Recursive inference for individual's identification in video data based on the description

Dearo Kim, Jiyoun Lim, Jeongwoo Son, and Namkyung Lee
Media Intellectualization Research Section
Electronics and Telecommunications Research Institute
Deajeon, Korea
eofh4817@gmail.com, {kusses, jwson, nklee}@etri.re.kr

***Abstract*—This study aims to enhance the Textual and Visual Representations for the identification of the same individuals within video descriptions. The character inference and ID generation for identification of individuals based on Transformer. A scene of a video visually represents that 'who' acts 'what' and 'how' in 'where' and 'when'. Among entities composing a scene, a person plays a crucial role in representing the context of the scene. Thus, in the video description problem, various ways to identify the persons in the scene have been studied. This paper deals with the problem of 'fill-in-the characters' that aims to predict the local IDs of characters that appeared through several consecutive scenes. In this task, it is demanded to predict local IDs of character that it is not required to recognize each character globally (in an entire movie), but locally (within a set of 5 clips). Due to the restriction of the problem definition, global identifications of characters cannot be obtained with contemporary methods while it is often required to deploy them in services and applications. To resolve this problem, we propose the method of recursive inference of local ID. Additionally, we propose optimizing Bert embedding for mask tokens from video descriptions to infer character's local IDs. According to the experimental results, the recursive process allows the acquisition of coherent representations among unique individuals.**

***Index Terms*—person re-ID, visual representation, language representation, recursive inference**

I. Introduction

To fully comprehend what happens in a video based on the description of the film, it is crucial to identify the individuals involved. In recent research, Text and Visual representations have been extracted to facilitate the re-identification of characters appearing in N consecutive scenes, and **local ID**s for a certain individual within video descriptions have been generated using Generative Transformers. However, conventional fill-in-the-blank models of this nature solely focus on identifying **local ID**s within short segments of N consecutive scenes, thereby failing to extend this identification to cover **local ID**s corresponding to the entire video sequence, which could be a movie or TV series. Addressing this challenge necessitates the expansion of the scope of **local ID** inference within video sequences. This study optimizes the Bert embedding that was previously proposed for character prediction [1] in order to increase character identification performance.

The objective of our study is to enhance the coherence of Textual and Visual representations among identical local IDs and predict **local ID**s encompassing the entire video sequence. To achieve this goal, this research introduces a recursively learned Identity Inference model trained on the Large Scale Movie Description Challenge (LSMDC) dataset. Initially, our model infers local IDs for short video segments. It heightens the coherence between the Visual and Textual representations of inferred identities and regenerates representations for characters. By progressively broadening the video segments and iteratively performing identity inference and representation generation, we engage in recursive inference until the video segment encompasses the entirety of the video sequence. Ultimately, employing the inferred characters, we employ a transformer model to generate local IDs. The proposed recursive inference approach in this study demonstrated superior performance in identifying local IDs for the same characters and effectively recognizing local IDs for the complete video sequence.

II. Relate Works

A. Identification of Person's ID

One of the tasks presented at LSMDC 2021 was the person reID problem [1] , proposed as a baseline model for this challenge, presented the following transformer-based character identification method. Part et al [1]. focused on re-identifying characters using local IDs for short segments. In this research, they aimed to incorporate gender information along with text and visual information about the characters. Alongside character face cluster classification, gender annotation was performed on the LSMDC dataset, which contains annotations for gender. The study utilized the classified face clustering label and gender information through a transformer model to generate local IDs.

[2] aimed to identify the Global IDs of individuals appearing in the video. To achieve this, they mapped verbs associated with characters in the description and the body tracks of characters to the Textual-Visual Embedding Space. Furthermore, each character's face cluster was assigned a unique proper noun, which served as the global ID. By utilizing verbs related to characters within the description, they retrieved the mapped body track and assigned the global ID from the corresponding face cluster within that body track to the character.

