MVFNet: Multi-View Fusion Network for Efficient Video Recognition

Wenhao Wu\textsuperscript{1}, Dongliang He\textsuperscript{1}, Tianwei Lin\textsuperscript{1}, Fu Li\textsuperscript{1}, Chuang Gan\textsuperscript{2}, Errui Ding\textsuperscript{1}

\textsuperscript{1}Department of Computer Vision Technology (VIS), Baidu Inc.
\textsuperscript{2}MIT-IBM Watson AI Lab

AAAI 2021
Task

**Video Recognition:** classify the short clip or untrimmed video into pre-defined class.
Task

**Video Recognition:** classify the short clip or untrimmed video into pre-defined class.

- More than simply recognizing objects
- Complex person-person interaction & people-object interactions
- Videos bring motions
Key Observations

• Efficient spatial-temporal modeling is the key to action recognition

• Classical C2D: temporal modeling unexplored but simple

• 3D CNN, e.g., SlowFast or SlowOnly: effective but expensive

• TSM enables C2D to model temporal relationship at nearly zero cost
  • fixed channel-wise 3x1x1 conv
  • kernel of [0,0,1] for forward shift and [1,0,0] for backward shift
Is TSM our ultimate choice?

NO!
We CAN have better choice:

- From the regular viewpoint of HW-T: TSM can be improved to have arbitrary learnable “shift” kernels
- Why not model relationships from other viewpoints of WT-H and TH-W?
- With careful designing, better effectiveness-efficiency trade-off is possible
Key Innovation

MVFNet = Learnable Temporal "Shift" + Learnable Horizontal "Shift" + Learnable Vertical "Shift"

Viewpoint of HW-T
Viewpoint of WT-H
Viewpoint of TH-W
Why MVFNet will work

MVFNet is a **generalized** architecture of existing frameworks

- $\alpha = 0$, MVFNet specializes to be C2D
- $\alpha = 1, \beta_H = \beta_W = 0$, MVFNet is a channel-wise 3x1x1 Conv version of SlowOnly/C3D
- $\alpha = 1/4, \beta_H = \beta_W = 0$, and half of channel-wise 3x1x1 conv kernels are [0,0,1] and the rest kernels are [1,0,0], then MVFNet becomes TSM
Ablation Experiments

Design choice of $\alpha$: MVFBlock is inserted into res_4 and res_5

| Setting | Sth-sth v1 | Kinetics-400 |
|---------|------------|--------------|
|         | #F | Top-1 | Top-5 | FLOPs | #F | Top-1 | Top-5 | FLOPs |
| $\alpha=0$ | 8  | 17.12 | 43.46 | 32.88G | 4  | 71.87 | 90.02 | 16.44G |
| $\alpha=1/8$ | 8  | 49.74 | 78.09 | 32.90G | 4  | 74.21 | 91.34 | 16.45G |
| $\alpha=1/4$ | 8  | 49.24 | 77.91 | 32.92G | 4  | 74.18 | 91.46 | 16.46G |
| $\alpha=1/2$ | 8  | 50.48 | 79.14 | 32.96G | 4  | 74.21 | 91.42 | 16.48G |
| $\alpha=1$  | 8  | 49.73 | 77.94 | 33.04G | 4  | 73.75 | 91.40 | 16.52G |

(a) Parameter choices of $\alpha$. Backbone: R-50.
Ablation Experiments

Design choice of how many and where MVFBlocks are inserted:
\[ \alpha = 1/2 \text{ and } 1/8 \text{ for Sth-v1 and K400, respectively} \]

| Stages   | Blocks | Sth-sth v1, \( \alpha=1/2 \) | Kinetics-400, \( \alpha=1/8 \) |
|----------|--------|-------------------------------|---------------------------------|
| None     | 0      | #F 17.12 43.46 32.88G         | #F 71.87 90.02 16.44G          |
| res\{5\} | 3      | 8 46.02 75.60 32.90G         | 4 73.46 91.09 16.44G          |
| res\{4,5\} | 9      | 8 **50.48** 79.14 32.96G     | 4 74.21 91.34 16.45G          |
| res\{3,4,5\} | 13    | 8 49.72 78.82 33.04G       | 4 74.08 91.51 16.46G          |
| res\{2,3,4,5\} | 16    | 8 49.95 77.96 33.12G       | 4 **74.22** 91.56 16.47G     |

(b) The number of MVF Blocks inserted into R-50.
Ablation Experiments

Design choice of fusing multiple viewpoints:
\[ \alpha = \frac{1}{2} \text{ and } \frac{1}{8} \text{ for Sth-v1 and K400, respectively;} \]

MVFblocks in res_4, res_5

| Views  | Sth v1 #F Top-1 | K400 #F Top-1 |
|--------|-----------------|---------------|
|        | Sth-v1          | K400          |
| T      | 8 49.13         | 4 73.72       |
| T-H    | 8 49.22         | 4 74.01       |
| T-W    | 8 49.31         | 4 73.88       |
| T-H-W  | 8 50.48         | 4 74.21       |
| T-H-W (S) | 8 47.21 | 4 73.81       |

