Following the Lecturer: Hierarchical Knowledge Concepts Prediction for Educational Videos

Xin Zhang[1,2](✉), Qi Liu[1,2], Wei Huang[1,2], Weidong He[1,2], Tong Xiao[1,2], and Ye Huang[1,2]

[1] Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China, Hefei, China
{zx2020,ustc0411,hwd,xt20020109,huangyehy}@mail.ustc.edu.cn, qiliuql@ustc.edu.cn
[2] State Key Laboratory of Cognitive Intelligence, Hefei, China

Abstract. With an irresistible trend of intelligent learning, predicting knowledge concepts for educational videos turns out to be a fundamental and essential task, which benefits personalized recommendation, retrieval, and learning. Prior studies of videos mainly focus on relatively short human actions and object recognition, while educational videos are minutes long and have heterogeneous elements such as texts, formulas, and hand-drawn graphics that serve lecturers' narration. Owing to the characteristics of education, most of the segmentation strategies for long-term videos do not apply well to educational videos. In addition, educational videos consist of progressive or referential sections and contain multimodal information. Thus, we propose a novel framework called *Spotlight Flow Network* (SFNet) to obtain hierarchical knowledge concepts for educational videos with multi-modality. Specifically, we first adopt an effective text-to-visual section segmentation strategy. Then, we model the mechanism that the viewers' spotlight follows the lecturer and leverage the associations between sections to enhance multimodal representation. We also consider explicit inter-level constraints of the hierarchical knowledge structure and associations between sections and concepts to get better predicting performance. Extensive experimental results on real-world data demonstrate the effectiveness of SFNet.

Keywords: Educational videos · Multi-modality · Hierarchical multi-label classification

1 Introduction

With the rapid development of online video platforms and intelligent educational systems like Coursera and Khan Academy [12], an enormous amount of students and knowledge seekers browse educational videos to consolidate their understanding of courses and broaden their horizons. Knowledge concepts prediction

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-20500-2_13.

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 159–170, 2022.
https://doi.org/10.1007/978-3-031-20500-2_13

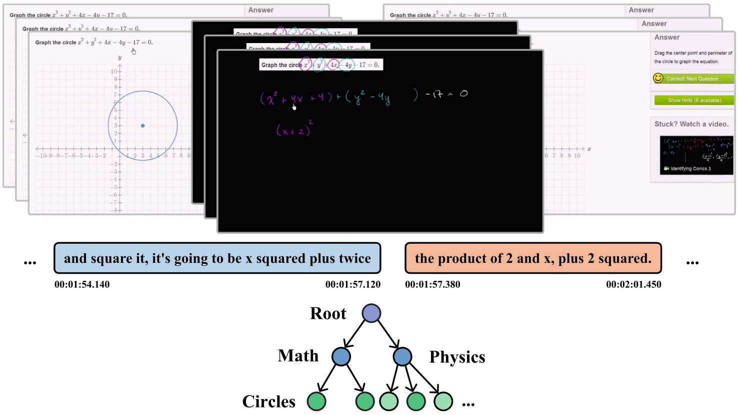
Graph the circle $x^2+y^2+4x-4y-17=0$.
Answer
Stuck? Watch a video.
...
and square it, it's going to be x squared plus twice
the product of 2 and x, plus 2 squared.
...
00:01:54.140
00:01:57.120
00:01:57.380
00:02:01.450
Root
Math
Physics
Circles
...

Fig. 1. An example of a math video from Khan academy and its related hierarchical knowledge concepts.

for educational videos is a fundamental task and very promising for organizing and managing educational videos with great quantity and diversity.

Figure 1 shows an example of a math video and related knowledge concepts with part of the knowledge structure. The video consists of multiple frames and a series of closed captions, and can be split into three different sections, i.e., introduction, problem solving, conclusion. In the last section, the lecturer refers to the problem and reviews the problem-solving process again, which demonstrates a common characteristic that educational videos are combined with sections (such as introduction, concept explanation, analysis, conclusion), and draws the importance of considering context of sections. As a key element of education, knowledge concepts are usually in the form of tree or Direct Acyclic Graph (DAG). As shown in Fig. 1, if we take the root node as level 0, sub-concepts are separated into two different routes from level 2, which describes Hierarchical Multi-label Classification (HMC). This type of problem has drawn more attention in industry and education with the trend of disciplinary crossover.

