Theoretical part

Barabanshchikova Polina, Protasov Dmitry, Shulgan Nikita

MIPT

25 октября 2022 г.

Формулировка задачи

Теоретическая модель

Пусть $V = (v_1, \ldots, v_N)$ — видеопоток, то есть $v_i \in \mathbb{R}^{K_v \times C \times H \times W}$, где K_v — число кадров в t секунд, а C, H, W — количество каналов, высота и ширина изображения. И пусть $F = (f_1, \ldots, f_N)$ — fMRI сигнал, состоящий из последовательности измерений $f_i \in \mathbb{R}^{K_f \times X \times Y \times Z}$, где K_f — число измерений за t секунд, а X, Y, Z — размерность одного измерения. Также для каждой пары (V, F) известна метка испытуемого $u \in \{1, \ldots, M\}$.

Задача состоит в предсказании fMRI сигнала F по паре (V, u).

Формально, необходимо построить отображение H, которое видеоряду V и участнику u сопоставляет сигнал $\tilde{F}$, причём

$$p(\tilde{F}|V, u) = p(F|V, u).$$

Формулировка задачи

Мы будем работать в предположении, что элемент f_t зависит только от текущего кадра v_t, предыдущих L кадров v_{t-L}^{t-1} и последних L сгенерированных элементов f_{t-L}^{t-1}. В таком случае возможна факторизация

$$p(\tilde{F}|V,u) = \prod_{n=1}^{N} p(f_t|f_{t-L}^{t-1}, v_{t-L}^{t}, u).$$

Теоретическое решение

Conditional GAN

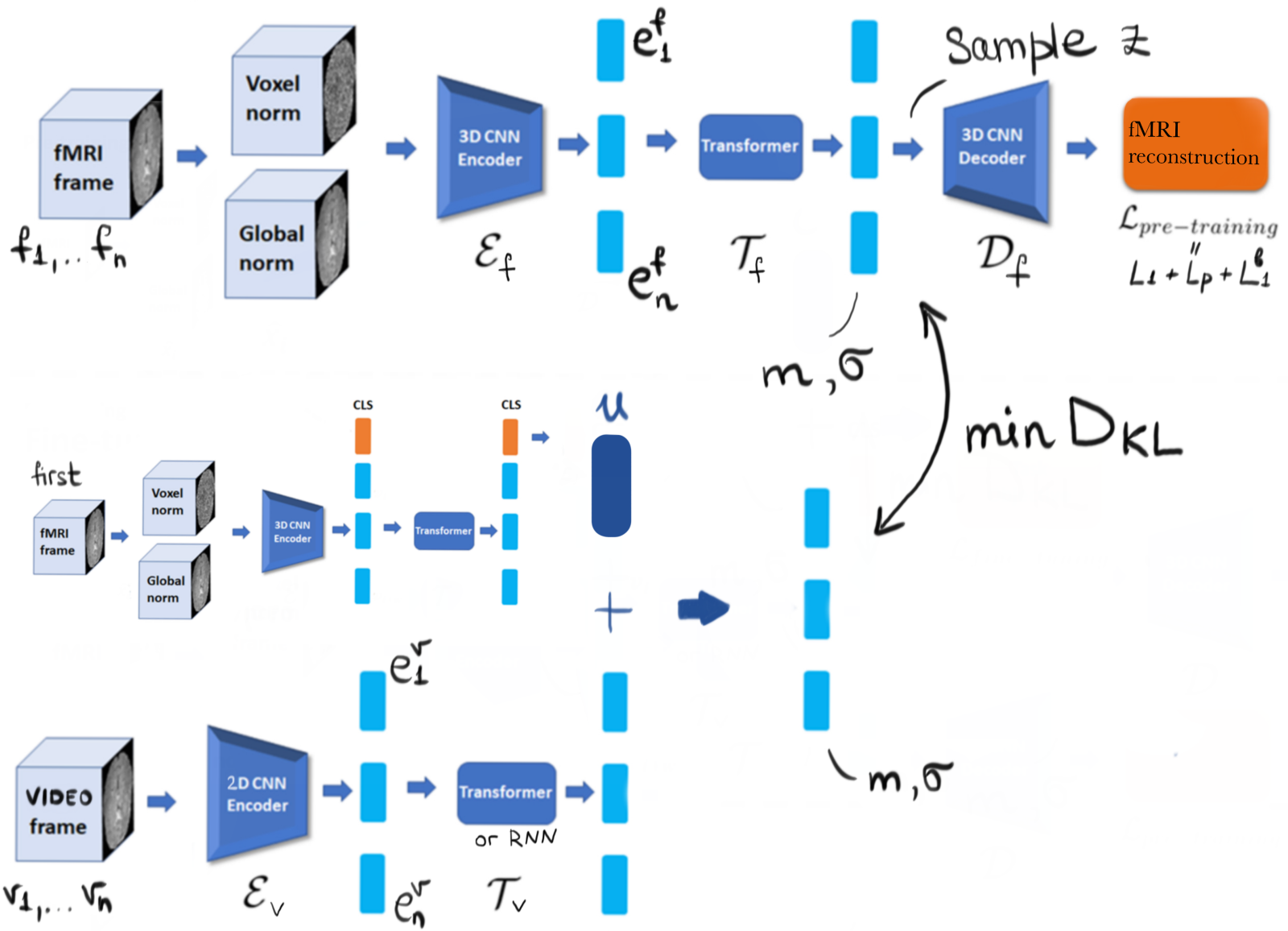
$$\min_G \max_D E_{(F,V,u)} \log D(F, V, u) + E_{V,u} \log(1 - D(G(V, u), V, u)).$$

