Disentangled Representation Learning in Sign Language Recognition

İpek Erdoğan

January 31, 2021

1 Introduction

My aim in this project is to provide disentangled representation learning in sign language recognition tasks. The paper I chose for this purpose is "Gait Recognition via Disentangled Representation Learning" [1]. Gait is a person's manner of walking. In gait recognition, they try to identify persons based on their walking styles. Yet this paper is from a completely different field, the problem they try to solve is so similar.

In sign language recognition tasks, results effect from differences in appearances of signers, since model creates a relationship between the label of the video and the features coming from signer identity. Traditional sign language recognition methods suffer from degraded recognition performance when handling confounding variables. To solve this problem, we need to disentangle necessary features and redundant (signer identity related) features in feature extraction part.

Disentangled representation learning problem is not a specific problem for sign language recognition. It is also a need in different fields, with similar reasons. For face recognition, extracting pose-invariant features is important. On the other hand in action recognition task, it's important to get identity-invariant features. In the paper I chose, their aim is to learn disentangled representations to improve gait recognition performance.

2 Literature Review

There are different approaches to provide disentangled representation learning. One of these approaches is to use generative adversarial networks. Tran et al. proposed a GAN model, DR-GAN, [2] to extract pose-invariant features for face recognition tasks. Liu et al. also proposed an adversarial network model, MTAN [3], after DR-GAN is proposed. Their aim is accomplishing disentangled feature learning to extract style-invariant features. They are both based

on an encoder-discriminator-generator architecture, yet there are differences in between. There are two discriminators for two different tasks (classifying the image by its content and classifying the image by its style) in MTAN, whereas there is one discriminator in DR-GAN, which tries to understand if its input image is real or fake. Also, in DR-GAN, discriminator's input is an image but in MTAN, discriminators' input is extracted feature representation.

Another approach for disentangled representation learning is to extract different components of images and infer them separately at the representation learning part. Yang and Yao proposed a variational auto encoder based method [4] for learning disentangled representations of hand poses. By using different factors of variations (hand pose, viewpoint of the camera and image content), VAE framework learns a disentangled representation for RGB hand images. Zhao et al. proposed another method [5] for disentangled representation learning in action recognition. They use a convolutional neural network architecture for video representation learning, which derives different components of dynamics from raw video frames. These components are static appearance, apparent motion, and appearance changes. They use three modules to achieve that result: 3D pooling, cost volume processing and warped feature differences. The method Jiang et al. proposed a different strategy to get a disentangled 3D face shape representation. 3D face shape is degraded into two parts: identity part and expression part. These parts are both encoded in a nonlinear way.

Disentangled feature learning problem can be considered as a domain adaptation problem. In domain adaptation, the aim is sharpening the effects of the target domain while decreasing the effects of the source domain. In disentangled feature learning, the aim is learning a feature representation which highlights the features related to the target domain and grays the features related to the source domain.

In this point of view, Ganin and Lempitsky [6] proposed a valuable domain adaptation approach which can adapt an architecture which belongs to a domain, in another domain.There are two classifiers which come after a feature extractor, in the proposed approach. One of them is for target domain and other one is for source domain. They use a new layer which is called "gradient reversal layer" between the feature extractor and source domain classifier in their model to manipulate the back propagation phase of training. Thus they train the feature extractor in a way that will increase the performance of the target domain classifier whereas decrease the performance of the source classifier. Ferreira et al. [7] use exactly same method in their model to provide a solution for the signer-dependency problem. Their model gives the feature it extracts with an encoder to two different classifiers: one for sign gloss prediction and one for signer prediction.