In the literature, prior works on video classification [3,4,22] have achieved great success. Most of these works mainly focus on relatively short video clips and recognize human actions and objects, while long-term video understanding has not been explored a lot yet. For long videos, prior studies [30] choose to evenly or randomly sample certain frames, or detect shot-boundaries [31] to break down whole videos into sections. Philip et al. [12] studied different types of educational videos and how video production decisions affect student engagement. Typical styles of educational videos include classroom lectures, slide presentations, "talking head" shots of an instructor and digital tablet drawings. Long-term content and more complex composition structure make the above strategy ineffective in educational videos. In addition, most recent HMC works [14] combine local and global approaches, and utilize hierarchical dependencies in the form of a feed-forward network. However, these studies fail to model explicit inter-level hierarchical constraints, and are currently limited to textual content.

In summary, there are the following challenges: (1) How to make use of multi-modal information from frames and subtitles. (2) How to consider finer-grained characteristics of educational videos that are relatively long, such as the section-level contexts. (3) How to effectively split educational videos. (4) How to explicitly model inter-level constraints in hierarchical knowledge structure.

To tackle the above challenges, we propose a novel framework named *Spotlight Flow Network* (SFNet). Specifically, we adopt a text-to-visual uniform segmentation strategy by utilizing progressiveness within a section and uniformity provided by timecodes of closed captions. Then, we model the mechanism of viewers' spotlight following the lecturers by leveraging different information from the pre-processing step. We also utilize explicit inter-level constraints of the hierarchical knowledge structure and associations between sections and concepts to improve the performance of knowledge concepts prediction. A real-world dataset of 7,521 educational videos is constructed and extensive experimental results address the effectiveness of our proposed method.

2 Related Work

Long-Term Video Understanding. In the literature, there have been many efforts to understand video content [2,5,6,13], including 2D and 3D CNN networks [9,26,32], two-stream methods [22], and well-known transformer-based methods [3,4] in recent years. Most of the prior works mainly focus on relatively short video clips (normally within 30 s) and recognize human actions, objects and scenes, etc., while long-term video understanding has not been explored a lot yet. Donahue et al. [8] proposed an end-to-end recurrent convolutional network for learning long-term dependencies. Wu et al. [31] proposed an object-centric transformer framework that recognizes, tracks, and represents objects and actions of long videos. In summary, most existing studies casually or equally sample certain frames from videos [17,27] or detect shot-boundaries [31] to breakdown whole videos into sections, yet they cannot apply well on educational videos due to the diversity and complexity of the contents.

Multimodal Video Representation. Aside from visual frames, videos also contain multimodal information such as audio and captions texts, which have complementary semantics and could enhance representation [15,23]. Shang et al. [19] utilized timestamps of closed captions to incorporate multimodal signals with a short-term order-sensitive attention mechanism. Gabeur et al. [11] developed a transformer-based architecture that jointly encodes different modalities' appearance by exploiting cross-modal cues. Nagrani et al. [16] added Multimodal Bottlenecks to input of transformer encoder and limited exchange of multimodal data in the middle of self-attention layers, and obtained more effective representation. For educational videos, VENet proposed by Wang et al. [28] exploited the static and incremental characteristics and modeled the fixed reading order of human, yet like other studies, is inadequate to fuse intra-section multi-modalities at a fine-grained level, which is emphatically concerned in our framework.

Hierarchical Multi-label Classification. There have been efforts for HMC in the literature [1,10]. Flat-based methods ignore the hierarchical structure and only leverage the last level. Local approaches adopt classifiers for each hierarchy, while global methods predict all classes with a single classifier. Recently, many hybrid methods that combine both the local and global manner have been proposed. Sun et al. [24] transformed the label prediction problem to optimal path prediction with structured sparsity penalties. Shimura et al. [21] addressed the data sparsity problem that data from the lower level is much sparser than that from upper levels and developed HFT-CNN to optimize. Wehrmann et al. [29] proposed a hybrid method called HMCN while penalizing hierarchical violations. Huang et al. [14] proposed HARNN, an attention-based recurrent network that models the correlation between texts and hierarchy. Recently, Shen et al. [20] presented TaxoClass that utilizes the core classes mechanism of humans. However, most prior studies are limited to texts and not adequate to capture the inter-level constraints of hierarchical structure.

