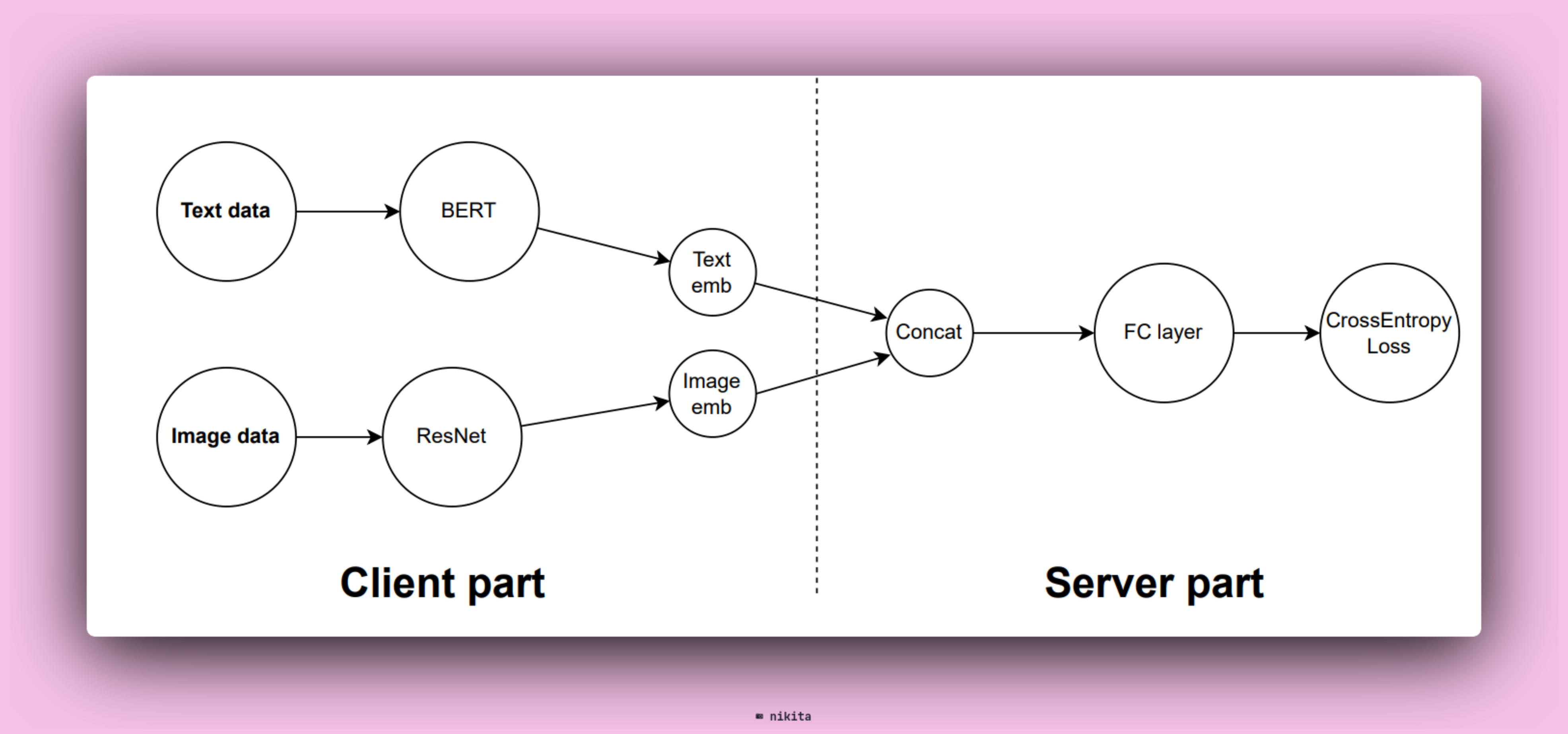
backprop

Architecture of the network:



Loss:

$$L(x, y) = -\sum_{i=1}^{m} y_i \log\left(\frac{\exp(x_i)}{\sum_{j=1}^{\|x\|} \exp(x_j)}\right)$$

- T - text data, I - image data
- F - weights matrix of FC layer
- E_t & E_i - embeddings of text and image

Total formula

$$L\left[F^T(BERT(T) \oplus ResNet(I))\right]$$

Backprop:

$$\frac{\partial L}{\partial L} = 1$$

Let

$$F^T(BERT(T) \oplus ResNet(I)) = FE$$

$$S = \frac{\exp(FE_i)}{\sum_{j=1}^{\|FE\|} \exp(FE_j)}$$

Gradient from the loss fuction to FC layer:

$$\frac{\partial L}{\partial FC} = S - y$$

Gradient of Loss with respect to concatenated Embeddings (E):

$$\frac{\partial L}{\partial E} = \frac{\partial L}{\partial FC} \times \frac{\partial FC}{\partial E} = (S - y) \times F^T$$

Resulting gradient gradient

$$\frac{\partial L}{\partial E} = \left(\frac{\exp(FE_i)}{\sum_{j=1}^{\|FE\|} \exp(FE_j)} - y \right) \times F^T$$

And this gradient comes to the users, to optimie parameters of the models(BERT & ResNet).