The BR method simply hypothesizes that the probabilities of each labels are independent given x:

$$p(\mathbf{y}|x) = \prod_{i=1}^{q} p(y_i|x) \tag{6}$$

The BR method is simple and shows a reasonable performance, but it cannot reflect correlation between labels due to its independence assumption. To avoid

losing information of dependencies between labels, the joint probability can be factorized and obtained in a chaining manner.

$$p(\mathbf{y}|x) = \prod_{i=1}^{q} p(y_i|x, y_{<i}) \tag{7}$$

Most of the chaining approaches model the chaining property via building q-classifier for each term of RHS in equation 7 [7, 15, 18]. More specifically, the function f_i is learned on an augmented input space $\mathbf{X} \times \{0,1\}^{i-1}$ which is taking $y_{<i}$ as additional attributes to determine the probability of y_i. Then the $p(\mathbf{y}|x)$ can be obtained as follows:

$$p(\mathbf{y}|x) = \prod_{i=1}^{q} f_i(x, y_{<i}) \tag{8}$$

However, to estimate the above probability, 2^q-combinations of labels need to be searched or specific order of labels must be pre-defined. Instead, we learn a single function f to map from a given x and an additional label information l to y ($f : \mathbf{X} \times L \rightarrow \mathbf{y}$), where the l is a vector $\{0,1\}^q$ and represents previously observed labels with 1 values.

In detail, at first, conditional probabilities over all labels $\mathbf{y}$ given x are predicted by function f, and then a label which is the most probable to 1 is chosen as a first observed label. Next, given the same x and previously predicted labels l, conditional probabilities are again predicted and the second observed label is chosen in a same manner. This procedure is iteratively performed and the number of iterative step is selected based on empirical performances. Figure 3(a) illustrates the mechanism with five labels and two iterative steps.

For function f, the neural network architecture is designed to capture the dependencies among x, observed y, and predicted y. It provides a richer representation with low-rank bilinear pooling [11] followed by context gate mechanism [13] which is shown in Figure 3(b).

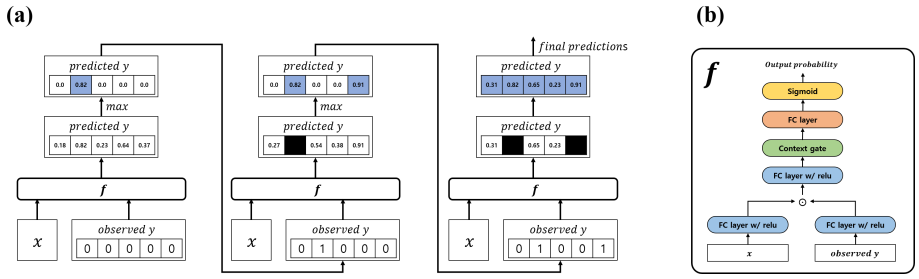

Fig. 3. (a): An illustration of the conditional inference procedure on 5-labels and 2-steps situation. (b): Core neural network architecture of conditional inference.

4 Experiments

4.1 Youtube-8M Dataset

The YouTube-8M dataset consists of 6.1M video clips collected from YouTube. The average length of the clips is 230.2 seconds and the maximum/minimum lengths are 303, 1 seconds respectively (statistics of the 3.9M training clips). From each clip, image sequences and audio signals are extracted. Visual features are extracted using Inception-V3 image annotation model[20] and audio features are extracted using a VGG-inspired acoustic model[9]. After preprocessing steps including PCA-ed and quantization, a 1024-dimensional image vector and a 128-dimensional audio vector are obtained for every second.

Each clip of the dataset is annotated with multiple labels. The average number of labels annotated for a clip is 3.0, and the maximum and the minimum are 23 and 1 respectively, out of 3862 possible labels. In the YouFurthermore, the number of examples per label is not uniformly distributed. As a specific example, 788,288 clips are annotated with GAME, while only 123 clips are annotated with Cylinder. More than half of the total labels (2086 of 3862 labels) contain less than 500 clips.

4.2 Training Details

Adam optimizer[12] with two parameters, i.e., a learning rate of 0.001 and a learning rate decay of 0.95, is utilized to train models. We also find it helpful to set the gradient clipping value to 5.0 for Bi-directional LSTM models and to 1.0 for NetVLAD models.

4.3 Experimental Results

Effects of Spatio-Temporal Attention First of all, the effectiveness of the suggested attention methods in Section 3.1 is verified. The quantitative results are summarized in Table 1. After applying the temporal attention methods to the original inputs, it is fed into Bi-directional LSTM(BLSTM) models with one layer and a cell per layer. Each output of the LSTM steps undergoes average pooling.

The table shows that models that selectively combine the features with attention values perform better than a naive BLSTM model. It is interesting to note that giving an attention to current audio features is not helpful. It may be possible that the label set of the YouTube-8M dataset is constructed to classify the video with "visual cues" rather than "auditory cues", meaning that the audio features may contain irrelevant information to predict the labels.

Effects of Conditional Inference To evaluate the effect of conditional inference mechanism for multi-label classification, comparative experiments are conducted with baseline models using video-level features. As shown in Table

Table 1. Validation Accuracy with Various Attention Methods with BLSTM

Attention Method	Window Size w	Accuracy (GAP)
None	None	0.858
Image Guided Attention	5	0.86071
Image Guided Attention	9	0.86078
Image Guided Attention	13	0.85920
Image Guided Attention	all	0.86129
Audio Guided Attention	5	0.85670

2, the proposed mechanism outperforms other variant baseline models. In addition, the GAP score increases as the number of steps increases, and it begins to decrease after the fourth step. It can be interpreted as the number of step hyper-parameter can be derived by average number of labels in a instance.