3 Preliminaries

3.1 Problem Definition

The input of our task is an educational video $V = \{F, C\}$ composed of multiple frames $F = \{f_1, f_2, ..., f_n\}$ and closed captions $C = \{c_1, c_2, ..., c_m\}$, where each frame f_i is an RGB image in width W and height H, and a caption is made up with texts, start and end timecodes, i.e. $c_j = \{t_j, tc_j^{start}, tc_j^{end}\}$. Texts of captions can be described as a word sequence $t = \{w_1, w_2, ..., w_k\}$. The Hierarchical Knowledge Structure is denoted as $\gamma = (K_1, K_2, ..., K_H)$, where H represents the depth of hierarchy and $K_i = \{k_1, k_2, ...\}$ is the set of knowledge concepts of level i. The Predicted concepts are $L = \{l_1, l_2, ..., l_H\}$ where $\forall i \in \{1, 2, 3, ..., H\}$, and $l_i \subset K_i$. Given an educational video V and the hierarchical knowledge structure γ, our goal is to predict the knowledge concepts L for the video.

3.2 Text-Visual Uniform Section Segmentation

Unlike previous works [31] that detect shot boundaries of visual frames and then guide the segmentation of captions, we preprocess sequential frames and closed captions by exploiting timecodes of captions. We first complement closed captions for videos using ASR (Automatic Speech Recognition) tools. We observe that educational visual content within a section is progressive and later frames tend to contain more information. Thus, inspired by Adaptive Block Matching (ABM) [28] and Dynamic Frame Skipping [18], we develop an efficient section segmentation strategy that fits well in educational videos:

1. Select the center frames of timecode gaps as candidates of sections.
2. Merge sections. Replace adjacent candidates within t_{min} by the latter ones.
3. Calculate the difference matrix $diff$ and score σ of all adjacent candidate frames by pixel-wise value subtraction.

4. Merge sections if corresponding difference score σ is less than threshold θ_{min}.
5. Calculate all ABM scores δ for adjacent candidates if difference score σ is greater than threshold θ_{max}.
6. Select top $n_{sections}$ candidates of δ as the keyframes representing each section with uniform pairs of caption and difference matrices between sections.

It is worth noting that the ABM score is calculated by dividing two frames into patches and measuring how the latter patches cover the previous ones. The difference score σ of the k-th candidate can be expressed as:

$$\sigma(k) = \frac{1}{W * H} \sum_{i=0,j=0}^{i<W,j<H} |f_{ij}^{k+1} - f_{ij}^{k}|, \tag{1}$$

where f_{ij}^{k} denotes the scaled pixel value of the k-th candidate frame. As a result, each input video is split into fixed number of sections. Each section comprises a keyframe and several uniform pairs of difference matrix and caption texts, serving the modeling of fine-grained spotlight flow within section.

4 Spotlight Flow Network

In this section, we introduce the details of SFNet, as shown in Fig. 2. We will discuss the two main parts, especially present the modeling of the Spotlight Flow Mechanism and specify the loss function used to train the model.

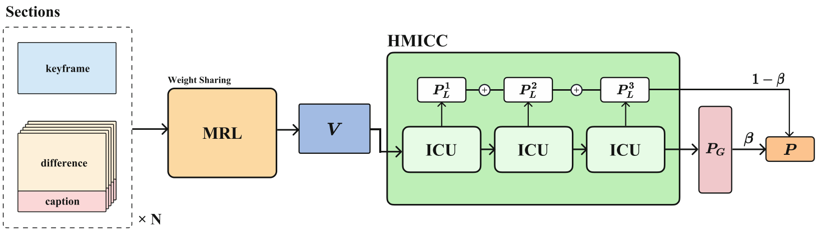

Fig. 2. The SFNet framework.

