迁移学习：基于动态视觉传感器（DVS）的图像识别模型在行为识别中的应用

第一作者
刘佳伟
1753070@tongji.edu.cn

第二作者
温鑫
1751918@tongji.edu.cn

Abstract

基于深度学习的方法在图像识别领域取得了巨大成功。行为识别是一项较高级别的任务，很难轻易获得满意的表现。图像识别领域的先进模型经过研究人员多年的精心设计，具有很强的学习能力。一方面，行为识别模型要么较慢（基于双流的方法，2-stream ConvNet）要么不甚准确（基于三维卷积的方法，3D convolotion）。另一方面，图像识别模型具有只能接受 3 帧输入的限制。

为了解决这些问题，我们将动态视觉传感器（DVS）的单帧数据堆叠成 3 帧输入，DVS 可以以惊人的速度捕获运动数据，这些运动数据可以作为训练数据“喂”给图像识别模型。

在我们的论文中，我们发现用 DVS 数据训练的图像识别模型可以在行为识别领域有优秀的表现，并在 UCF-50 DVS 测试集中获得了 100% 的准确率 [8]。同时我们还发现，当基于 DVS 数据训练时，冻结预训练模型前数层权重的迁移学习策略产生了较差的结果。令人惊喜的是，DVS 数据对于小规模模型非常友好，趋向于具有更好的性能。

1. 引言

基于深度学习的图像识别算法在计算机视觉领域已经取得了巨大的成功。近年来，数百种图像识别体系结构（如 ResNet[7]、GoogLeNet[20]）和数据集（例如 ImageNet[3]、CIFAR-10/100[12]）旨在将机器的识别能力提升到人类的水平。另一方面，行为识别是一个更高层次的任务，越来越多的研究人员正试图设计新的模型来提高这一任务的准确性和速度。与图像识别不同，行为识别的研究对象是视频而非图像。因此，其输入是一系列的图像，这意味着相关模型需要处理更多的信息，不仅包含空间成分（spatial component），还有时间成分（temporal component）。这些因素都导致了行为识别领域的研究结果不尽人意。2015 年设计的一种朴素的 ResNet[7] 模型在 ImageNet 测试数据集中轻易达到了 3.57% 的错误率。尽管 2-stream ConvNet[17] 对时间信息进行了仔细的提取，但在 2014 年使用 HMDB-51[13] 数据集测试时，仍然获得了超过 40% 的错误率。

基于以上讨论，我们可以看到，由于图像识别领域的研究更为充分，其中最先进的模型在特征提取和训练方面似乎有更好的表现。将经过良好训练的先进图像识别模型应用于行为识别可能是个不错的想法。然而，这两个任务的输入规模不同。图像识别模型的全部输入就是 3 通道图像，而行为识别模型的输入是一组图像序列，其尺度可以是多种多样的，而不一定为 3。此外，有时我们需要使用多帧密集光流提取时间信息，这意味着需要更多帧图片。因此，将空间成分和时间成分同时作为图像识别模型的输入似乎是不可能的。这就驱使我们探索其他的迁移学习策略。

随着诸多算法在行为识别领域的蓬勃发展，可以产出高质量数据的新型硬件设备也逐渐涌现。传统的 RGB 相机产生的 RGB 图像具有较大的像素块，帧率较低。这种 RGB 帧包含大量冗余的空间信息，但时间信息较少，时间分辨率较低。从 RGB 图像序列中提取时间信息代价巨大，使用最先进的密集光流算法（dense optical flow algorithm[2]），每帧的计算时间高达 1.2 秒到 23.4 秒不等。而新型硬件设备的数据有可能催生一种新型网络结构，从而解决这些问题。以动态视觉传感器（DVS）[14] 为例，它是一种事件驱动的视觉传感器。DVS 只对运动感兴趣，这意味着它能够分