Table 2. Experimental results of conditional inference modules with video-level features

Attention Method	Accuracy (GAP)
Logistic model	0.7942
Mixture of expert (# of expert: 2)	0.8282
Mixture of expert (# of expert: 3)	0.8296
Mixture of expert (# of expert: 4)	0.8305
Mixture of expert (# of expert: 6)	0.8324
Conditional Inference(# of steps: 1)	0.8385
Conditional Inference(# of steps: 2)	0.8398
Conditional Inference(# of steps: 3)	0.8407
Conditional Inference(# of steps: 4)	0.8410
Conditional Inference(# of steps: 5)	0.8403

5 The Final Ensemble Model

Unfortunately, it was hard to find the optimal combination of the suggested methods described in Section 3. In this section, the final model that ranked 5th in the final leaderboard of the Kaggle competition is described, which may not be directly related to the methods in Section 3.

Based on the three criteria (Figure 4), we designed basic modules. As basic modules for temporal aggregation, vanilla RNN, GRU, LSTM, BLSTM, hierarchical RNN[5] and NetVLAD are tested with various methods on multimodal learning and MLC methods suggested in Section 3. Various number of layers,

Criteria	Methods
Multimodal Inputs	Concatenate, Element-wise summation, Attention, Differential Features, Bilinear Pooling
Temporal Aggregation	LSTM, GRU, Bidirectional LSTM, Hierarchical RNN, NetVLAD, CBHG
Classification Modules	Logistic Regression, Mixture of Experts, Class Chaining, Conditional Inference
Additional Modules	Layer Normalization, Skip Connection, Dropout, Gradient Clipping

Fig. 4. Various methods with three criteria which are postulated to solve this competition and additional options for the methods.

hidden states, and well-known techniques such as dropout, zoneout and skip-connection are tested with the temporal aggregation models. Among more than 100 experimental results with the possible combinations of those techniques, six of the experiments were selected for the final ensemble model by using a beam search method with a validation dataset.

The final six models selected for the final ensemble model are as follows:

1. MC-BLSTM-MoE2
2. MA-BLSTM-MoE2
3. MC-BLSTM-CG-MoE2
4. MC-NetVLAD-diff-C64-MoE4
5. MS-NetVLAD-C64-MoE4
6. MC-NetVLAD-C128-MoE4

where *MC, MA, MS* represent the methods to construct multimodal representation. *MC* represents an early fusion with a concatenation, *MA* and *MS* represent method 3 and 2 in section 3.1 respectively. For the attention methods, we set the window size to 5 based on the empirical performance. *CG* represents the context gating method[13], *C* stands for cluster size, and *MoE* is the number of experts.

diff means that the differential feature is concatenated. As the NetVLAD model could lose the temporal relationship between successive frames, the differences $\mathbf{x}^t_{diff}$ between the frames are concatenated to original inputs.

$$\mathbf{x}^t_{diff} = [\mathbf{x}^t : \frac{2 \times \mathbf{x}^t - \mathbf{x}^{t-1} - \mathbf{x}^{t+1}}{2}] \tag{9}$$

The logits of the six models are combined with different weight values(ensemble weights), which are learned by a single-layer neural network, and the exact values are 0.21867326, 0.22206327, 0.13936463, 0.16840834, 0.14120385, and 0.11028661.

The final test accuracy of the ensemble model is 0.88527, which is ranked at 5th in the final leader board.

We should note that there was a strict constraint on the final model size with 1GB, so the models are searched with this constraints, and the sizes of the selected models are 162M, 163M,168M, 138M, 136M, and 200M.

6 Conclusion

Even though the suggested methods in Section 3 could not be selected for the final ensemble model, we think that the newly suggested attention and MLC method might be helpful to improve the performance if we can find more suitable model architectures with more intensive exploring. From the competition point of view, we observe that the ensemble method dramatically improves the performance. Performances of NetVLAD models alone were not better than those of BLSTM models. But the ensemble of NetVLAD and BLSTM outperformed the ensemble of BLSTM models alone.

Acknowledgement

This work was partly supported by the Institute for Information & Communications Technology Promotion (R0126-16-1072-SW.StarLab, 2017-0-01772-VTT , 2018-0-00622-RMI) and Korea Evaluation Institute of Industrial Technology (10060086-RISF) grant funded by the Korea government (MSIP, DAPA).

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
3. Arandjelovic, R., Zisserman, A.: All about vlad. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1578–1585 (2013)
4. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
5. Chung, J., Ahn, S., Bengio, Y.: Hierarchical multiscale recurrent neural networks. arXiv preprint arXiv:1609.01704 (2016)
6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
7. Dembczynski, K., Cheng, W., Hüllermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: ICML. vol. 10, pp. 279–286 (2010)
8. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
9. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R., Wilson, K.: Cnn architectures for large-scale audio classification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017), `https://arxiv.org/abs/1609.09430`
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
11. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classification. arXiv preprint arXiv:1706.06905 (2017)
14. Na, S., Yu, Y., Lee, S., Kim, J., Kim, G.: Encoding video and label priors for multi-label video classification on youtube-8m dataset. arXiv preprint arXiv:1706.07960 (2017)
15. Nam, J., Mencía, E.L., Kim, H.J., Fürnkranz, J.: Maximizing subset accuracy with recurrent neural networks in multi-label classification. In: Advances in Neural Information Processing Systems. pp. 5413–5423 (2017)
16. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. pp. 1–8. IEEE (2007)
17. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 254–269. Springer (2009)
18. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333 (2011)

19. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: null. p. 1470. IEEE (2003)
20. Tensorflow: Tensorflow: Image recognition. /https://www.tensorflow.org/tutorials/images/image_recognition