Fast AI Lesson 8: NLP

QUEENSLAND HUB

QUEENSLAND AI

Language models
And
Much much more.

About me: Jaco du Plessis

Mechatronic Engineer

Head of ML at Notiv

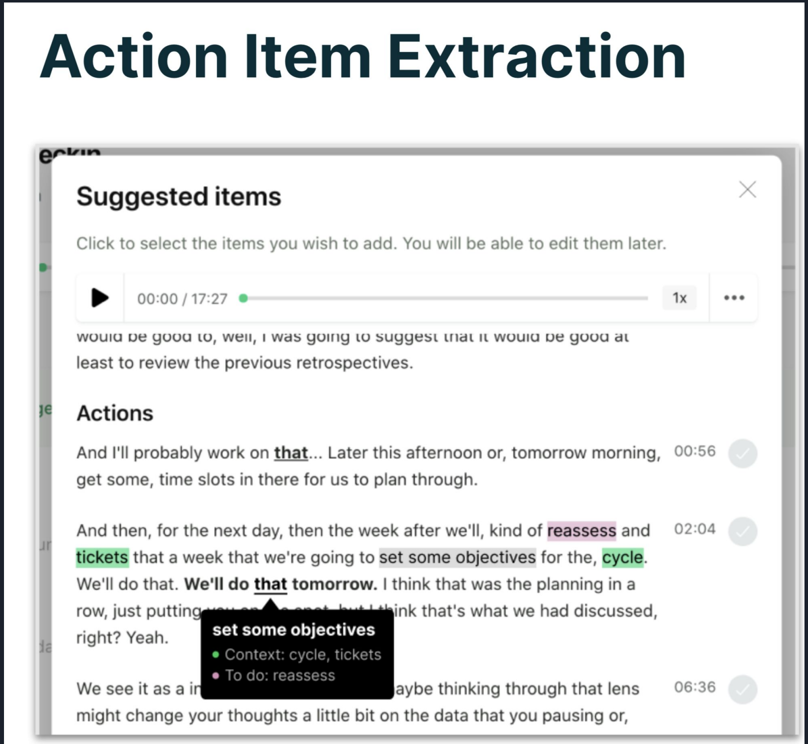
Focused building nlp and audio deep learning products

Hiring for NLP

Twitter

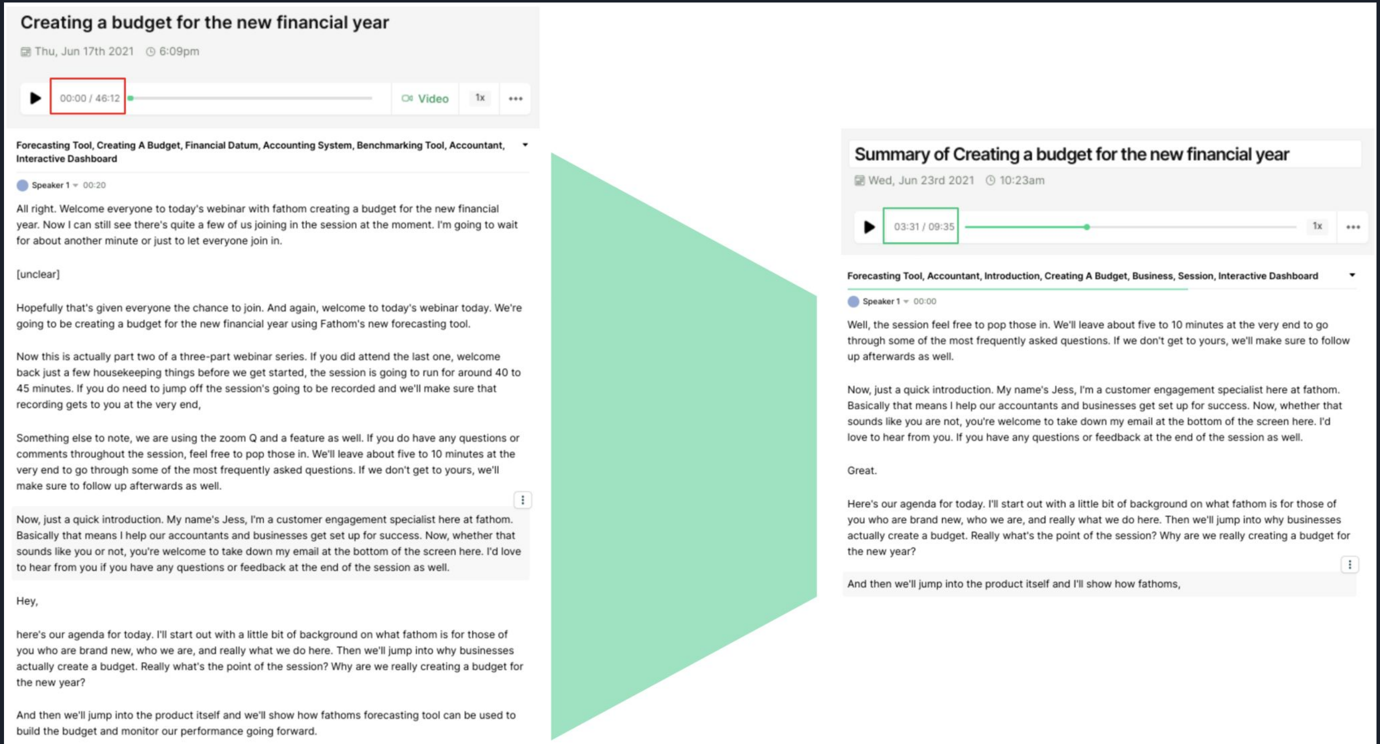
linkedin.com/in/jaco-du-plessis-566b61105