Conditional VAE

$$\mathcal{L}(\theta, \psi) = E_{q_\psi} \log p_\theta(F|z, u) - D_{KL}[q_\psi(z|F, u)||p_\theta(z|u)],$$

$$\mathcal{L}(\theta, \psi) = E_{q_{\psi_f}} \log p_\theta(F|z, u) - D_{KL}[q_{\psi_f}(z|F, u)||q_{\psi_v}(z|V, u)].$$

Модель cVAE


fMRI frame
Voxel norm
Global norm
3D CNN Encoder
Transformer
Sample z
3D CNN Decoder
fMRI reconstruction
$\mathcal{L}_{pre-training}$
$L_1 + L_p + L_1^6$
$\mathcal{E}_f$
$\mathcal{T}_f$
$\mathcal{D}_f$
m, σ
min D_{KL}
first
CLS
u
VIDEO frame
2D CNN Encoder
Transformer
or RNN
$\mathcal{E}_v$
$\mathcal{T}_v$

Модель cVAE

Формальное описание

Модель состоит из двух частей: автоэнкодера $(\mathcal{E}_f, \mathcal{T}_f, \mathcal{D}_f)$ и сети $(\mathcal{E}_v, \mathcal{T}_v)$. Автоэнкодер тренируется из предобученного состояния $(\mathcal{E}_f^o, \mathcal{T}_f^o, \mathcal{D}_f^o)$.

Входные данные имеют вид $(\overline{f}^o, \overline{f}, \overline{v})$, где $f_i^o, f_i \in \mathbb{R}^{K_f \times X \times Y \times Z}$, а $v_i \in \mathbb{R}^{K_v C \times H \times W}$.

Сначала сигнал fMRI нормируется и нормированные копии конкатенируются по оси каналов: $\hat{f}_i^o, \hat{f}_i \in \mathbb{R}^{3K_f \times X \times Y \times Z}$. Далее энкодеры, состоящие из 3D CNN, обрабатывают каждое измерение независимо: $e_i^f = \mathcal{E}_f(f_i), e_i^o = \mathcal{E}_f^o(f_i^o)$. Полученные эмбеддинги подаются на вход трансформеру, который учитывает временные зависимости между измерениями. Результат применения трансформера $\mathcal{T}_f$ к $\overline{e}_f$ — это множество пар (m, σ) для каждого элемента эмбеддинга.

Модель cVAE

Продолжение

Вектор z семплируется из распределения $\mathcal{N}(m, \sigma)$ и подаётся на вход декодеру, которые восстанавливает fMRI. Качество восстановления сигнала контроллируется функцией потерь из статьи "Self-Supervised Transformers for fMRI representation".
По эмбеддингу $\overline{e}_f^o$ трансформер $\mathcal{T}_f^o$ строит вектор пользователя u. Энкодер $\mathcal{E}_v$, состоящий из 2D CNN, строит эмбеддинги кадров v_i видеоряда $\overline{v}$: $e_i^v = \mathcal{E}_v(v_i)$. Далее сеть $\mathcal{T}_v$ используется для преобразования эмбеддингов и вектора u в множество пар (m_v, σ_v) той же размерности, что и выход трансформера $\mathcal{T}_f$. Минимизируется KL дивергенция между $\mathcal{N}(m, \sigma)$ и $\mathcal{N}(m_v, \sigma_v)$.