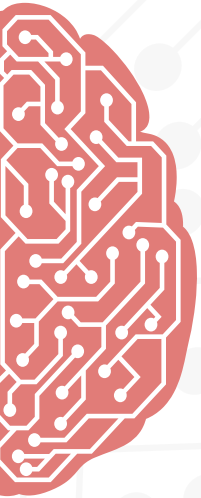


HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

On the differences between BERT and MT embedding spaces

Raúl Vázquez

FOTRAN
Found in Translation

March 2021

Outline

01. Motivation

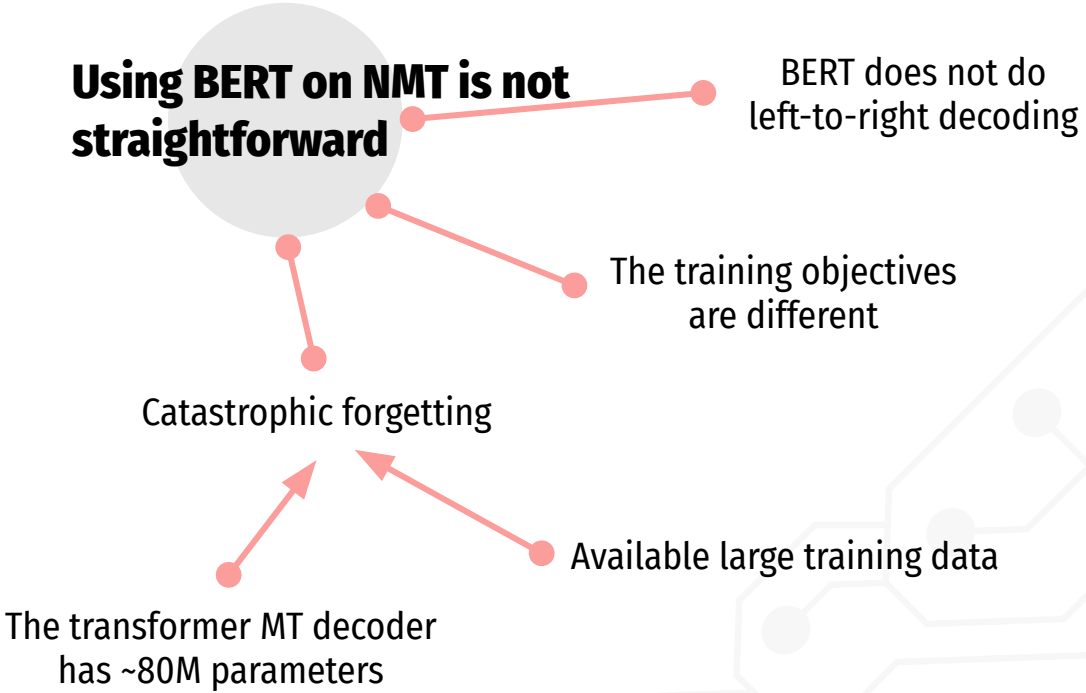
02. What's been done?

03. Analysis of the embedding spaces

04. Alignment Methods

05. Conclusions

Motivation



?

Are the two tasks really incompatible?

Motivation

Claims of efficient ways to incorporate BERT into MT abound.

What has been done?

Claims of efficient ways to
incorporate BERT into MT abound.

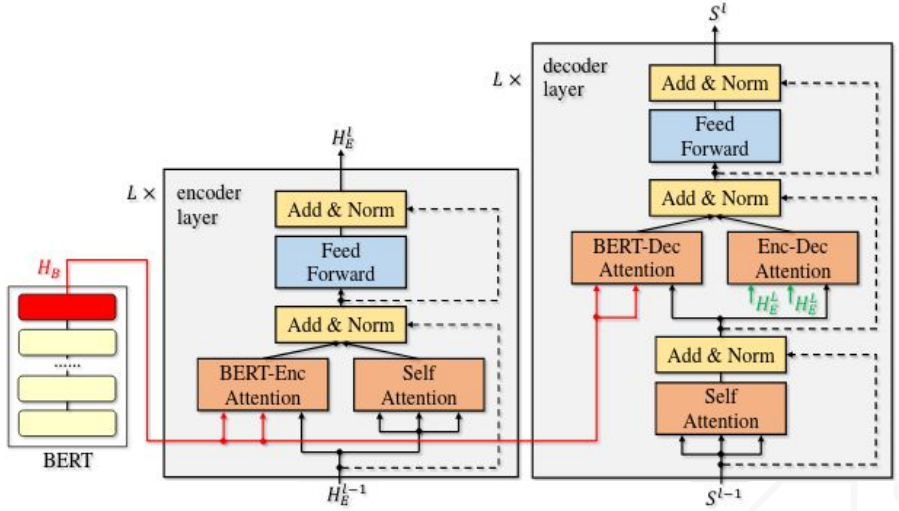
01. Incorporating BERT into NMT (Zhu et al., 2020)

02. On the use of BERT for NMT (Clinchant et al., 2019)

03. Towards Making the Most of BERT in NMT (Yang et al., 2020)

04. Recycling a Pre-trained BERT Encoder for NMT (Imamura and Sumita, 2019)

Incorporating BERT into NMT (Zhu et al., 2020)



1. train an NMT model until convergence.
2. initialize the enc and dec of the BERT-fused model with this model.