图 1. 一张从 DVS 二进制文件中提取的可视化图像，表示刘佳伟正竖起大拇指赞扬温鑫。

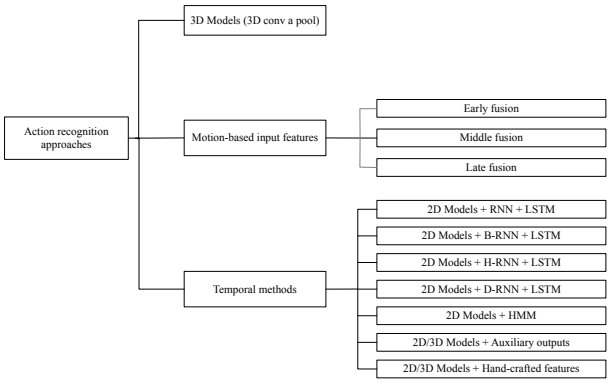

图 2. 主要行为识别方法。

辨当前帧的哪一部分在运动，哪一部分是静止的。这种功能直接在一帧图像中生成运动信息，而不需后续计算。其次，它在性能上也有不小优势，3.6μ 的延迟比光流算法快得多。

与传统的 RGB 相机使用多位二进制数据表示颜色信息的方式不同，DVS 生成一位二进制数据来描述运动，如图 1 所示。

2. 相关方法

多数基于深度学习的行为识别网络结构被设计用于处理时间或空间维的信息。它们主要有 3 种架构：3D-CNN 模型 [10]、多流 2D-CNN 模型 [17]、结合序列模型的 CNN，如图 2 所示。通常，3D-CNN 模型往往更快，而多流 2D-CNN 模型可以获得更高的准确度。

3D-CNN 模型使用固定图像序列作为三维输入，并使用三维卷积核进行三维卷积以提取运动和时空特征。多流 2D-CNN 模型不仅使用时空序列（主输入流），而且还使用其他基于运动的特征（例如密集光流）作为其他流的输入。多流模型需要在输入或输出中融合多流。融合策略主要有 3 种变体：早融合（在将数据馈入模型之前进行融合），中融合（在某些中间层进行融合）和晚融合（为每个流的输出进行融合）。融合主要采用一些朴素的方法，如 SVM，多层感知器，平均等。对于结合序列模型的 CNN，它们通常将一系列图像设置为输入，然后使用 2D 或 3D CNN 来提取特征作为序列模型的输入。序列模型主要是基于 RNN [5] 的模型。为了解决 RNN 短期记忆的缺点，长短期记忆模型（LSTM）[6] 更加常用于 RNN 模型来解决这一问题。还有许多新的 RNN（例如，双向 RNN（BRNN）[16]，分层 RNN（H-RNN）[4] 和差分 RNN（D-RNN）[22]）旨在提升朴素 RNN 的表现。