Notiv



Action Item Extraction
Suggested items
Click to select the items you wish to add. You will be able to edit them later.
00:00 / 17:27
1x
least to review the previous retrospectives.
Actions
And I'll probably work on that... Later this afternoon or, tomorrow morning, get some, time slots in there for us to plan through.
00:56
And then, for the next day, then the week after we'll, kind of reassess and tickets that a week that we're going to set some objectives for the, cycle. We'll do that. We'll do that tomorrow. I think that was the planning in a row, just putting
ink that's what we had discussed, right? Yeah.
02:04
set some objectives
Context: cycle, tickets
To do: reassess
We see it as a i
aybe thinking through that lens might change your thoughts a little bit on the data that you pausing or,
06:36

Notiv - summarization



Agenda

- Recap of content
- Transformers aka sesame street
- How things go wrong in nlp
- Make your models better
- Code examples
- Paper review

Recap

- Tokenizer
- numericalization
- Batching
- Embedding
- Model -> NN - with for loop
- Classification head

Tokenization

```
first(spacy(['The U.S. dollar $1 is $1.00.']))
```

```
(#9) ['The','U.S.','dollar','$','1','is','$','1.00','.']
```

```
coll_repr(tkn('&copy;   Fast.ai www.fast.ai/INDEX'), 31)
```

```
"(#11) ['xxbos','©','xxmaj','fast.ai','xxrep','3','w','.fast.ai','/','xxup','index'...]"
```

Numericalization

```
nums = num(toks)[:20]; nums
```

```
tensor([  2,   8,  21,  28,  11,  90,  18,  59,   0,  45,   9, 351, 499,  11,  72, 533, 584, 146,  29,  12])
```

This time, our tokens have been converted to a tensor of integers that our model can receive. We can check that they map back to the original text:

```
' '.join(num.vocab[o] for o in nums)
```

```
'xxbos xxmaj this movie , which i just xxunk at the video store , has apparently sit around for a'
```