The BERT-encoder and BERT-decoder attention are randomly initialized

Figure 1: The architecture of BERT-fused model. Dash lines denote residual connections. H_B(red part) andH^L_E(green part) denote the output of the last layer from BERT and encoder.

Incorporating BERT into NMT (Zhu et al., 2020)

Algorithm	En→De	En→Fr
DynamicConv (Wu et al., 2019)	29.7	43.2
Evolved Transformer (So et al., 2019)	29.8	41.3
Transformer + Large Batch (Ott et al., 2018)	29.3	43.0
Our Reproduced Transformer	29.12	42.96
Our BERT-fused model	30.75	43.78

Table 1: BLEU scores of newest WMT14 translation

"Experiments on WMT tasks are conducted on 8 M40 GPUs ... It takes 1, 8 and 14 days to obtain the pre-trained NMT models, and additional 1, 7 and 10 days to finish the whole training process."

On the use of BERT for NMT (Clinchant et al., 2019)

A systematic comparison of 4 different BERT+NMT architectures:

- **Baseline:** A transformer-big model
- **Emb:** The embedding layer is replaced by the BERT
- **FT:** The encoder initialized by the BERT parameters
- **Freeze:** FT with BERT frozen

Using pre-trained LMs for NMT should be assessed beyond BLEU scores on in-domain datasets.

On the use of BERT for NMT (Clinchant et al., 2019)

	news14	news18	iwslt15	wiki	kde	OpenSub
Baseline	27.3	39.5	28.9	17.6	18.1	15.3
NMTsrc.FT	27.7	40.1	28.7	18.3	18.4	15.3
Wiki.FT	27.7	**40.6**	28.7	18.4	**19.0**	15.4
News.FT	**27.9**	40.2	29.1	18.8	17.9	15.7
News.Emb	27.7	39.9	**29.3**	**18.9**	18.2	**16.0**
News.Freeze	23.6	35.5	26.5	15.0	15.1	13.8

Table 2: BLEU scores for in-domain and out-of-domain testing.

- Gains: ~1 BLEU in-domain and ~2 BLEU out-of-domain
- Cost: Train your own 6-layered BERT

Towards Making the Most of BERT in NMT (Yang et al., 2020)

CT_{NMT} := a combination of 3 techniques:

- asymptotic distillation (AD)
- dynamic switch for knowledge fusion (DS)
- rate-scheduled updating

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{nmt}} + (1-\alpha) \cdot \mathcal{L}_{\text{kd}}$$
$$\mathcal{L}_{\text{kd}} = -||\hat{h}^{lm} - h_l||_2^2$$

$$h = g \odot h^{lm} + (1-g) \odot h^{nmt}$$

$$\theta_t^{lm} = \theta_{t-1}^{lm} - \eta^{lm} \nabla_{\theta^{lm}} \mathcal{L}(\theta^{lm})$$
$$\theta_t^{nmt} = \theta_{t-1}^{nmt} - \eta^{nmt} \nabla_{\theta^{nmt}} \mathcal{L}(\theta^{nmt})$$

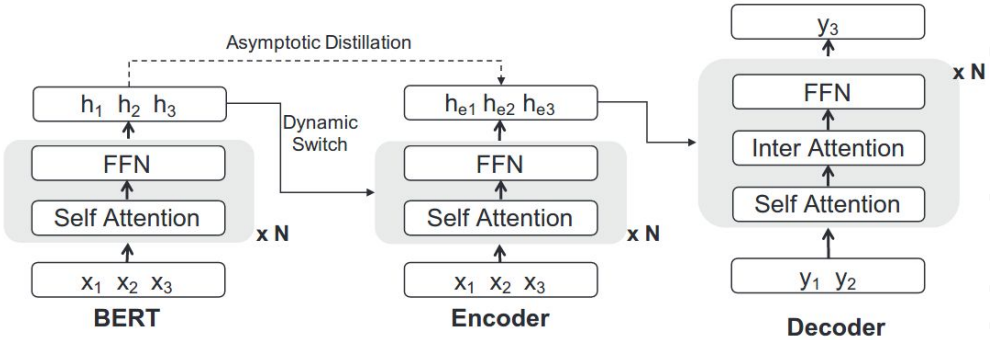


Figure 2: The overall CT_{NMT} with asymptotic distillation and dynamic switch.