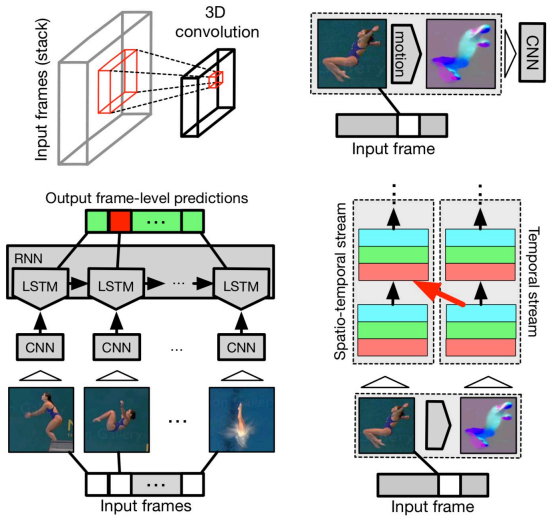

图 3. [1] 不同的神经网络架构和融合策略。左上：三维卷积；右上：运动前处理；左下：基于 LSTM 的序列模型；右下：时空流融合。

图 3 展示了在处理一系列输入图像时，主要的神经网络架构是如何工作的。

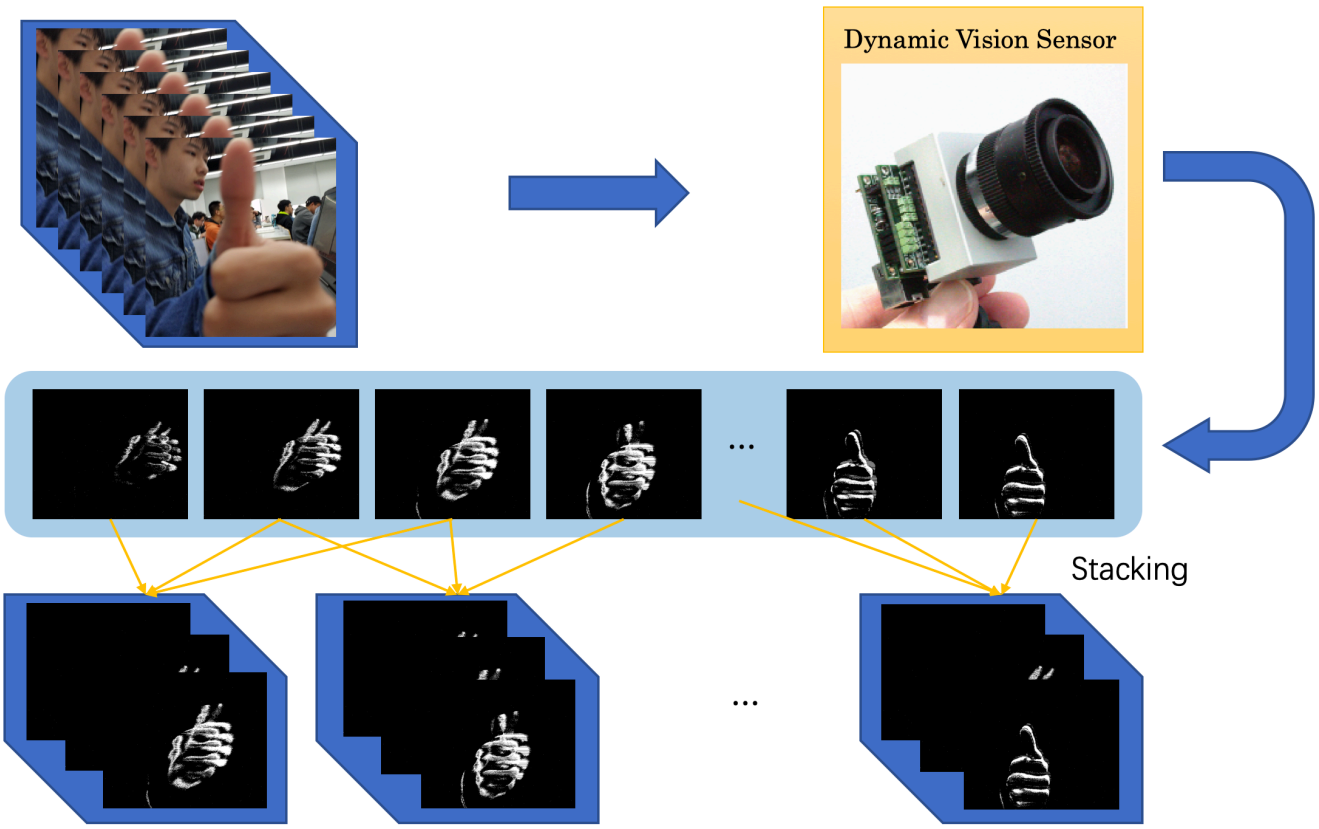

图 4. 帧堆叠的过程。

3. 基于 DVS 数据的图像识别模型

3.1. 帧堆叠策略

为了让图像识别模型能够获取运动和时空数据，我们设计了一个帧叠加策略，如图 4 所示。

首先，我们将 DVS 二进制文件转换为单帧二值图像，其大小为 640×768。这种图像仅包含 2 个值：高（图 1 中的亮像素）和低（图 1 中的暗像素）。然后，我们将连续的 3 帧图像堆叠在一起，这使得它能够作为传统的 3 帧图像识别模型的输入数据。