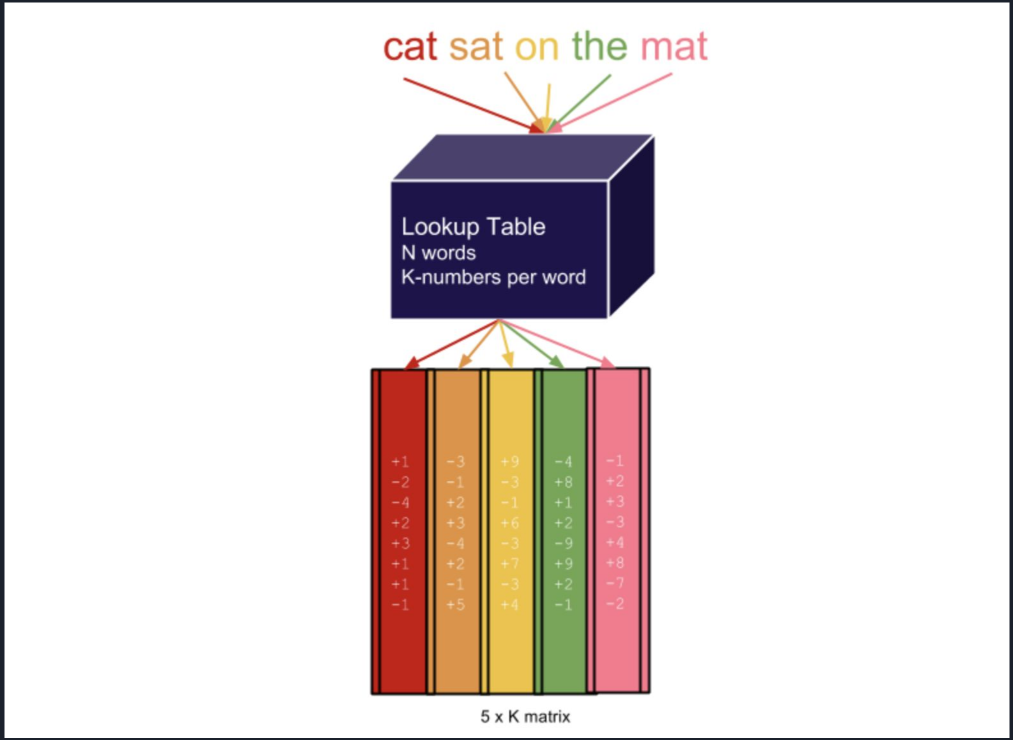
Batching - Language model

xxbos	xxmaj	in	this	chapter	,	we	will	go	back	over	the	example	of	classifying
movie	reviews	we	studied	in	chapter	1	and	dig	deeper	under	the	surface	.	xxmaj
first	we	will	look	at	the	processing	steps	necessary	to	convert	text	into	numbers	and
how	to	customize	it	.	xxmaj	by	doing	this	,	we	'll	have	another	example
of	the	preprocessor	used	in	the	data	block	xxup	api	.	\n	xxmaj	then	we
will	study	how	we	build	a	language	model	and	train	it	for	a	while	.

xxbos	xxmaj	in	this	chapter
movie	reviews	we	studied	in
first	we	will	look	at
how	to	customize	it	.
of	the	preprocessor	used	in
will	study	how	we	build

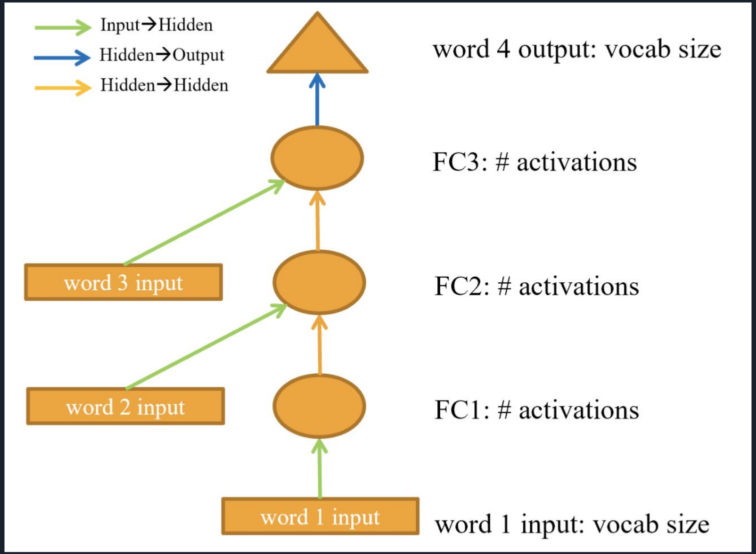
,	we	will	go	back
chapter	1	and	dig	deeper
the	processing	steps	necessary	to
xxmaj	by	doing	this	,
the	data	block	xxup	api
a	language	model	and	train