[1]https://sites.google.com/site/describingmovies/

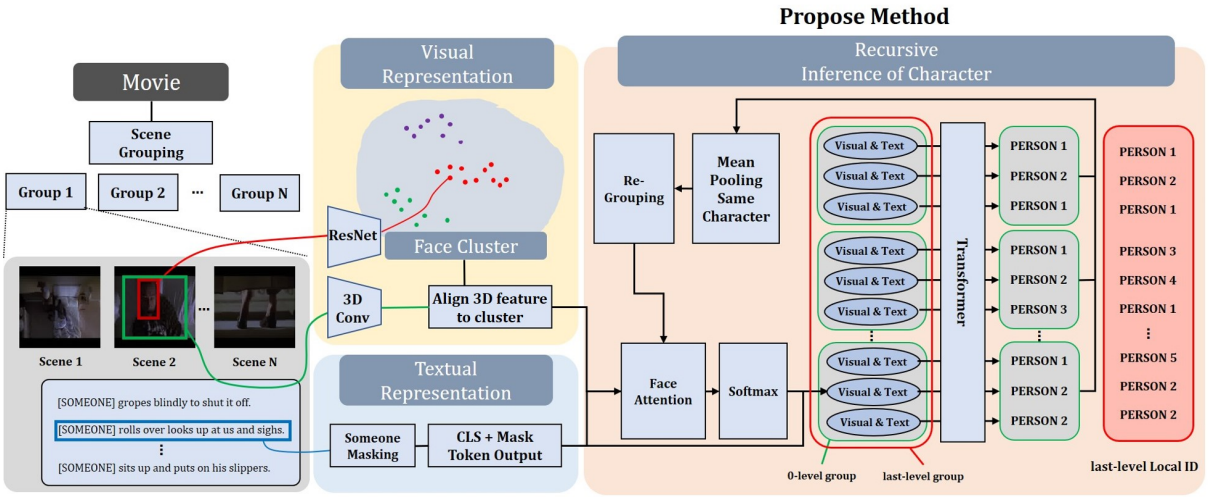

Figure 1. Recursive Fill-in Model Architecture based on [1]

III. Recursive Inference Model

Our model is structured into three main stages of **Fig.1**. Firstly, we create segments (T scenes) of consecutive scenes and extract Textual and Visual representations. Secondly, we classify Character classes for individuals and consolidate representations of identical individuals to generate new representations. These steps are recursively repeated by grouping the set of scenes into T segments until the entire video is grouped. Finally, based on the Character classes, we generate local IDs for the entire video sequence.

A. Extracting Textual and Visual Representation

To extract text representations containing contextual information from video descriptions, a pre-trained Bert Model [3] is employed. The Bert model not only captures word similarities but also excels in context understanding due to its bidirectional analysis of context. To extract Bert embeddings containing contextual information, the representation of the **CLS** tokens and MASK tokens are optimized. The Bert Model comprises 12 hidden layers, each consisting of multi-head self-attention and **PFFN** (position-wise feed-forward networks), with each layer's output nodes being 768-dimensional. As layers progress, more specific abstractions are produced. To prevent excessive abstraction and enhance generalization, embedding vectors are obtained by mean pooling the states of the last H hidden layers. In our model, the outputs of the **CLS** token and MASK token are combined creating a 1536-dimensional text representation to capture the character context of a sentence.

For Visual representation, we utilize face cluster embeddings and **I3D** (Inflated 3D ConvNet) features. We track faces throughout the set of T scenes and extract 512-dimensional face features using **FaceNet** [4], a model trained on the **VGGFace2** dataset. These extracted face features are used to create face clusters through the **DBSCAN** algorithm, and the average value of each cluster serves as the cluster's feature. Additionally, to link Visual representation with text representation through the context of scenes, we utilize the **I3D** model [5]. **I3D** features are extracted, containing spatiotemporal information about individual actions. By mapping Face clusters and **I3D** features based on scene timing, we integrate Visual Representations, adding action information to face clusters.

B. Recursive Inference Local IDs

The core of the **Recursive Inference of Character** stage is the recursive repetition of character inference, representation mean pooling, and scene grouping.

The model to predict character classes contains visual features such as face cluster corresponding to SOMEONEs, using extracted text, face features, and I3D features. To enhance the coherence of representations between identical individuals, mean pooling is applied to the text and visual representations of each individual. The model proposed in this study infers individuals within the segments of grouped scenes, considering that scenes are grouped into N segments. Consequently, mean pooling for representations is performed for individual identification within the same group.

The initial 0-level group is composed of N scenes to infer local IDs spanning over segments corresponding to groups. After performing local ID inference and representation mean pooling, local ID identification within a group possesses the same representation. Subsequently, the 0-level group is grouped again into 1-level groups, each consisting of N segments. By recursively repeating the steps of inferring local ID, representation mean pooling, and scene grouping, we ultimately obtain local ID inference results for all scene segments of the video. Notably, **Recursive Inference of Local IDs** is conducted during the inference stage and not during training; local ID inference alone is executed.

Table I
EVALUATION WITH VARIOUS BERT REPRESENTATIONS

Extraing Embedding Method	Same–Acc	Diff–Acc	Class–Acc	Inst–Acc
BERT Gender Language Model, Park et al. **[1]** (Pooled output[CLS] + MASK token output)	62.8	68.0	65.3	69.7
Pooled output[CLS] + Pooling MASK token of last 2 layers	58.5	72.0	64.5	70.3
Pooled output[CLS] + Pooling MASK token of last 4 layers	56.5	73.0	63.7	70.3
Pooling all tokens of last layer + Mask token output	55.6	72.0	62.7	69.3
Pooling all tokens of last 2 layers + Mask token output	56.8	70.6	63.0	69.0
Pooling all tokens of last layer + Pooling MASK token of 2 layer	58.6	71.1	64.2	70.2

C. Generating Local ID

Once character inference for the entire video sequence has been performed, to sequentially generate appropriate local IDs, the inferred character classes and text representations are used as input for the transformer model [6]. This transformer model then generates the appropriate local IDs.