3 Proposed Method

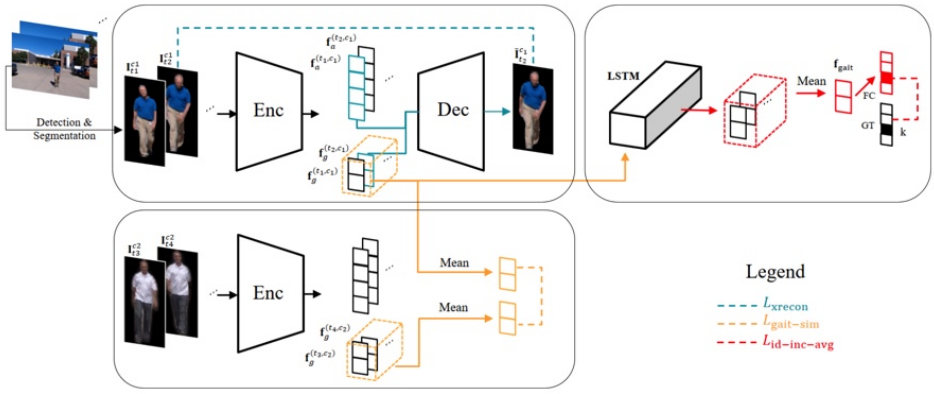

Figure 1: Proposed Model

The model proposed in this paper is in Figure 1. Input of the model is a video frame. Its background has been removed using Mask R-CNN [8]. There is an encoder-decoder network, with novel loss functions, is used to disentangle the appearance and pose features for each video frame. Then, to consider the temporal relation of pose features and combine them into a gait feature for gait identification purpose, a multi-layer LSTM [9] is being used. You can see the overall loss function to minimize, in equation(1).

$$\mathcal{L} = \mathcal{L}_{id-inc-avg} + \lambda_r \mathcal{L}_{xrecon} + \lambda_s \mathcal{L}_{gait-sim} \tag{1}$$

The components of the overall loss function are, gait similarity loss, cross reconstruction loss, and incremental identity loss. Details of loss functions will be explained in next chapter.

4 Paper Contributions and Analysis

4.1 Novel Loss Functions

In the generative models, for example in DR-GAN [2] model, adversarial loss only confuses the discriminator by using real or fake labels. There is no adversarial strategies are adopted between different class categories. Yet, determining being "real" or "fake" is not really fits our targeted measurement here. Another popular loss function which is popular to compare images in generative adversarial networks is reconstruction loss. But again, using reconstruction loss alone doesn't ensure appearance information and pose information (which are referred

as f_a and f_g in the formula) are being learned separately.

$$\mathcal{L}_{xrecon} = \|\mathcal{D}(f_a^{t_1}, f_g^{t_2}) - I_{t_2}\|_2^2 \tag{2}$$

This paper proposes the cross reconstruction loss(2), using an appearance feature f_a of frame 1 and pose feature f_g of frame 2 to reconstruct frame 2. The cross reconstruction loss can ensure the two features are representative enough to reconstruct video frames. Since we can use a pose feature of a one frame and the appearance feature of any frame in the same video together, to reconstruct the same result, it enforces the appearance features to be similar for all frames.

$$\mathcal{L}_{gait-sim} = \left\| \frac{1}{n_1} \sum_{t=1}^{n_1} f_g^{t,c_1} - \frac{1}{n_2} \sum_{t=1}^{n_2} f_g^{t,c_2} \right\|_2^2 \tag{3}$$

Although the cross reconstruction loss is being used, some appearance information may still be leaked into pose feature. To make pose feature "cleaner", this paper also proposes gait similarity loss(3). In the input dataset, they considered multiple videos for same subject. Their aim is to provide same gait information between two videos while appearance changes. So in sign language field, it refers: While signer changes for the same gloss, the gloss information should be consistent between two videos.

This part is very important: It's not possible to enforce and measure similarity on gait feature vectors between different videos' frames, as it requires strict frame-level alignment. In order to overcome this issue, they enforce the similarity between two videos' averaged pose features. This is a very novel and meaningful approach since it solves the problem we encounter while trying to use GAN solutions for disentangled representation learning in sign language recognition tasks.

$$\mathcal{L}_{id-inc-avg} = \frac{1}{n} \sum_{t=1}^{n} -w_t log\Big(C_k\big(\frac{1}{t} \sum_{s=1}^{t} h^s\big)\Big) \tag{4}$$

The third loss function they propose is incremental identity loss(4). They propose to use the averaged LSTM output as the gait feature since output of the LSTM is heavily effected by its last input.

4.2 Results

Methods	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°	Average
CPM [12]	13	14	17	27	62	65	22	20	15	10	24.1
GEI-SVR [29]	16	22	35	63	95	95	65	38	20	13	42.0
CMCC [28]	18	24	41	66	96	95	68	41	21	13	43.9
ViDP [26]	8	12	45	80	**100**	**100**	81	50	15	8	45.4
STIP+NN [30]	–	–	–	–	84.0	86.4	–	–	–	–	–
LB [46]	18	36	67.5	93	99.5	99.5	92	66	36	18	56.9
L-CRF [12]	38	**75**	68	93	98	99	**93**	67	76	39	67.8
GaitNet (ours)	**68**	74	**88**	91	99	98	84	**75**	76	**65**	**81.8**

Figure 2: Recognition accuracy cross views

For disentangled representation learning, it's hard to compare the results since the way to compare extracted representations is to process them in their own field. It's not appropriate to compare the DR-GAN's face recognition results with GaitNet's gait recognition results. So in this part, we will compare the results of this paper on its own field, gait recognition. As you can see in Figure 2, for almost every movement angle, GaitNet over performed other gait recognition methods or very close to the ones who over performed.