Embedding

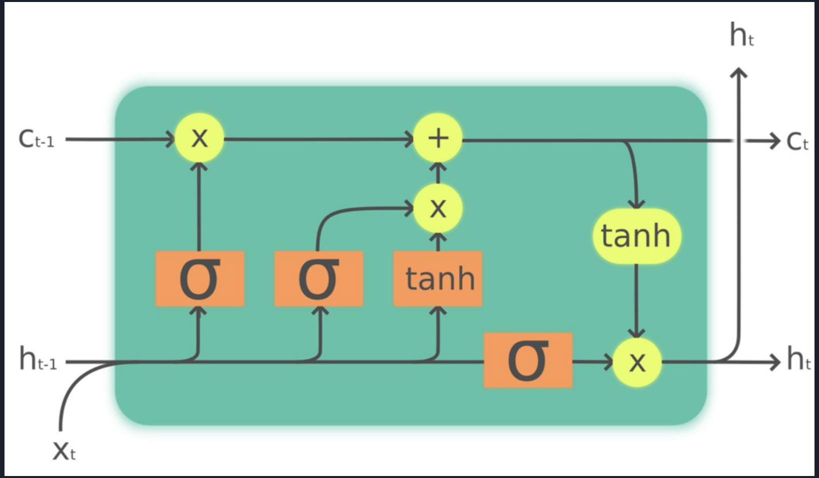


https://developer.nvidia.com/blog/understanding-natural-language-deep-neural-networks-using-torch/

Model



Input→Hidden
Hidden→Output
Hidden→Hidden
word 4 output: vocab size
FC3: # activations
FC2: # activations
FC1: # activations
word 3 input
word 2 input
word 1 input
word 1 input: vocab size



h_t
C_{t-1}
C_t
x
+
x
tanh
σ
σ
tanh
σ
x
h_{t-1}
h_t
X_t

Transformers

Attention is all you need.

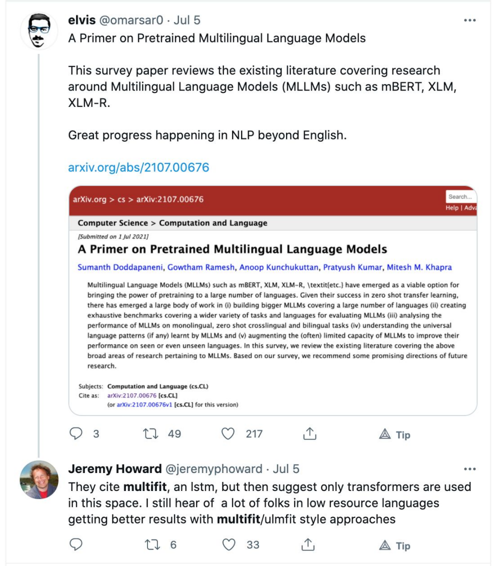
BERT, ALBERT, GPT series

Large scale expensive
to train

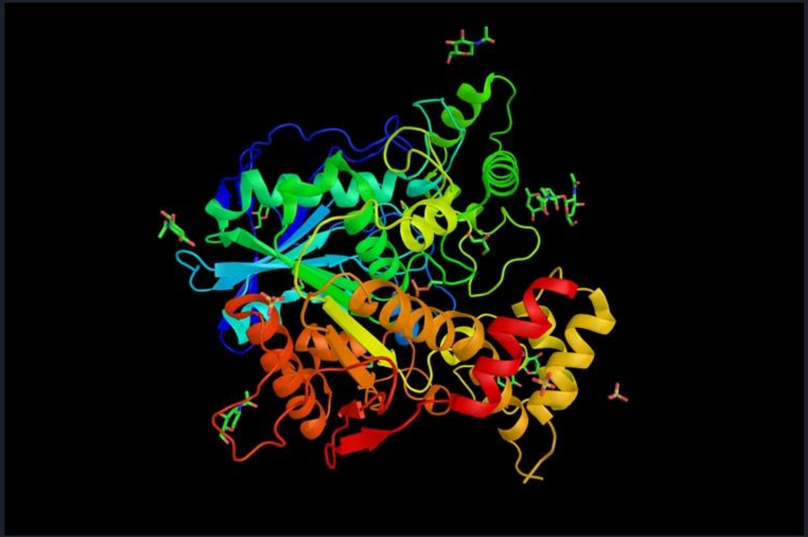
Not always needed

Transformers

[Single Headed Attention RNN: Stop Thinking With Your Head]



Transformers - do learn about them



https://jalammar.github.io/illustrated-transformer
http://nlp.seas.harvard.edu/2018/04/03/attention.html
The Transformer for language translation (NLP video 18)

How things go wrong - Noise

Repetition	When is Eas ugh Easter this year?
Correction	When is Lent I meant Easter this year?
Restarts	How much no wait when is Easter this year?