3.2. 迁移学习策略

在将输入图像堆叠成几个 3 帧块后，我们将使用这些 3 帧块训练预训练模型。

传统的迁移学习方法通常冻结前数层的权值，并更新后几层的权值。这大大减少了反向传播的计算量，提高了训练的速度。然而，在我们的任务中，它很可能收效甚微。传统方法使用带有多通道的 RGB 图像，然而我们的图像帧并不表示 RGB 分量，而是运动和时空信息。因此，我们认为锁定前几层的权重并非好策略。

为了证明我们的观点，我们设计了对比实验来验证权值冻结是否适合这样的转移学习任务。我们将使用从 UCF-50 DVS 数据集中提取的小数据集，比较两种情况下的识别精度和收敛速度。

此外，为了证明基于 DVS 数据的行为识别模型不仅速度快，而且准确性较高，我们将对不同输入特征的结果进行比较。

4. 实验结果

4.1. 实现细节

我们从 UCF-50 DVS 数据集中提取了 3 类数据:

- 弹吉他（PlayingGuitar）
- 跳水（Diving）
- 打篮球（Basketball）

超参数表	
超参数	值
学习率（Learning rate）	1×10^{-3}
优化器（Optimizer）	Adamax[11]
权重衰减（Weight decay）	1×10^{-3}
批大小（Batch size）	32
训练代数（Epochs）	3

表 1. 训练预训练模型时的超参数

行为识别方法	准确率
DVS	97.7%~100%
IDT	91.2%
DT	84.5%
HOG	76.1%
HOF	64.3%

表 2. 不同输入特性的比较。除“DVS”外的实验数据均来自[15]。DVS 的性能测试采用 UCF-50 数据集的部分样本，其余方法的性能测试采用 UCF-50 数据集 [19] 的所有样本。

在进行如图4所示的帧堆叠后，我们得到了 4031 个用于训练的样本和 1728 个用于验证的样本（训练集:验证集 ≈7:3）。我们将这些数据提供给预训练模型。在训练过程中，我们记录了冻结参数的模型和没有冻结参数的模型的损失值（交叉熵损失）。对于冻结策略，我们将模型的线性层（Linear Layer）输出数设为 3，冻结除输出层外的所有参数。

我们实验中的超参数如表1所示，在训练时我们使用相同的超参数。

4.2. 结果

在相同的超参数下，各种 state-of-art 方法的训练结果如表3所示。

根据表 3 我们可以看出：

- 未冻结参数的模型比冻结参数的模型具有更好的表现。
- 未冻结参数的模型的损失相对不易随训练收敛。
- 经训练的预训练模型的高准确率表明，该模型具有较高的工业实用性