4.1 Multimodal Representation Layer

In the first stage of SFNet, we aim to represent each section by encoding multimodal data and modeling Spotlight Flow Mechanism, and obtain video-level representation. The input of each section is a keyframe and several uniform pairs of difference matrices and caption texts. We first utilize a variant of ResNet [25] to extract keyframe feature $r^f \in \mathbb{R}^{d_1}$. A base version of BERT is used to get sequential semantic vectors $r^{cs} \in \mathbb{R}^{t \times d_1}$ for all captions within the section, where t denotes the number of diff-caption pairs.

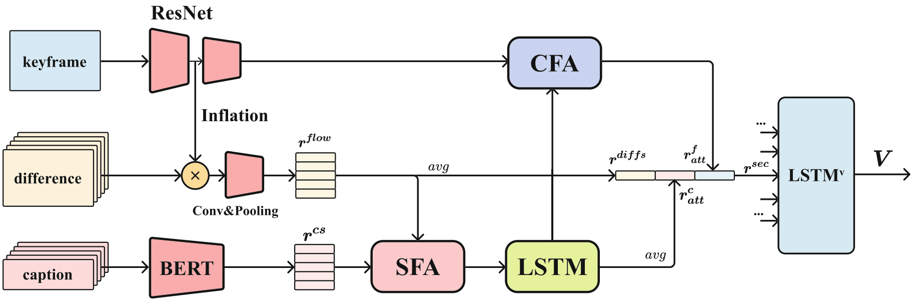

Fig. 3. Multimodal Representation Layer.

Spotlight Flow Attention (SFA). We observe that lecturers tend to conduct viewers to focus on certain visual regions. Content that periodically comes out or is regularly referenced by underlines, circle drawings, etc., strongly indicates the correlation of different time periods and connects time and moving regions. Thus, SFA is designed to model the above mechanism. Inspired by I3D [7], we inflate the feature maps from the middle of the backbone and get $r^{mid} \in \mathbb{R}^{t \times d_2 \times w \times h}$. We resize difference matrices with interpolation and apply element-wise multiplication as follows:

$$r^{flow}_{(i,j)} = r^{mid}_{(i,j)} \cdot diff_{(i,j)}, \tag{2}$$

and through the latter part of the feature extractor, we get the corresponding features of moving regions $r^{flow} \in \mathbb{R}^{t \times d_1}$. Then SFA can be formulated as:

$$r^{c}_{att} = SFA(r^{flow}, W_{sf}, r^{cs}) = softmax(r^{flow} \cdot W_{sf})r^{cs}, \tag{3}$$

where matrix $W_{sf} \in \mathbb{R}^{t \times d_1}$ is the hidden matrix. Considering the association between the sequential captions, we utilize Bi-LSTM that is capable of learning dependencies across the sequence forward and backward at the same time. We input $r^{c_{att}}$ and r^{flow} with the same size on temporal dimension, and the final representation of caption r^{c} is calculated by average pooling the hidden state.

Caption Frame Attention (CFA). We propose CFA by taking the correlation between captions and related parts of the visual content across time. We exploit CFA by using the hidden states of Bi-LSTM h as the query of attention input:

$$r^{f}_{att} = avg(CFA(h, r^{f}, r^{f})) = avg(\frac{softmax(h \cdot r^{f})}{\sqrt{d_k}} r^{f}), \tag{4}$$

where $avg()$ denotes the average pooling operation, and d_k represents the scaled factor. Therefore, the representation of the section is as follows:

$$r^{sec} = r^{c} \oplus avg(r^{diffs}) \oplus r^{f}_{att}, \tag{5}$$

and the final output of MRL is calculated by inputting all the section-level features to a video level Bi-LSTM to model the correlation among sections.

4.2 Hierarchical Multi-label Inter-level Constrained Classifier

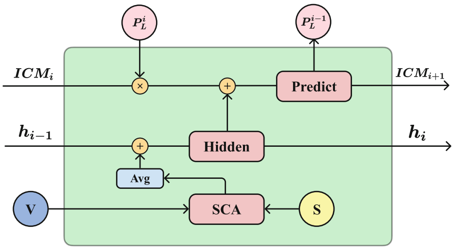

Fig. 4. Inter-level constrained unit.