Model	Train	Eval	HasAns-F1	NoAns-F1	Overall-F1
BERT-QA	ALL	SQuAD	83.87	70.55	77.46
		Heuristics	51.45 ↓ 32.42	74.49 ↑ 3.94	62.53 ↓ 14.93
		DISFL-QA	40.97 ↓ 42.90	75.97 ↑ 5.42	57.81 ↓ 19.65
	ANS	SQuAD	89.63	-	89.63
		Heuristics	80.52 ↓ 9.11	-	80.52 ↓ 9.11
		DISFL-QA	78.88 ↓ 10.75	-	78.88 ↓ 10.75

https://arxiv.org/pdf/2106.04016.pdf

How things go wrong - Sentiment Classifier

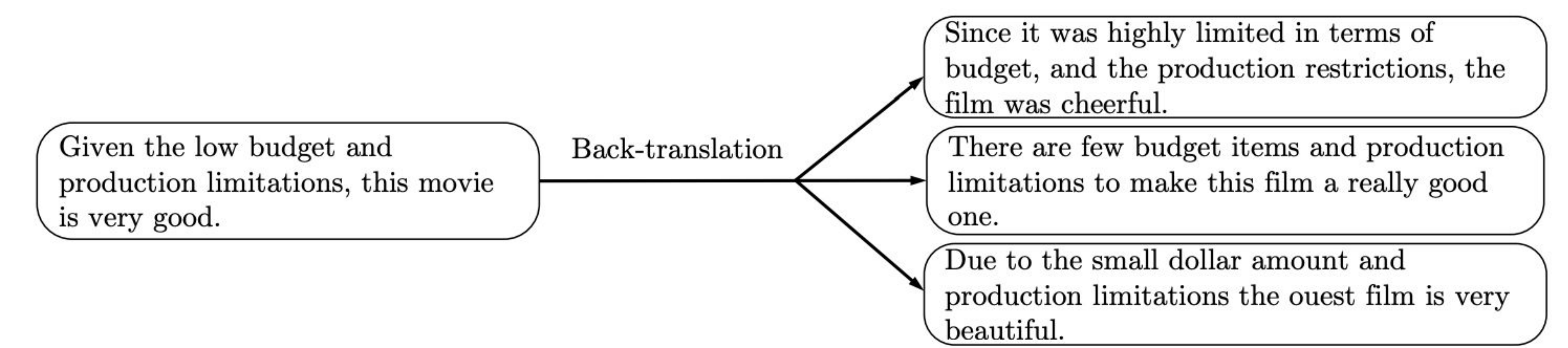
1904.12848 : Unsupervised Data Augmentation for Consistency Training

Datasets (# Sup examples)	IMDb (25k)
Pre-BERT SOTA	*4.32*
$\text{BERT}_{\text{LARGE}}$	4.51