Fusing multi-view information is beneficial

Channel-wise 3x1x1 temporal / horizontal / vertical convolution must have independent kernels

(c) Study on the different views of MVF module. Backbone: R-50. S denotes weight sharing.
Ablation Experiments

Impact of MVFBlocks when different backbones are used: 
\( \alpha = 1/2 \) and \( 1/8 \) for Sth-v1 and K400, respectively; MVFblocks in res_4, res_5

| Model | Top-1 FLOPs |
|-------|-------------|
| Mb-V2 | 64.4 1.25G  |
| MVF   | 67.5 1.25G  |

| Model | Top-1 FLOPs |
|-------|-------------|
| Mb-V2 | 71.9 16.44G |
| MVF   | 74.2 16.48G |

(e) Advanced backbones for MVFNet on Kinetics-400. 
(f) Different backbones for MVFNet on Kinetics-400. Mb-V2 denotes MobileNet-V2.
Comparison with Similar Variants

$\alpha = 1/2$ and 1/8 for Sth-v1 and K400, respectively; MVFblocks in res_4, res_5

| Method    | Sth v1 Top-1 | K400 Top-1 | FLOPs | Params |
|-----------|--------------|------------|-------|--------|
| C2D       | 17.1         | 71.4       | 32.9G | 24.3M  |
| TSM       | 47.2         | 74.1       | 32.9G | 24.3M  |
| SlowOnly  | -            | 74.9       | 41.9G | 32.4M  |
| CoST*     | -            | -          | 45.8G | 24.3M  |
| MVFNet    | **50.5**     | **76.0**   | 32.9G | 24.3M  |

