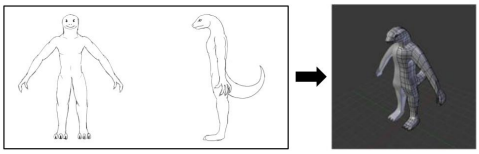
2D Character Sketches to 3D Models

Andrew Yu <acyu@stanford.edu>

Introduction

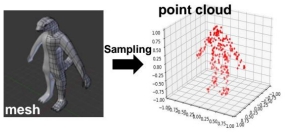


Concept: convert grayscale front + side 2D character sketch into 3D model using deep learning

→ **Motivation:** speed up asset generation for video games or animation

→ **Prior Work:** mostly single image reconstruction [2-6], whereas here, the hope is that 2 views will be more generalizable

3D Model Format



3D Models represented by **point cloud** set of (x,y,z):

→ Typical 3D models in mesh format (edges and vertices)

→ But mesh requires a graph loss (difficult to define)

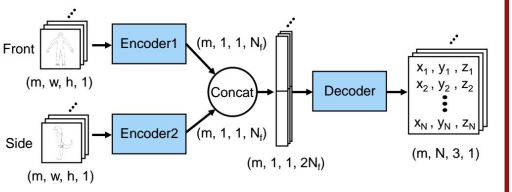
→ Point cloud loss is simpler

Define: **Chamfer distance (CD)** loss between two point sets $S, \hat{S} \in \mathbb{R}^3$ [3]

$$\mathcal{L}_{\text{CD}}(S,\hat{S}) = \sum_{y \in S} \min_{\hat{y} \in \hat{S}} \{\|y-\hat{y}\|_2^2\} + \sum_{\hat{y} \in \hat{S}} \min_{y \in S} \{\|y-\hat{y}\|_2^2\}$$

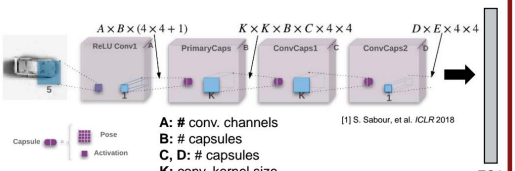
→ For $S, \hat{S}$ CD calculates distance to **nearest neighbor** point in other set

Network Architecture



Encoder & Decoder Architectures

Encoder: Modified Matrix Capsule Network (CapsNet)



[1] S. Sabour, et al. *ICLR* 2018

A: # conv. channels
B: # capsules
C, D: # capsules
K: conv. kernel size

modification to [1]

CapsNet like a convolutional network except:

- activations are matrices ("capsules") instead of scalars
- non-linearity between layers from a trained dynamic routing
  - network "figures out" how to route activation outputs to next layer

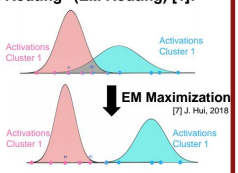
Routing by "Expectation-Maximization Routing" (EM Routing) [1]:

procedure EM ROUTING(a, V)
$\forall i \in \Omega_L, j \in \Omega_{L+1}$: $R_{ij} \leftarrow 1/|\Omega_{L+1}|$
for t iterations **do**
$\forall j \in \Omega_{L+1}$: M-STEP($a, R, V, j$)
$\forall i \in \Omega_L$: E-STEP(μ, σ, a, V, i)
return a, M

a, M : activation and pose
Ω_L : capsules in layer L
R_{ij} : assignment probabilities
β_u, β_a, *cost* : related to gaussian cost

