Self-supervised learning (SSL)

Self-supervision

$\{(x_i, y_i)\}_{i=1}^{\ell}$

$$\underbrace{\frac{1}{\ell}\sum_{i=1}^{\ell} L(f(x_i), y_i) \to \min_f}_{\text{supervised}}$$

Pre-train without manual labels

- Embeddings
- Language models

GPT (Generative Pre-training of Transformers)

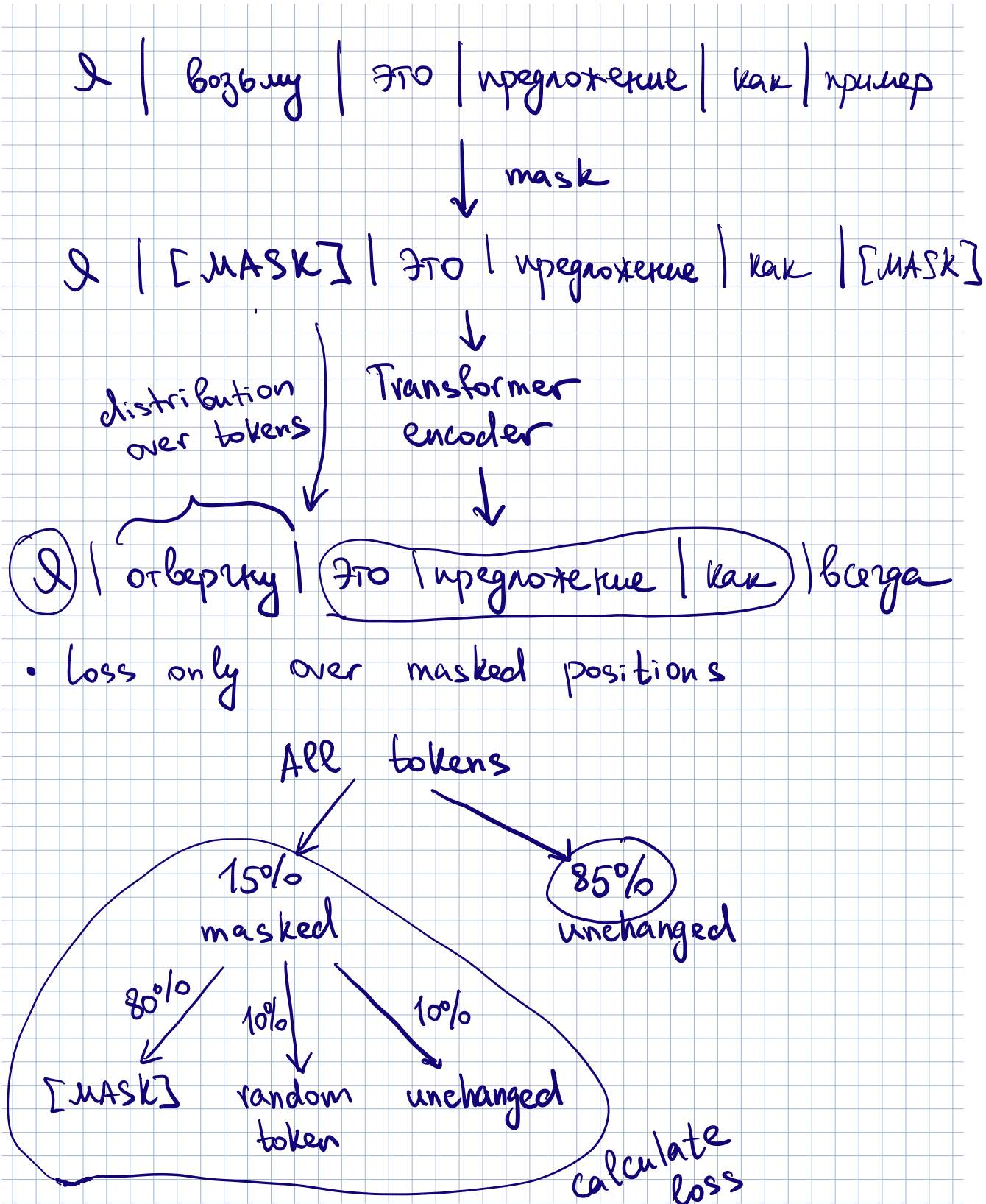
Architecture – transformer decoder

Task – language modeling

BERT (Bidirectional Encoder Representations from Transformers)

Architecture – transformer encoder

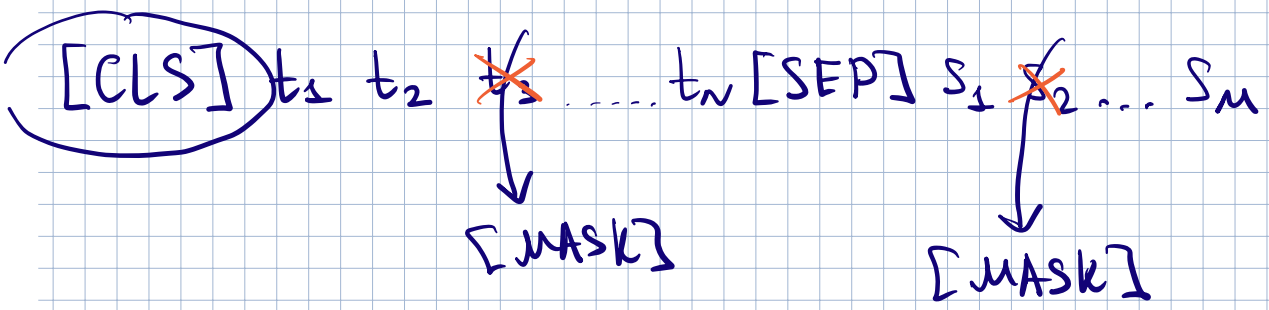
Task – Masked Language Modeling


я | возьму | это | предложение | как | пример
mask
я | [MASK] | это | предложение | как | [MASK]
distribution over tokens
Transformer encoder
я | отвертку | это | предложение | как | всегда
• Loss only over masked positions
All tokens
15% masked
85% unchanged
80%
10%
10%
[MASK]
random token
unchanged
calculate loss