各个预训练模型经过 378 次迭代（3 个 epoch）的损失值如图5，6，7和8所示。

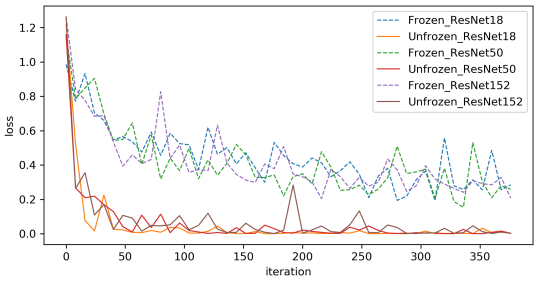

图 5. ResNet 模型经过 378 次迭代的损失值

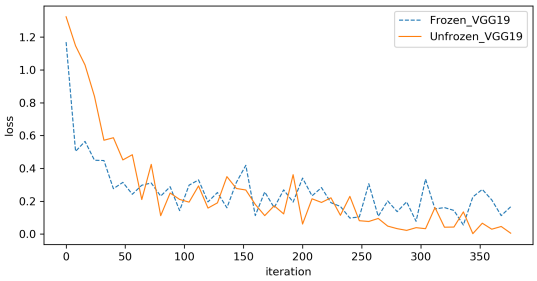

图 6. VGG 模型经过 378 次迭代的损失值

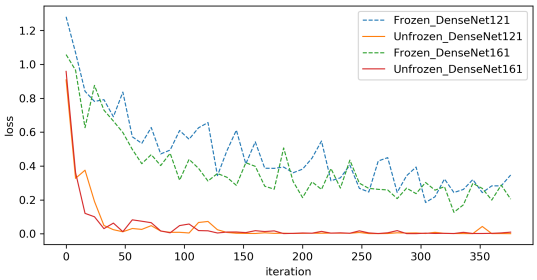

图 7. DenseNet 模型经过 378 次迭代的损失值

至于不同输入特征的比较，我们可以通过观察表2发现：

- DVS 数据有助于模型在保持较高准确率的同时快速收敛。
- 除了高准确率之外，我们还需要指出的是，同样高性能的方法——IDT，需要大量的预处理时间，并且很难做到实时工作。然而 DVS 不仅记录数据迅速，而且节省内存（前者预处理需要数百毫秒，而 DVS 的帧率总是在百万级）。

实验结果		
方法	采用冻结策略的准确率	不采用冻结策略的准确率
ResNet-18[7]	92.2%	100.0%
ResNet-50[7]	94.0%	98.9%
ResNet-152[7]	92.7%	99.5%
VGG-19[18]	94.3%	97.7%
DenseNet-121[9]	91.7%	99.9%
DenseNet-161[9]	94.0%	99.8%
Inception v3[21]	90.3%	99.6%

表 3. 各模型在验证集上的准确率表。

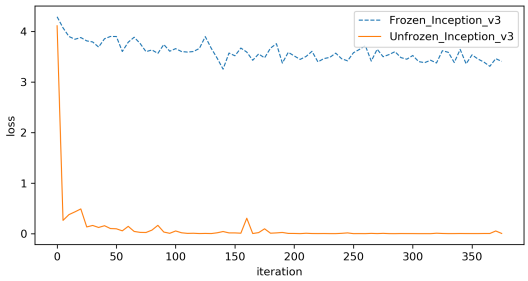

图 8. Inception v3 模型经过 378 次迭代的损失值

5. 总结分析

在这篇论文中，我们得出以下结论：

- 单帧 DVS 数据可以堆叠到 3 帧，并输入到 3 帧图像识别模型中。
- 基于 DVS 数据的图像识别模型在行为识别领域有很好的应用前景。
- 在基于 DVS 数据训练预训练模型时，冻结参数的迁移学习策略是不合适的，因为大多数预训练模型都只擅长提取 RGB 数据，而非 DVS 数据。
- DVS 数据非常简单，因此对小行模型非常友好，往往具有更好的性能。
- ResNet18 在我们的实验中表现最好。

参考文献

[1] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera. A survey on deep learning based approaches for action and gesture recognition in image sequences. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pages 476–483. IEEE, 2017.

[2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In T. Pajdla and J. Matas, editors, Computer Vision - ECCV 2004, pages 25–36, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[4] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1110–1118, 2015.

[5] J. L. Elman. Finding structure in time. Cognitive science, 14(2):179–211, 1990.

[6] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. Learning precise timing with lstm recurrent networks. Journal of machine learning research, 3(Aug):115–143, 2002.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[8] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. Frontiers in Neuroscience, 10:405, 2016.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.

[10] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1):221–231, 2013.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[12] A. Krizhevsky. Convolutional deep belief networks on cifar-10.

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb51: A large video database for human motion recognition. pages 2556–2563, 11 2011.

[14] J. A. Leñero-Bardallo, T. Serrano-Gotarredona, and B. Linares-Barranco. A 3.6μs latency asynchronous frame-free event-driven dynamic-vision-sensor. IEEE Journal of Solid-State Circuits, 46(6):1443–1455, 2011.

[15] J. Miao, X. Jia, R. Mathew, X. Xu, D. Taubman, and C. Qing. Efficient action recognition from compressed depth maps. In 2016 IEEE International Conference on Image Processing (ICIP), pages 16–20. IEEE, 2016.

[16] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. International Journal of Computer Vision, 126(2-4):430–439, 2018.

[17] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pages 568–576, 2014.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[19] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[22] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In Proceedings of the IEEE international conference on computer vision, pages 4041–4049, 2015.