Since we have obtained the multimodal representation of the video $V \in \mathbb{R}^{n \times d}$, where n is the number of sections, a Hierarchical Multi-label Inter-level Constrained Classifier (HMICC) is proposed to predict knowledge concepts for educational videos based on the feed-forward manner of current hybrid methods. The network consists of several Inter-level Constrained Units (ICU) shown in Fig. 4. Each one utilizes Section-Concept Attention (SCA) and Inter-level Constrained Matrix (ICM) to model each level's dependencies and feed the hidden information to the next unit. Specifically, $S^i \in \mathbb{R}^{C^i \times d}$ denotes the hidden representation of the i-th level and is input to SCA together with V. We apply the dot-product scores to measure the similarity of categories and video sections:

$$\begin{aligned} V_{att} &= softmax(S^i \cdot V) \cdot V, \\ r_{att}^v &= avg(V_{att}), \end{aligned} \tag{6}$$

where we operate average pooling on temporal dimension to get r_{att}^v. Then we concatenate r_{att}^v and the previous hidden state h_{i-1} to obtain h_i by:

$$h_i = \varphi(W_h(r_{att}^v \oplus h_{i-1}) + b_h), \tag{7}$$

where $\oplus$ denotes concatenation. Here we adopt the Inter-level Constrained Matrix $ICM_i \in \mathbb{R}^{C^i \times C^{i+1}}$ with each icm_{jk} representing the influence of the j-th category on the k-th one to the next level. We initialize all ICMs by calculating the conditional probabilities from the training set. The result of product between ICMs and previous prediction is added up to get the local prediction through a hidden layer, and the global output is obtained by inputting the last hidden state through a fully-connected layer:

$$\begin{aligned} P_L^i &= \sigma(W_L((P_L^{i-1} ICM_i) \oplus h_i) + b_L), \\ P_G &= (W_G \cdot h_H + b_G), \end{aligned} \tag{8}$$

where W_h, W_L and b_H, b_L are weight matrices and bias vectors. Therefore, we can calculate the final predictions P with a parameter $\beta \in [0, 1]$ for balancing the local and global outputs:

$$P = \beta \cdot P_G + (1 - \beta) \cdot (P_L^1 \oplus P_L^2 \oplus, \ldots, \oplus P_L^H), \tag{9}$$

4.3 Training SFNet

In this section, we specify a hybrid loss function for training SFNet to learn both global and local information. We calculate the global loss($\mathcal{L}_G$) and the local loss($\mathcal{L}_L$) for each hierarchical level, which can be formulated as:

$$\mathcal{L}_G = \varepsilon(P_G, Y_G), \mathcal{L}_L = \sum_{h=1}^{H} \varepsilon(P_L^h, Y_L^h), \tag{10}$$

where Y_G denotes the binary label vector for all categories of the knowledge structure and Y_L^h contains only the categories of the h-th level. We utilize the binary cross-entropy loss as $\varepsilon(\hat{Y}, Y)$ and formulate the final loss function as:

$$\mathcal{L}(\Omega) = \mathcal{L}_L + \mathcal{L}_G + \lambda ||\Omega||^2, \tag{11}$$

where Ω denotes the parameters of SFNet and λ is the hyper-parameter for L2 regularization. Thus, we can train SFNet by minimizing the loss function $\mathcal{L}(\Omega)$.

5 Experiments

5.1 Data Description

To evaluate the performance of our framework, we construct the dataset by collecting 7,521 educational videos, corresponding closed captions and hierarchical knowledge concepts from **Khan Academy**[1] The dataset involves a three-level hierarchical knowledge structure with 6, 42, 351 concepts in each level, and 399 in total. Averagely, a video is 436.4s long and has 1151 words of captions.

5.2 Baseline Approaches and Experimental Setup

We compare our proposed model with state-of-the-art works including unimodal and multimodal approaches. It is worth noting that all baseline models are pre-trained on ImageNet, Kinetics dataset, etc., according to the categories, and tuned to obtain the best results.