But the important part of these results is GaitNet's performance at 0 and 180 degree angle. In this categories, people walk making 0 or 180 degree angles to the camera. This means they either walk directly straight to the camera or they turn their back and walk straight away from the camera. In both cases, images would consist of high amount of appearance information of people and very less amount of movement information. We can see remarkable effect of disentangled representation learning in these results. Recognition results are much more higher in these categories than the nearest competitor.

4.3 Encoder Choice

In GaitNet, they use a 4-layered Convolutional Neural Network [10] as encoder. Considering popular sign language recognition, action recognition, video classification models, since there are no additional features, researchers aim to extract feature as much as they can from the frames. They use deep networks such as VGG [11], ResNet [12] etc. Comparing that, this 4-layered CNN may not be capable to extract enough features. But also, it need to be considered that inverse of this encoder will be used as decoder. It may be hard to implement and costly for computation.

Here, my suggestion is implementing a middle-depth custom CNN. Also it seems there are already implemented encoder-decoder networks which are inspired from deep convolutional neural networks(i.e. RedNet [13] which is inspired from ResNet). This models deserves a chance.

4.4 Using LSTM

For extracting temporal features, LSTM is one of the most popular networks to use. We can see lots of models consist of CNN and LSTM to consider both spatial and temporal features, especially in cases where inputs are videos. But there is an alternative for LSTM which is called Tranformer Network [14]. It's more popular in natural language processing because of the attention mechanism it provides. Sign Language videos can be considered as sentences or words in the language processing. With this perspective, while implementing this GaitNet model into sign language field, using Transformer instead of LSTM seems like a promising alternative.

5 Conclusion and Future Work

I intended to train the network with BosphorusSign22k [15] dataset. But after a time I realized there was not a ready-to-run code for this model. There was a code in researchers' Github repository, yet the code was spaghetti and there are different versions of it for different conferences and datasets, which made it harder to generalize. So I couldn't train the network by myself. But at this time I found a chance to deeply criticize the paper and compare it with different disentangled representation learning approaches. I realized this approach's weak and strong points.

As a very near future work, I will build my own model with this approach. I will use a different encoder (like ResNet or another custom deeper model). I will use same loss functions, because as far as I observed, the most important contribution of this paper is novel loss functions they propose. I will both try LSTM and Transformer Network to see if we can use the advantage of multi-head attention while generating sign language videos' representations.

References

[1] Zhang, Z., L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan and N. Wang, "Gait Recognition via Disentangled Representation Learning", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4705–4714, 2019.

[2] Tran, L., X. Yin and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1283–1292, 2017.

[3] Liu, Y., Z. Wang, H. Jin and I. Wassell, "Multi-task Adversarial Network for Disentangled Feature Learning", pp. 3743–3751, 06 2018.

[4] Yang, L. and A. Yao, "Disentangling Latent Hands for Image Synthesis and Pose Estimation", pp. 9869–9878, 06 2019.

[5] Zhao, Y., Y. Xiong and D. Lin, "Recognize Actions by Disentangling Components of Dynamics", pp. 6566–6575, 06 2018.

[6] Ganin, Y. and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation", , 09 2014.

[7] Pedro M. Ferreirai, A. R., Diogo Pernes and J. S. Cardoso, "Signer-Independent Sign Language Recognition with Adversarial Neural Networks", *International Journal of Machine Learning and Computing (IJMLC)*, 2019, `publications/journals/2019PedroFerreiraIJMLC.pdf`.

[8] He, K., G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN", , 03 2017.

[9] Hochreiter, S. and J. Schmidhuber, "Long Short-term Memory", *Neural computation*, Vol. 9, pp. 1735–80, 12 1997.

[10] LeCun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard and L. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition", *Neural Computation*, Vol. 1, pp. 541–551, 1989.

[11] Simonyan, K. and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *arXiv 1409.1556*, 09 2014.

[12] He, K., X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", , 2015.

[13] Jiang, J., L. Zheng, F. Luo and Z. Zhang, "RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation", , 06 2018.

[14] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need", , 06 2017.

[15] Özdemir, O., A. Kındıroğlu, N. Camgoz and L. Akarun, "BosphorusSign22k Sign Language Recognition Dataset", , 04 2020.