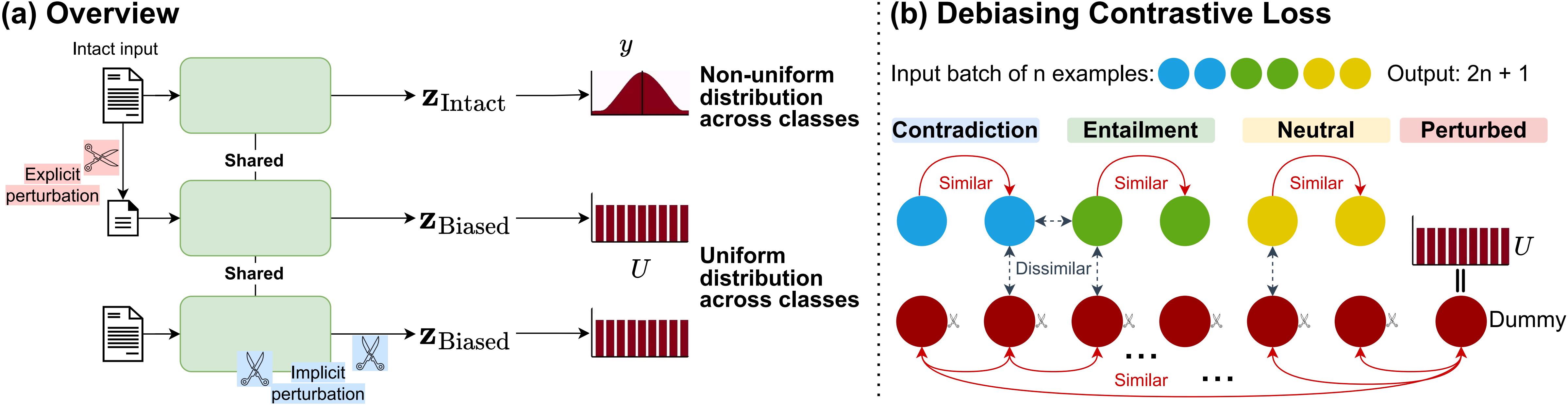
(a) Overview

(b) Debiasing Contrastive Loss

Intact input
Explicit perturbation
Shared
Shared
Implicit perturbation
$\mathbf{z}_{\text{Intact}}$
$\mathbf{z}_{\text{Biased}}$
$\mathbf{z}_{\text{Biased}}$
y
U
Non-uniform distribution across classes
Uniform distribution across classes
Input batch of n examples:
Output: 2n + 1
Contradiction
Entailment
Neutral
Perturbed
Similar
Similar
Similar
Dissimilar
U
Dummy
...
Similar
...