- **R3D** [26] is a deep 3D convolution network with residual connection across layers and enables a very deep network structure while retaining performance improvement.

[1] All Khan Academy content is available for free at www.khanacademy.org.

- **SlowFast** [9] is a two-stream 3D CNN network that consists of two different paths that separately focus more on temporal and spatial information.
- **TimeSformer** [4] is a video transformer network that uses frame patches with positional encoding as input and exploits divided spatial and temporal self-attention.
- **R3D+BERT** is the combination of R3D and BERT. We leverage BERT to obtain the feature of captions and fuse the visual feature from frames.
- **HMCN-F** [29] is a feed-forward network that models the top-down hierarchical relationship and optimizes both local and global performance with penalties of hierarchical violations.

We implement all the methods using Pytorch. To train SFNet, we first set $n_{sections}$ as 8 and the maximum length of words for each caption as 64. We use ResNet34 and BERT-base as the feature extractor backbones and set the output dimension to 256. Hidden sizes of Bi-LSTM and HMICC are 128. We use the Adam optimizer and set up the initial learning rate to 0.0005 with cosine annealing scheduler that periodically adjust the value to 0.00005 for every 60 epochs. We also set $\beta = 0.5$, $\lambda = 0.00005$ and dropout rate as 0.5 to mitigate over-fitting. We used $Precision$, $Recall$, $F1 - score$, and mAP (mean Average Precision) as criteria for performance comparison. Whether a model considers the knowledge hierarchy or not, we calculated the performance at each hierarchical level and globally as well to further compare the differences.

Table 1. Performance Comparison on khan academy dataset. V and T denote the visual and textual modalities of the input data.

Model	Input	mAP	Precision	Micro-F1	Recall
R3D	V	0.6089	0.6745	0.5897	0.4591
SlowFast	V	0.6433	0.6936	0.6124	0.5431
TimeSformer	V	0.6799	0.7126	0.6295	0.5982
HMCN-F	T	0.7321	0.7640	0.6724	0.6213
R3D+Bert	V+T	0.8125	0.8391	0.7204	0.6428
SFNet	V+T	**0.8351**	**0.8712**	**0.7628**	**0.6787**

5.3 Experimental Results

Performance Comparison. From the results shown in Table 1 and Fig. 5, we can get several observations. First, models with textual input tend to outperform those visual-only models. In educational videos, visual content serves the lecturers' explanation. Due to the complexity and variance of visual elements such as hand-drawn graphics, it is harder to understand the semantics than textual content. It also indicates the significance of spotlight flow attention. Second, it is obvious that the performance decreases as the level gets lower. Hierarchical structure has a natural identity that higher levels have fewer categories and more

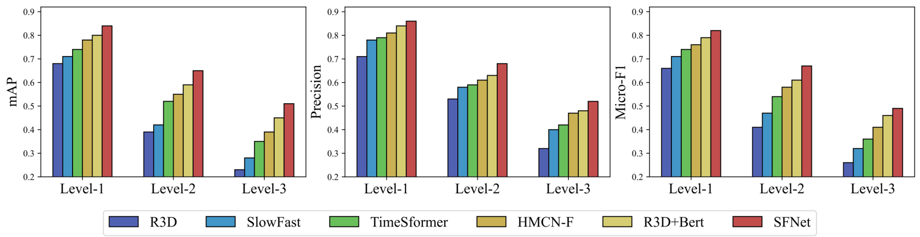

Fig. 5. Performance of SFNet and baseline models on different hierarchical levels.

data, which might explain the step down of performance. The results show that SFNet is more efficient by considering the inter-level association.

Ablation Study. To further assess how each part of our model donates to the performance, we remove each key module once at a time and construct several variants of SFNet. In Table 2, all the key modules do have contribution to better-predicting performance. The greater difference indicates more impact of the removed module. In addition, the variant without textual input has the greatest performance drop, which once again showing the above characteristics.

Table 2. The results of ablation study. V and T represent visual and textual input.