(d) Study on the effectiveness of MVFNet. Backbone: R-50, 8f input. * indicates our implementation.
| Method                        | Backbone      | Frames × Crops × Clips | GFLOPs   | Top-1  | Top-5  |
|------------------------------|---------------|------------------------|----------|--------|--------|
| I3D (Carreira et al. 2017)  | Inception V1  | 64×N/A×N/A             | 108×N/A  | 72.1%  | 90.3%  |
| S3D-G (Xie et al. 2018)     | Inception V1  | 64×3×10                | 71.4×30  | 74.7%  | 93.4%  |
| TSN (Wang et al. 2016)      | Inception V3  | 25×10×1                | 80×10    | 72.5%  | 90.2%  |
| ECO-RGB_{En} (Zolfaghari et al. 2018) | BNIncep+Res3D-18 | 92×1×1           | 267×1    | 70.0%  | -%     |
| R(2+1)D (Tran et al. 2018)  | ResNet-34     | 32×1×10                | 152×10   | 74.3%  | 91.4%  |
| X3D-M (Feichtenhofer 2020)  | -             | 16×3×10                | 6.2×30   | 76.0%  | 92.3%  |
| STM (Jiang et al. 2019)     | ResNet-50     | 16×3×10                | 67×30    | 73.7%  | 91.6%  |
| TSM (Lin, Gan, and Han 2019) | ResNet-50     | 8×3×10                 | 33×30    | 74.1%  | 91.2%  |
| SlowOnly (Feichtenhofer et al. 2019) | ResNet-50     | 8×3×10                 | 41.9×30  | 74.9%  | 91.5%  |
| TEInet (Liu et al. 2020)    | ResNet-50     | 8×3×10                 | 33×30    | 74.9%  | 91.8%  |
| TEA (Li et al. 2020b)       | ResNet-50     | 8×3×10                 | 33×30    | 75.0%  | 91.8%  |
| Slowfast (Feichtenhofer et al. 2019) | R50+R50       | (4+32)×3×10            | 36.1×30  | 75.6%  | 92.1%  |
| NL+I3D (Wang et al. 2018b)  | ResNet-50     | 32×3×10                | 70.5×30  | 74.9%  | 91.6%  |
| NL+I3D (Wang et al. 2018b)  | ResNet-50     | 128×3×10               | 282×30   | 76.5%  | 92.6%  |
| MVNet                        | ResNet-50     | 8×3×10                 | 32.9×30  | 76.0%  | 92.4%  |
| MVNet                        | ResNet-50     | 16×3×10                | 65.8×30  | 77.0%  | 92.8%  |
| ip-CSN (Tran et al. 2019)   | ResNet-101    | 32×3×10                | 82×30    | 76.7%  | 92.3%  |
| SmallBig (Li et al. 2020a)  | ResNet-101    | 32×3×4                 | 418×12   | 77.4%  | 93.3%  |
| SlowOnly (Feichtenhofer et al. 2019) | ResNet-101     | 16×3×10               | 185×30   | 77.2%  | -%     |
| NL+I3D (Wang et al. 2018b)  | ResNet-101    | 128×3×10               | 359×30   | 77.7%  | 93.3%  |
| Slowfast (Feichtenhofer et al. 2019) | R101+R101     | (8+32)×3×10            | 106×30   | 77.9%  | 93.2%  |
| Slowfast (Feichtenhofer et al. 2019) | R101+R101     | (16+64)×3×10           | 213×30   | 78.9%  | 93.5%  |
| TPN (Yang et al. 2020)      | ResNet-101    | 32×3×10                | 374×30   | 78.9%  | 93.9%  |
| MVNet                        | ResNet-101    | 8×3×10                 | 62.7×30  | 77.4%  | 92.9%  |
| MVNet                        | ResNet-101    | 16×3×10                | 125.4×30 | 78.4%  | 93.4%  |
| MVNet_{En}                   | R101+R101     | (16+8)×3×10            | 188.1×30 | 79.1%  | 93.8%  |
| Method              | Backbone                | Frames × Crops × Clips | FLOPs       | Pre-train | V1 Val Top-1 (%) | V2 Val Top-1 (%) |
|---------------------|-------------------------|------------------------|-------------|-----------|------------------|------------------|
| I3D (Wang et al. 2018) | 3D ResNet50             | 32 × 3 × 2             | 153G × 3 × 2| ImageNet   | 41.6             | -                |
| NL I3D (Wang et al. 2018) | 3D ResNet50             |                        | 168G × 3 × 2| +          | 44.4             | -                |
| NL I3D+GCN (Wang et al. 2018) | 3D ResNet50+GCN       |                        | 303G × 3 × 2| K400       | 46.1             | -                |
| ECO (Zolfaghari et al. 2018) | BNIncep+3D Res18         | 81 × 1 × 1             | 32G × 1 × 1 | K400       | 39.6             | -                |
| ECO_{En} (Zolfaghari et al. 2018) | Inception             | 92 × 1 × 1             | 267G × 1 × 1| K400       | 46.4             | -                |
| S3D-G (Xie et al. 2018) | Inception             | 64 × 1 × 1             | 71G × 1 × 1 | K400       | 48.2             | -                |
| TSN (Wang et al. 2016)  | ResNet50                | 8 × 3 × 2              | 33G × 3 × 2 | ImageNet   | 20.5             | 30.4             |
| TSM (Lin et al. 2019)  | ResNet50                | 8 × 3 × 2              | 33G × 3 × 2 | ImageNet   | 47.2             | 61.2             |
| STM (Jiang et al. 2019) | ResNet50                | 8 × 3 × 10             | 33G × 3 × 10| ImageNet   | 49.2             | 62.3             |
| TEInet (Liu et al. 2020) | ResNet50                | 8 × 3 × 10             | 33G × 3 × 10| ImageNet   | 49.2             | 62.3             |
| TEA (Li et al. 2020b)   | ResNet50                | 8 × 3 × 10             | 35G × 3 × 10| ImageNet   | 51.0             | 64.7             |
| MVFNet                | ResNet50                | 8 × 1 × 1              | 33G × 1 × 1 | ImageNet   | 48.8             | 60.8             |
|                      |                         | 8 × 3 × 2              | 33G × 3 × 2 | ImageNet   | 50.5             | 63.5             |
|                      |                         | 16 × 1 × 1             | 66G × 1 × 1 | ImageNet   | 51.0             | 62.9             |
|                      |                         | 16 × 3 × 2             | 66G × 3 × 2 | ImageNet   | 52.6             | 65.2             |
|                      |                         | (16+8) × 3 × 2         | 99G × 3 × 2 |            | 54.0             | 66.3             |
Mean class accuracy of RGB modality is reported, RGB models are pretrained on Kinetics400

| Method   | Backbone           | UCF-101 | HMDB-51 |
|----------|--------------------|---------|---------|
| ECO\textsubscript{En} | BNIncep+Res3D-18   | 94.8%   | 72.4%   |
| ARTNet   | ResNet-18          | 94.3%   | 70.9%   |
| I3D      | Inception V1       | 95.6%   | 74.8%   |
| R(2+1)D | Inception V1       | 96.8%   | 74.5%   |
| S3D-G    | Inception V1       | 96.8%   | 75.9%   |
| TSN      | BNInception        | 91.1%   | -       |
| StNet    | ResNet-50          | 93.5%   | -       |
| TSM      | ResNet-50          | 95.9%   | 73.5%   |
| STM      | ResNet-50          | 96.2%   | 72.2%   |
| TEINet   | ResNet-50          | 96.7%   | 72.1%   |
| MVFNet   | ResNet-50          | 96.6%   | 75.7%   |
Conclusion

• Upgrading fixed shift kernels of TSM to be learnable is more flexible

• Relationship modeling from multiple viewpoints is a strong boost

• MVFNet consistently outperforms existing solutions on Kinetics400, Something-Something-v1/v2

• Codes & models will be available

  https://github.com/whwu95/MVFNet
Thank you!

MVFNet: Multi-View Fusion Network for Efficient Video Recognition

Contact: Wenhao Wu
wuwenhao17@mails.ucas.edu.cn
https://github.com/whwu95/MVFNet