Axon AI's Solution to the 2nd YouTube-8M Video Understanding Challenge

Choongyeun Cho, Benjamin Antin, Sanchit Arora, Shwan Ashrafi, Peilin Duan, **Dang The Huynh**, Lee James, Hang Tuan Nguyen, Moji Solgi, Cuong Van Than

Axon AI

YouTube-8M challenge

- **6.1M** video IDs
- **3862** classes
- Multi-class Multi-label video classification.
- Evaluation metric: Global Average Precision (GAP).

$$GAP = \sum_{i=1}^{N} p(i)\Delta(i)$$

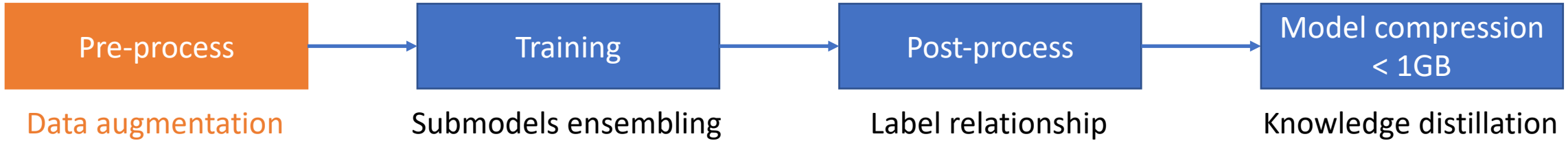
Challenge strategy

Data augmentation

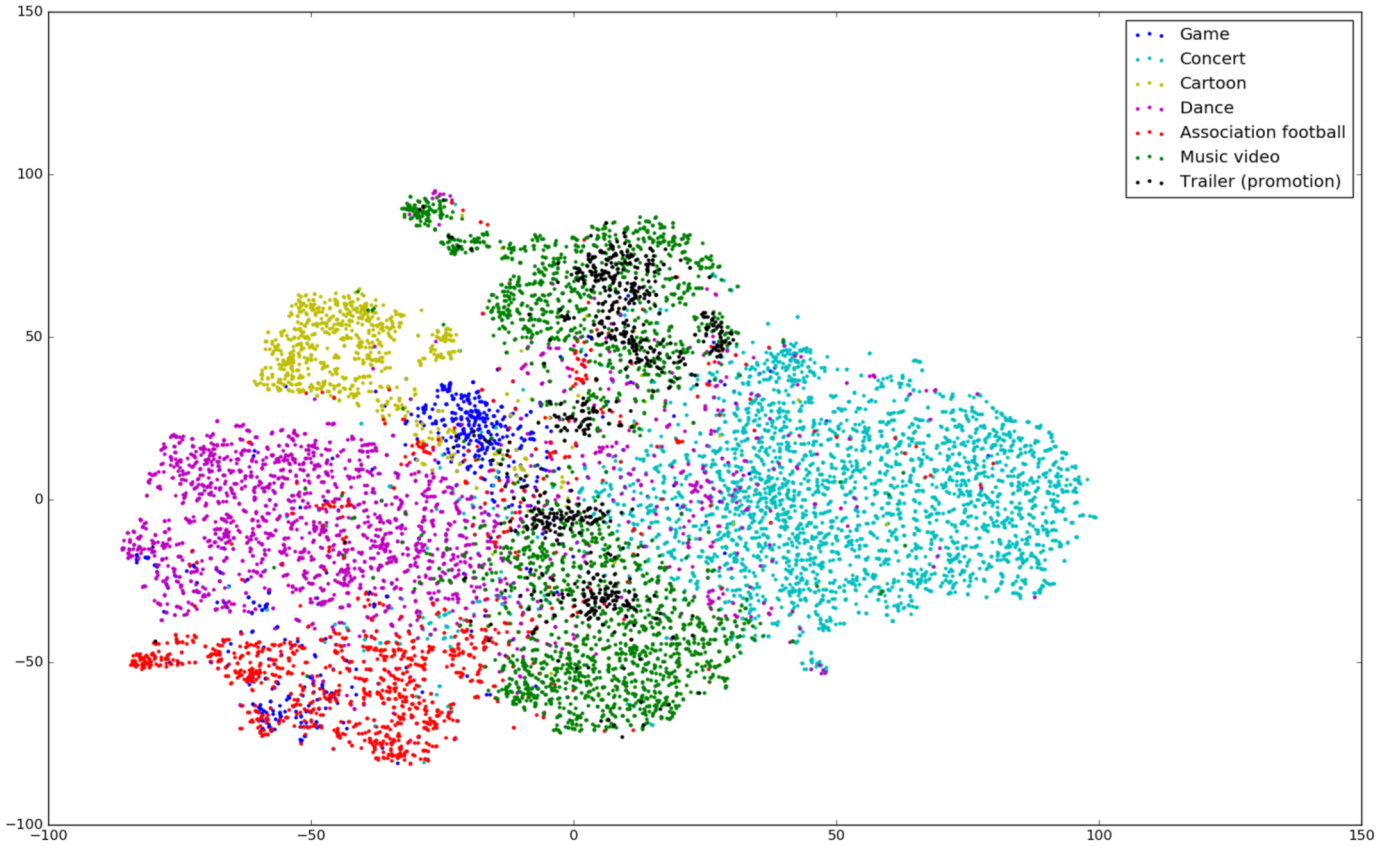
Submodels ensembling

Label relationship