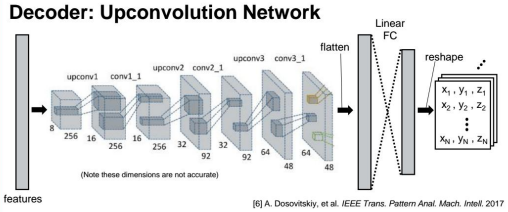
procedure M-STEP(a, R, V, j) ▷ for one higher-level capsule, j
$\forall i \in \Omega_L$: $R_{ij} \leftarrow R_{ij} * a_i$
$\forall h$: $\mu_j^h \leftarrow \frac{\sum_i R_{ij} V_{ij}^h}{\sum_i R_{ij}}$
$\forall h$: $(\sigma_j^h)^2 \leftarrow \frac{\sum_i R_{ij} (V_{ij}^h - \mu_j^h)^2}{\sum_i R_{ij}}$
$cost^h \leftarrow (\beta_u + log(\sigma_j^h)) \sum_i R_{ij}$
$a_j \leftarrow logistic(\lambda(\beta_a - \sum_h cost^h))$

procedure E-STEP(μ, σ, a, V, i) ▷ for one lower-level capsule, i
$\forall j \in \Omega_{L+1}$: $p_j \leftarrow \frac{1}{\sqrt{\prod_h^H 2\pi(\sigma_j^h)^2}} \exp\left(-\sum_h^H \frac{(V_{ij}^h - \mu_j^h)^2}{2(\sigma_j^h)^2}\right)$
$\forall j \in \Omega_{L+1}$: $R_{ij} \leftarrow \frac{a_j p_j}{\sum_{k \in \Omega_{L+1}} a_k p_k}$



- EM iteratively fits gaussian to feature activations
- Routes low-level features to be part of a higher level feature

Decoder: Upconvolution Network

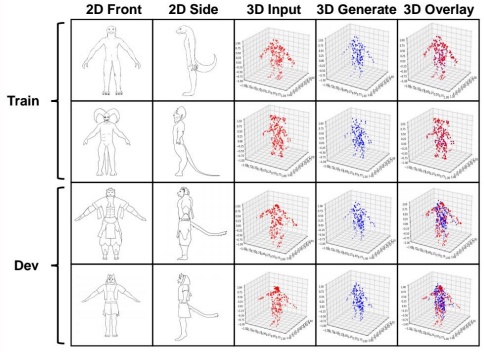


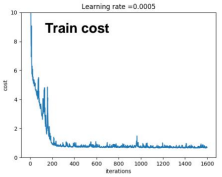
[6] A. Dosovitskiy, et al. *IEEE Trans. Pattern Anal. Mach. Intell.* 2017

Decoder series of (upconv → conv) layers

- upconv is a "transpose" of a (conv → max pool) encoding layer
- maps lower dimensional features to higher dimensional output
- finish with linear fully connected layer to scale or spread output

Results





- Cost saturates, need more hyperparameter tuning and training examples
- Dev predicted output is overfitting to lizards in training set
- Predicted points tend to cluster near (0,0,0), decoder needs to be tuned to give better spread over volume

General Conclusions

- **2D → 3D point cloud feasible, but <u>not very good</u>**
  - **Unordered output data bad for generation**
  - **Scales poorly with number of output points**
  - **Still need to do 3D mesh reconstruction (annoying)**
- **Future: End-to-end 2D → mesh approaches more desirable [2]**

References

[1] S. Sabour, et al., "Matrix capsules with EM routing," *ICLR 2018*, no. 2011, pp. 1–12, 2018.
[2] J. K. Pontes, et al., "Image2Mesh: A Learning Framework for Single Image 3D Reconstruction," 2017.
[3] H. Fan, et al., "A Point Set Generation Network for 3D Object Reconstruction from a Single Image," 2016.
[4] C. B. Choy, et al., "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction," vol. 1, pp. 1–17, 2016.
[5] P. Buchanan, et al., "Automatic single-view character model reconstruction," *Proc. - Sketch-Based Interfaces Model. SBIM 2013 - Part Expressive 2013*, pp. 5–14, 2013.
[6] A. Dosovitskiy, et al., "Learning to Generate Chairs, Tables and Cars with Convolutional Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 692–705, 2017.
[7] J. Hui, https://jhui.github.io/2017/11/14/Matrix-Capsules-with-EM-routing-Capsule-Network/, 2018

Acknowledgements: Xingyu Liu for TA project mentorship, Russell Kaplan for AWS GPU credits & advice, Andrew Ng for teaching class.