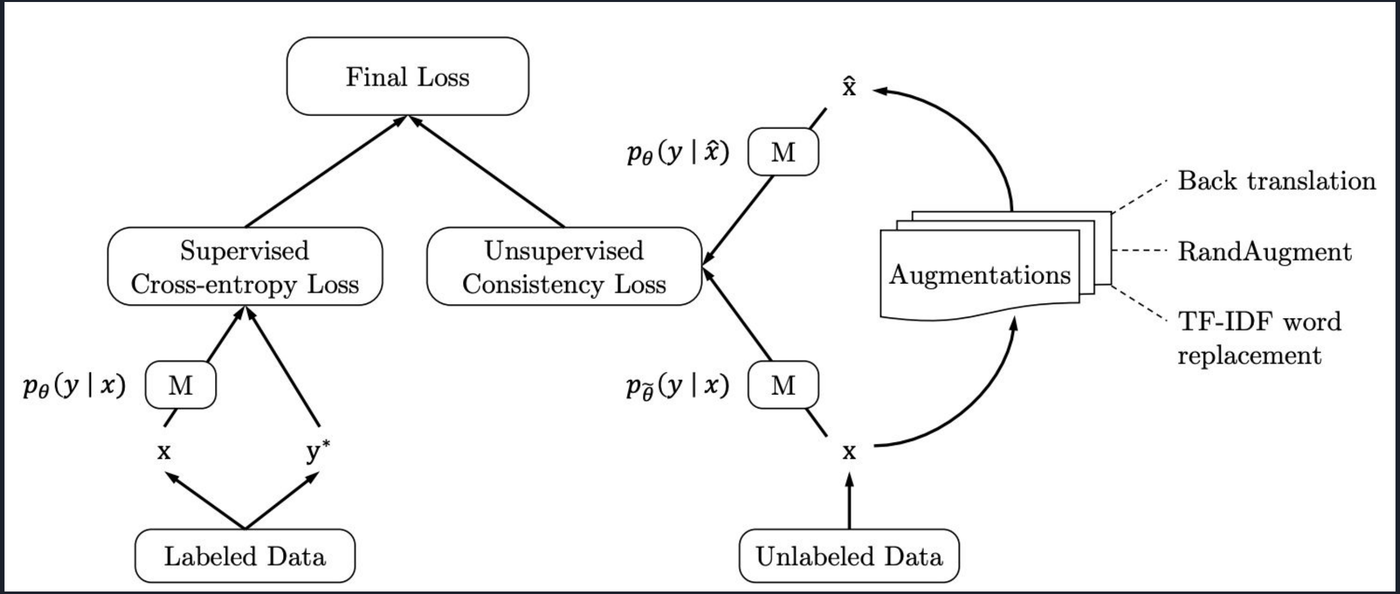
S

Initialization	**UDA**	IMDb (20)
Random	✗	43.27
	✓	25.23
$\text{BERT}_{\text{BASE}}$	✗	18.40
	✓	5.45
$\text{BERT}_{\text{LARGE}}$	✗	11.72
	✓	4.78
$\text{BERT}_{\text{FINETUNE}}$	✗	6.50
	✓	**4.20**

Unsupervised Data Augmentation for Consistency Training



Unsupervised Data Augmentation for Consistency Training



Unsupervised Data Augmentation for Consistency Training

Confidence-based masking. We find it to be helpful to mask out examples that the current model is not confident about. Specifically, in each minibatch, the consistency loss term is computed only on examples whose highest probability among classification categories is greater than a threshold β. We set the threshold β to a high value. Specifically, β is set to 0.8 for CIFAR-10 and SVHN and 0.5 for ImageNet.

Domain-relevance Data Filtering. Ideally, we would like to make use of out-of-domain unlabeled data since it is usually much easier to collect, but the class distributions of out-of-domain data are mismatched with those of in-domain data, which can result in performance loss if directly used [44]. To obtain data relevant to the domain for the task at hand, we adopt a common technique for detecting out-of-domain data. We use our baseline model trained on the in-domain data to infer the labels of data in a large out-of-domain dataset and pick out examples that the model is most confident about. Specifically, for each category, we sort all examples based on the classified probabilities of being in that category and select the examples with the highest probabilities.

Unsupervised Data Augmentation for Consistency Training

$$\min_{\theta} \mathcal{J}(\theta) = \mathbb{E}_{x_1 \sim p_L(x)} \left[-\log p_\theta(f^*(x_1) \mid x_1) \right] + \lambda \mathbb{E}_{x_2 \sim p_U(x)} \mathbb{E}_{\hat{x} \sim q(\hat{x} \mid x_2)} \left[\mathrm{CE}\left(p_{\tilde{\theta}}(y \mid x_2) \| p_\theta(y \mid \hat{x}) \right) \right] \quad (1)$$

$$\frac{1}{|B|} \sum_{x \in B} I(\max_{y'} p_{\tilde{\theta}}(y' \mid x) > \beta) \mathrm{CE}\left(p_{\tilde{\theta}}^{(sharp)}(y \mid x) \| p_\theta(y \mid \hat{x}) \right)$$

$$p_{\tilde{\theta}}^{(sharp)}(y \mid x) = \frac{\exp(z_y / \tau)}{\sum_{y'} \exp(z_{y'} / \tau)}$$