IV. EXPERIMENTS

For model training, the cross-entropy loss was computed using the output of the face attention. The Adam optimizer was utilized in this process. To evaluate the model, a pairwise comparison between the inferred characters and the generated IDs, and the ground truth IDs was conducted. For evaluation metrics, the following measures were used:

- **Same ID Accuracy (Same-Accuracy)**: This metric represents the accuracy of correctly identifying instances where the model determines the same individuals as having the same ID label.
- **Different ID Accuracy (Diff-Accuracy)**: This metric reflects the accuracy of accurately identifying instances where the model distinguishes different individuals by assigning different ID labels.
- **Instance Accuracy (Instance-Accuracy)**: This metric measures the accuracy of correctly distinguishing between the same and different individuals for each instance.
- **Class Accuracy (Class-Accuracy)**: This metric is obtained by calculating the average of Same-Accuracy and Diff-Accuracy and indicates the accuracy of differentiating between the same and different individuals across all instances.

These metrics were employed to assess the performance of the model in its ability to correctly identify the same and different individuals.

A. Extracting Text Representation

Park et al. **[1]** extracted text representation with BERT gender language model learned to include gender information such as 'his' and 'her'. We replace the 'SOMEONE' in the description with a Mask Token. Subsequently, the pre-trained BertModel was applied and values from the hidden state were employed. The pooled output which is usually used output of CLS token through dense layer and activation function with the Bert Model participated in this experiment as the baseline. The CLS token that contains the information on sentence classification, pays high attention to the words highly related to the context using attention weight. Therefore, the CLS token has a tendency to miss words that appear to not affect sentence classification. Based on this point, we propose a method for extracting sentence-level embedding, assuming that the context of the entire sentence affects character identification. The sentence-level embedding is extracted by mean pooling of output embedding of entire tokens in the sentence, whereas the MASK token embedding only comes from MASK tokens in the sentence. In the mean pooling of the last 2 layers, we apply mean pooling between hidden layers in advance. Then, the mean pooling of entire tokens is performed. The result using sentence-level embedding has shown a lower result than the case considering the token-level embedding. We extracted the embedding from the mean pooling of various combinations from 12 hidden layers, not using the last layer solely. According to the result as shown in **Table I**, The embedding constructed through the mean pooling with the last 2 hidden layers identifies the characters better than the other combinations. Even though the proposed model with 4-layer pooling shows better performance with 73.0 of **Diff-Acc**, one with 2-layer pooling achieved better performances with respect to other measures.

B. Recursive Inference of Local IDs

The performance of the 0-level is equivalent to the performances in **Table I**. Five segments in the previous level are tied as a single segment in the current level, e.g. i-level group includes 5^{i+1} scenes. As the scope of sequences expands, that much input representations of transformer are needed to infer the increased characters. However, the increased input representation are reduced back to the number of unique characters in the group through representation cohesion between the same IDs. With recursive structure, the model demonstrated the ability to infer local IDs effectively across higher-level group intervals, despite being trained on 0-level grouping. Moreover, the recursive procedure of the proposed model demands much less resources than contemporary methods to reveal globally consistent IDs for a video. Thus, the problem is applicable to more diverse services and applications.

V. Conclusion

In this paper, the textual representation to improve local ID identification performance and the recursive inference of person's Local IDs is proposed. According to the experiment on textual representation's effect on character recognition with the LSMDC dataset, token-level representation by pooling the states of several hidden layers is more effective than directly using the output of the final layer of BERT. We confirmed that local ID can be identified for the expansion video scenes by reducing input representation of transformer with a model trained on 0-level scene group accuracy.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00852, Development of Intelligent Media Attributes Extraction and Sharing Technology)

References

[1] J. Park, T. Darrell, and A. Rohrbach, "Identity-aware multi-sentence video description," *Computer Vision - ECCV*, pp. 360–378, 2020.

[2] S. Pini, M. Cornia, L. Bolelli, and R. Cucchiara, "M-vad names: a dataset for video captioning with naming," *Multimedia Tools and Applications 78*, pp. 14 007–14 027, 2019.

[3] D. Jacob, C. Ming-Wei, L. Kenton, and T. Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805v2*, 2019.

[4] S. Florian, K. Dmitry, and P. James, "Facenet: A unified embedding for face recognition and clustering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[5] C. Joao and Z. Andrew, "Quo vadis, action recognition? a new model and the kinetics dataset," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need." *In: Advances in Neural Information Processing Systems*, 2017.