BERT loss = $L_{MLM} + L_{NSP}$

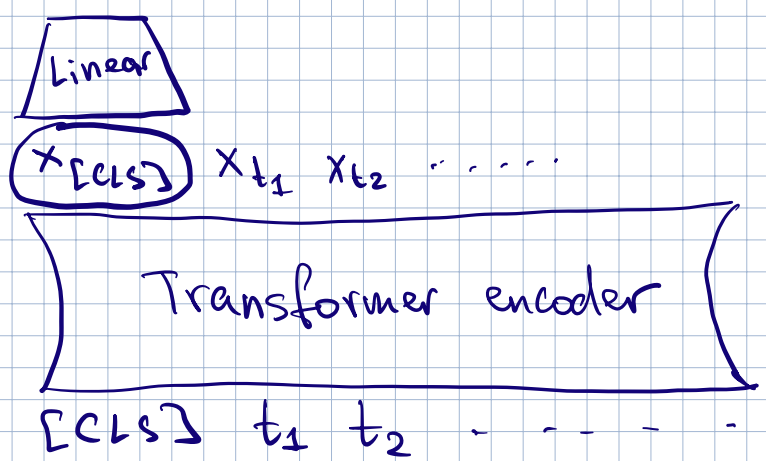
NSP – Next Sentence Prediction

Model input



$E_{[CLS]}$	E_{t_1}	E_{t_2}	...	E_{t_N}	$E_{[SEP]}$	E_{s_1}	....	E_{s_M}
E_0	E_1	E_2	...	E_N	E_{N+1}	E_{N+2}	....	E_{N+M+1}
E_A	E_A	E_A	...	E_A	E_A	E_B	....	E_B

Linear probing



GLUE (General Language Understanding Evaluation)

- 9 NLP tasks, classification
  - CoLA – Corpus of Language Acceptability
  - MNLI – Multi-genre Natural Language Inference

- NLP: ROBERTA, ALBERT
- Images: BEiT
- Audio: HUBERT

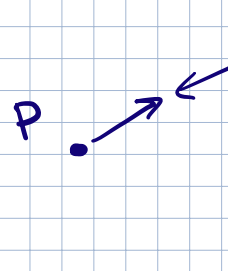
SSL for images

Pre-text tasks:

- Rotation angle prediction
- Jigsaw puzzle
- Image colorization

Triplet loss

$\|f(x) - f(y)\|_2^2$ – semantic distance



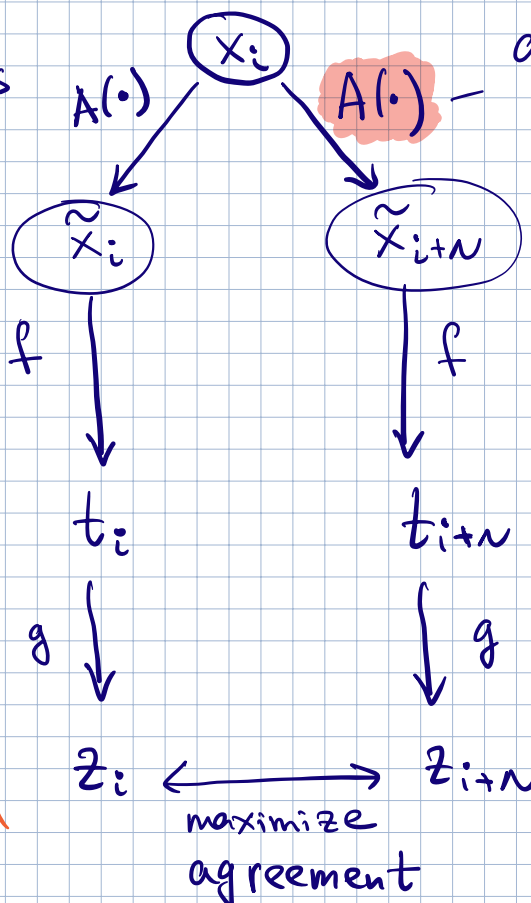
$$\max\left(\|f(A)-f(P)\|_2^2 - \|f(A)-f(N)\|_2^2 + m, 0\right) \to \min_f$$

$m > 0$

SimCLR (A Simple framework for Contrastive Learning of Representaions)

$\{x_i\}_{i=1}^{N}$ - batch of images

$\{\tilde{x}_i\}_{i=1}^{2N}$ - double batch of augmented images



Contrastive loss

$$\ell(z_i, z_{i+N}) = -\log \frac{e^{\text{sim}(z_i, z_{i+N})}}{\sum_{k=1}^{2N} \mathbb{1}_{\{k \neq i\}} e^{\text{sim}(z_i, z_k)}}$$

$$\text{sim}(x, y) = \frac{\langle x, y \rangle}{\|x\|_2 \cdot \|y\|_2}$$

$$L(z_i, z_{i+N}) = \frac{1}{2}\left(\ell(z_i, z_{i+N}) + \ell(z_{i+N}, z_i)\right)$$

- Random Crop
- Color Jitter
- Convert to Grayscale
- Gaussian Blur