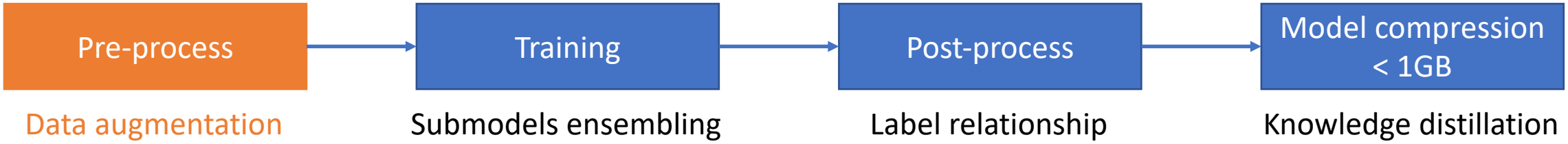
Knowledge distillation



Observation: videos associated with a same label form a cluster, whereas others are separated to some degree.



TSNE plot of visual features for a few selected classes



Data augmentation is for visual features only, by adding small noise to the feature vector.

$$x_i' = x_i + \gamma Z, Z \sim \mathcal{N}(0, \sigma^2)$$

Over-sampling: a single label with **less** than 10^4 samples. For each sample x_i, find K nearest neighbors x_j (L2-distance)

- Interpolation:

$$x_i' = x_i + \lambda_i(x_j - x_i)$$

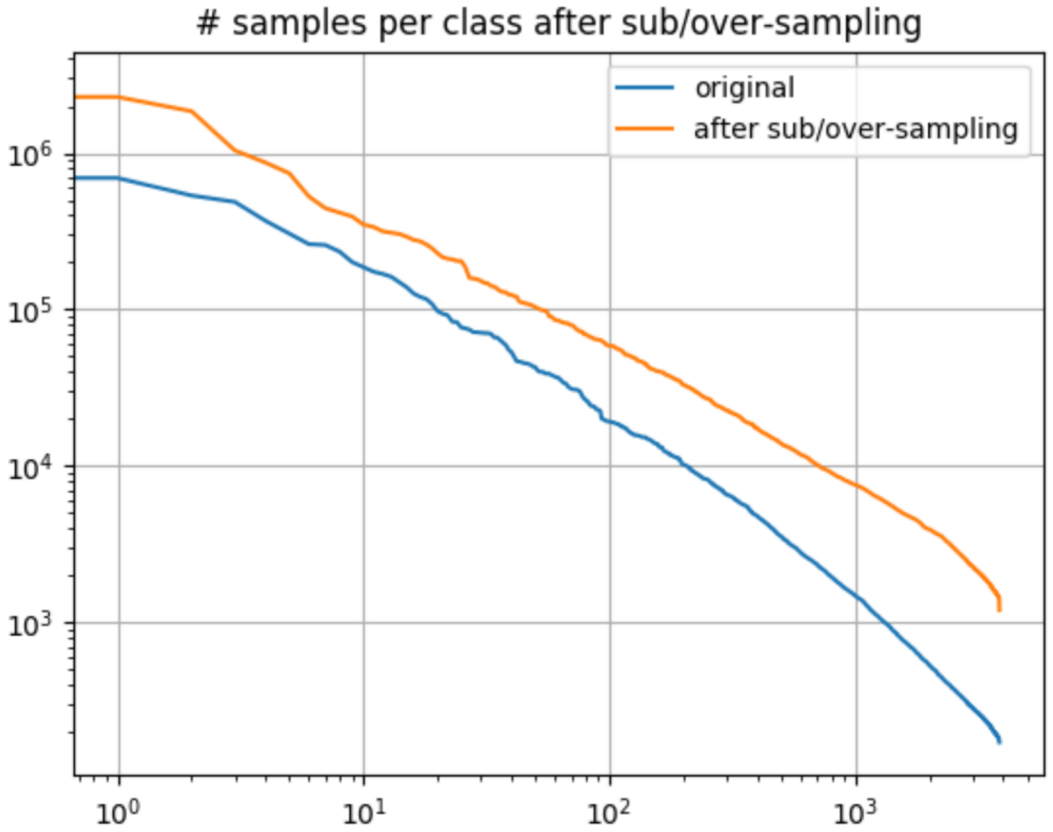
- Extrapolation:

$$x_i' = x_i + \lambda_e(x_i - x_j)$$

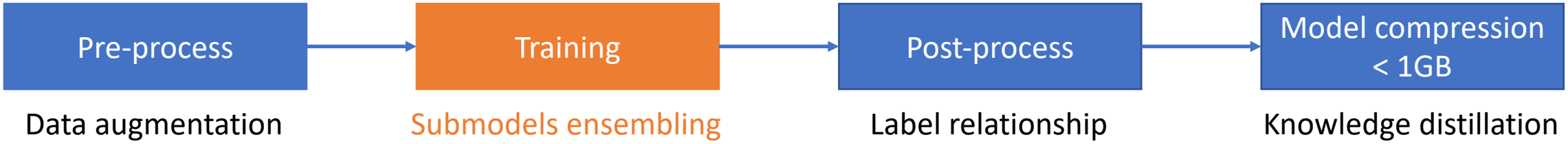
Sub-sampling (random-sampling): a single label with **more** than 10^4 samples.

Data augmentation | Submodels ensembling | Label relationship | Knowledge distillation

Before data augmentation: 5,001,275; After data augmentation: **23,590,464** (472%)



Label counts before and after data augmentation in feature space



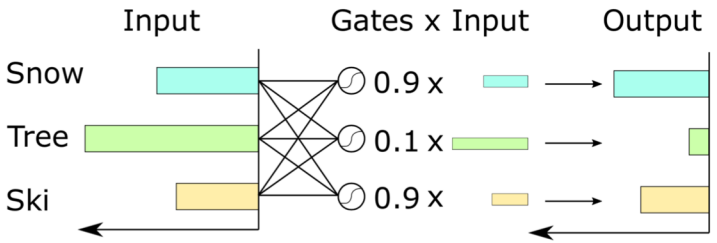
Identify powerful and efficient baseline models (last-year winners) regardless of their model sizes:

- Training set: **train????.tfrecord** + **validate???[0-4,6-9].tfrecord**
- Validation set: **validate???5.tfrecord**
- Baseline models:

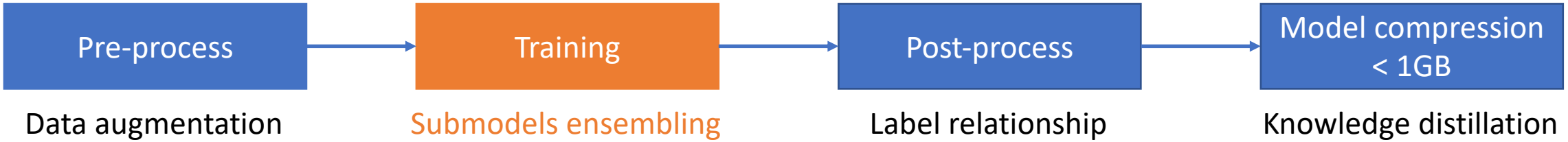
Model family	Brief description
Learnable Pooling	Gated NetVLAD with 256 clusters
Learnable Pooling	Gated NetFV with 128 clusters
Bag of Words	Gated soft-DBoW with 4096 clusters
Bag of Words	Soft-DBoW with 8000 clusters
Learnable Pooling	Gated NetRVLAD with 256 clusters
RNN	Gated recurrent unit (GRU) with 2 layers and 1024 cells per layer
RNN	LSTM with 2 layers and 1024 cells per layer

Context gating:

$$y = \sigma(W \cdot x + b) \circ x$$



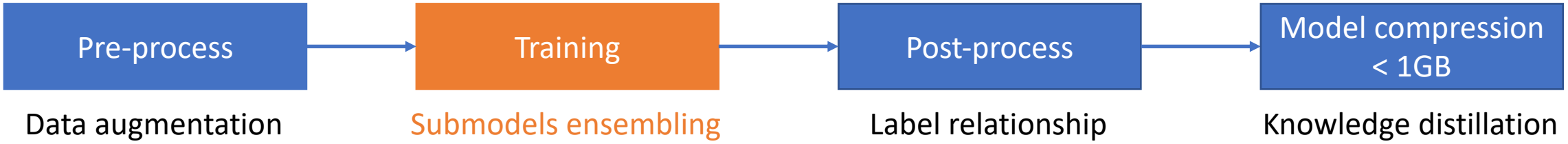
All models are kept original, trained with Adam optimizer (LR = 0.0002 with exponential decay 0.8 for every 4M samples)



Approach: combine efficient submodels to have a better performance.

