Bag-of-words (BoW) models are widely used in the field of computer vision and find application in texture and object classification, object discovery, and action recognition. In the traditional BoW model, kMeans or a Gaussian mixture model is used to generate a visual vocabulary based on which a histogram of visual words is computed. This histogram is then used for classification, e.g. in combination with a support vector machine (SVM). In our paper, we show an equivalent formulation of BoW as a recurrent neural network that allows to train the visual words discriminatively and directly on video level rather than on frame level only. We apply our method to four action recognition benchmarks as well as two small image recognition datasets and show its superiority over the classical model and some sparse coding methods such as LLC [3] and ScSPM [4].

In the classical approach, a set \( \mathbf{x} \) of \( T \) feature vectors is quantized into a histogram of \( K \) visual words \( v_1, \ldots, v_K \),

\[
H(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} h(x_t), \quad h(x_t) = \begin{pmatrix} p(v_1|x_t) \\ \vdots \\ p(v_K|x_t) \end{pmatrix},
\]

where \( p(v_k|x_t) = 1 \) if \( v_k \) is the visual word being closest to \( x_t \) and 0 otherwise. As an alternative to this hard assignment, soft assignment is frequently used, i.e. \( p(v_k|x_t) \) is defined as a posterior distribution of a Gaussian mixture model or the kMeans model. Utilizing that kMeans is a special case of Gaussian mixture models, where the visual word prior is uniform and all variances are set to 1, we formulate an equivalent representation as a single-layer neural network with softmax output,

\[
p_{\text{NN}}(v_k|x) := \text{softmax}_K(W^T x + b),
\]

where \( W \in \mathbb{R}^{D \times K} \) is a weight matrix and \( b \in \mathbb{R}^K \) the bias. Setting these to

\[
W = (v_1 \ldots v_K), \quad b = -\frac{1}{2}(v_1^T v_1 \ldots v_K^T v_K)^T
\]

results in a neural network that generates the same posterior distribution as kMeans with soft assignment.

In order to realize the summation over all \( x_t \) that is required in Equation (1), we add a recurrent layer with the unit matrix as weights to the network. The normalization over the sequence length is realized by the activation function

\[
\sigma_t(z) = \begin{cases} 
    z & \text{if } t < T, \\
    z & \text{if } t = T.
\end{cases}
\]

An additional softmax layer is finally added to model a posterior distribution of the action classes given the video, c.f. Figure 1.

Standard neural network training algorithms can now be applied to optimize the weights and biases in the model. The softmax layer we added on top of the network is only a linear classifier. In the classical BoW model, usually a SVM with non-linear kernel is applied. Thus, we use the softmax classifier only during training in order to obtain visual words that are optimized

1. discriminatively, i.e. separating the different classes as good as possible, and
2. on video level, i.e. the visual vocabulary is learned to best differentiate between collections of feature vectors that form the video rather than between single feature vectors or video frames only.

Once the network is trained, the output layer is discarded. The neural network histograms of an input sequence \( x = x_1, \ldots, x_T \) are then used in combination with an SVM after application of a non-linear kernel.