Model	mAP	Precision	Micro-F1	Recall
SFNet	0.8351	0.8712	0.7628	0.6787
V-only	0.5897	0.6528	0.5734	0.4345
T-only	0.7654	0.8133	0.7282	0.6571

6 Conclusion

In this paper, we presented Spotlight Flow Network to predict knowledge concepts for educational videos. We first adopted an effective text-to-visual section segmentation strategy for educational videos. Then, with different information paired with captions, we modeled the Spotlight Flow mechanism in which lecturers tend to conduct viewers' attention and moving regions help build up space-time connection. We also designed the HMICC to predict hierarchical knowledge concepts with implicit progressive impact and explicit inter-level constraints.

Acknowledgements. This research was partially supported by grants from the National Key Research and Development Program of China (No. 2021YFF0901003), and the National Natural Science Foundation of China (Grant No. 61922073).

References

1. Aly, R., Remus, S., Biemann, C.: Hierarchical multi-label classification of text with capsule networks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 323–330 (2019)
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. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4 (2021)
5. Bhardwaj, S., Srinivasan, M., Khapra, M.M.: Efficient video classification using fewer frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 354–363 (2019)
6. Cao, J., Mao, D.H., Cai, Q., Li, H.S., Du, J.P.: A review of object representation based on local features. J. Zhejiang Univ. Sci. C **14**(7), 495–504 (2013). https://doi.org/10.1631/jzus.CIDE1303
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
8. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
10. Feng, S., Fu, P., Zheng, W.: A hierarchical multi-label classification algorithm for gene function prediction. Algorithms **10**(4), 138 (2017)
11. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 214–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_13
12. Guo, P.J., Kim, J., Rubin, R.: How video production affects student engagement: an empirical study of MOOC videos. In: Proceedings of the first ACM Conference on Learning@ Scale Conference, pp. 41–50 (2014)
13. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition: a survey. Image Vis. Comput. **60**, 4–21 (2017)
14. Huang, W., et al.: Hierarchical multi-label text classification: an attention-based recurrent network approach. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1051–1060 (2019)
15. Liang, M., Cao, X., Du, J., et al.: Dual-pathway attention based supervised adversarial hashing for cross-modal retrieval. In: 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 168–171. IEEE (2021)
16. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Adv. Neural. Inf. Process. Syst. **34**, 14200–14213 (2021)
17. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3163–3172 (2021)

18. Seo, J.J., Kim, H.I., De Neve, W., Ro, Y.M.: Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection. Image Vis. Comput. **58**, 76–85 (2017)
19. Shang, X., Yuan, Z., Wang, A., Wang, C.: Multimodal video summarization via time-aware transformers. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1756–1765 (2021)
20. Shen, J., Qiu, W., Meng, Y., Shang, J., Ren, X., Han, J.: Taxoclass: hierarchical multi-label text classification using only class names. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4239–4249 (2021)
21. Shimura, K., Li, J., Fukumoto, F.: HFT-CNN: learning hierarchical category structure for multi-label short text categorization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 811–816 (2018)
22. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
23. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: a joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7464–7473 (2019)
24. Sun, Z., Zhao, Y., Cao, D., Hao, H.: Hierarchical multilabel classification with optimal path prediction. Neural Process. Lett. **45**(1), 263–277 (2017). https://doi.org/10.1007/s11063-016-9526-x
25. Targ, S., Almeida, D., Lyman, K.: Resnet in resnet: generalizing residual architectures. arXiv preprint. arXiv:1603.08029 (2016)
26. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
27. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
28. Wang, X., et al.: Fine-grained similarity measurement between educational videos and exercises. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 331–339 (2020)
29. Wehrmann, J., Cerri, R., Barros, R.: Hierarchical multi-label classification networks. In: International Conference on Machine Learning, pp. 5075–5084. PMLR (2018)
30. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 284–293 (2019)
31. Wu, C.Y., Krahenbuhl, P.: Towards long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1884–1894 (2021)
32. Zhang, S., Guo, S., Huang, W., Scott, M.R., Wang, L.: V4d: 4d convolutional neural networks for video-level representation learning. arXiv preprint. arXiv:2002.07442 (2020)