Experiment	Test GAP (%)
Single baseline model (gated NetVLAD)	85.75 (Val GAP)
Single gated NetVLAD model + video-level MoE model trained with augmented dataset in feature space	85.98 (Val GAP)
Single gated NetVLAD model + regularized DNN exploiting label relationship	87.88 (Val GAP)
A simple average ensembling of all of the 7 models	88.27
A simple average ensembling of two sets of all of the 7 models (14 models in total)	88.62
Ensembled using learned weights	88.73
Distilled model	87.29

GAP performance per experiment

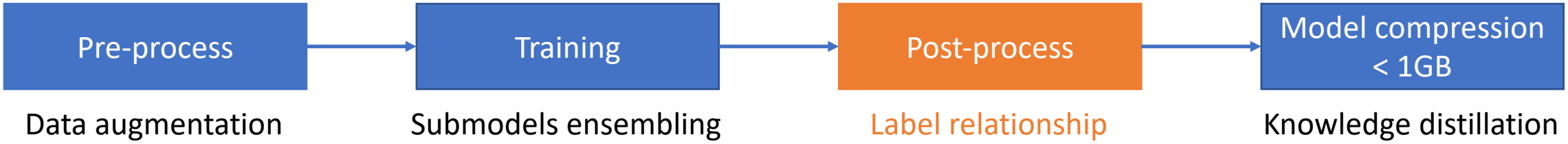


Approach: … but how to combine?

- Simple average
- **Per-model linearly-weighted average**
- Per-model and per-class linearly-weighted average

Model	Weight
Gated NetVLAD	0.2367
Gated NetFV	0.1508
Gated soft-DBoW	0.1590
Soft-DBoW	0.1000
Gated NetRVLAD	0.1968
GRU	0.1306
LSTM	0.0621

Learned weights for 7 baseline models.



- Exploit correlation and diversity of video label relationship, by using an extra regularization term.

$$\min_{\boldsymbol{W},\Omega} \sum_{i=1}^{N} l(f(x_i), y_i) + \frac{\lambda_1}{2} \sum_{l=1}^{L-1} ||\boldsymbol{W}_l||_F^2 + \lambda_2 \cdot \mathrm{tr}(\boldsymbol{W}_{L-1} \Omega^{-1} \boldsymbol{W}_{L-1}^T)$$

$$\text{s.t. } \Omega \succeq 0$$

where the optimal $\Omega \in \mathbb{R}^{C \times C}$ can be derived as:

$$\Omega = \frac{(\boldsymbol{W}_{L-1}^T \boldsymbol{W}_{L-1})^{\frac{1}{2}}}{\mathrm{tr}((\boldsymbol{W}_{L-1}^T \boldsymbol{W}_{L-1})^{\frac{1}{2}})}$$

Data augmentation | Submodels ensembling | Label relationship | Knowledge distillation

Approach: training a student model (< 1GB) based on a teacher model (ensemble of 7 baseline models).

- Student model: **NetVLAD** with the last FC of 800 hidden weights (instead of 1024)
- Loss function: weighted sum of two cross-entropy losses (with teacher model prediction $\widetilde{p}$ and with ground truth q)

$$L = \lambda \cdot CE(p, \widetilde{p}) + (1 - \lambda) \cdot CE(p, q)$$

Experiment	Test GAP (%)
Ensembled using learned weights	88.73
Distilled model	87.29

GAP performance after knowledge distillation

References

- **Google: Google cloud & youtube-7m video understanding challenge** (2017) https://www.kaggle.com/c/youtube8m.
- Miech, A., Laptev, I., Sivic, J.: **Learnable pooling with context gating for video classification** (2017) Computer Vision and Pattern Recognition (CVPR) Youtube-8M Workshop.
- Hinton, G., Vinyals, O., Dean, J.: **Distilling the knowledge in a neural network** (2014) NIPS 2014 Deep Learning Workshop.
- Miech, A., Laptev, I., Sivic, J.: **https://github.com/antoine77340/loupe**
- DeVries, T., Taylor, G.W.: **Dataset augmentation in feature space** (2017) https://arxiv.org/abs/1702.05538.
- Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: **Objects in context** (2007) IEEE ICCV.
- Bengio, S., Dean, J., Erhan, D., Ie, E., Le, Q., Rabinovich, A., Shlens, J., Singer, Y.: **Using web co-occurrence statistics for improving image categorization** (2013) Computer Vision and Pattern Recognition (CVPR).
- Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengi, S., Li, Y., Neven, H., Adam, H.: **Large-scale object classification using label relation graphs** (2014) ECCV.
- Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: **Exploiting feature and class relationships in video categorization with regularized deep neural networks** (2018) IEEE TPAMI 40.2.
- Bober-Irizar, M., Husain, S., Ong, E.J., Bober, M.: **Cultivating dnn diversity for large scale video labelling** (2017) Computer Vision and Pattern Recognition (CVPR) Youtube-8M Workshop.

Thank you