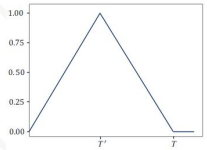


Figure 3: The LR used for η^{lm}. Then $\eta^{nmt} = \rho * \eta^{lm}$

Towards Making the Most of BERT in NMT (Yang et al., 2020)

System	Architecture	En-De	En-Fr	En-Zh
Existing systems				
Vaswani et al. (2017b)	Transformer base	27.3	38.1	-
Vaswani et al. (2017b)	Transformer big	28.4	41.0	-
Lample and Conneau (2019)	Transformer big + Fine-tuning	27.7	-	-
Lample and Conneau (2019)	Transformer big + Frozen Feature	28.7	-	-
Chen et al. (2018)	RNMT+ + MultiCol	28.7	41.7 -	
Our NMT systems				
CTNMT	Transformer (base)	27.2	41.0	37.3
CTNMT	Rate-scheduling	29.7	41.6	38.4
CTNMT	Dynamic Switch	29.4	41.4	38.6
CTNMT	Asymptotic Distillation	29.2	41.6	38.3
CTNMT	+ ALL	**30.1**	**42.3**	**38.9**

Table 3: Case-sensitive BLEU scores.

* *"... If not specified, we choose the second-to-last hidden states of BERT to help the training of NMT."*

Recycling a Pre-trained BERT Encoder for NMT (Imamura and Sumita, 2019)

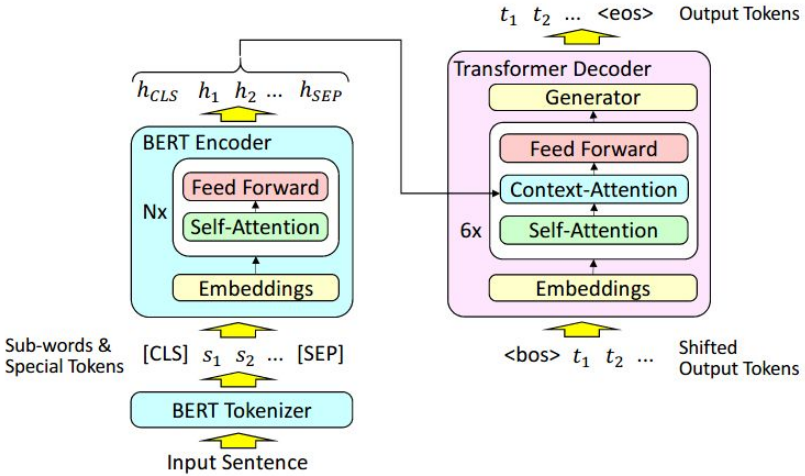


Figure 3: Structure of NMT using a BERT encoder.

Two-stage training:

1. train only unlearned parameters
2. apply fine-tuning

Recycling a Pre-trained BERT Encoder for NMT (Imamura and Sumita, 2019)

	System	LR	Dev. PPL ↓	BLEU ↑ 2013	2014	2015	Remark
Baselines	Transformer base	4.0×10^{-4}	4.23	26.29	27.22	29.48	Stat. test baseline
	Transformer BERT size	4.0×10^{-4}	4.04	26.15	27.09	29.32	
NMT with BERT	Direct fine-tuning	8×10^{-5}	4.28	0.13 (-)	0.10 (-)	0.12 (-)	# Epochs = 33
		4.0×10^{-4}	4.09	0.48 (-)	0.42 (-)	0.54 (-)	# Epochs = 29
	Proposed: Decoder training only	4.0×10^{-4}	4.76	24.13 (-)	23.62 (-)	25.74 (-)	# Epochs = 65
	+ Fine-tuning	4×10^{-5}	3.93	**27.14** (+)	28.27 (+)	30.68 (+)	# Epochs = 21
		8×10^{-5}	**3.92**	27.05 (+)	**28.90** (+)	**30.89** (+)	# Epochs = 9
		1.2×10^{-4}	3.93	27.03 (+)	28.50 (+)	30.51 (+)	# Epochs = 11

Table : Results of the WMT english-german data (~4.46M sent pairs)

Table : Results of the IWSLT-2015 english-vietnamese data (~133k sent pairs)

	System	LR	Dev. PPL ↓	BLEU ↑ 2012	2013
Baseline	Transformer base	4.0×10^{-4}	11.54	24.03	26.12
NMT with BERT	Proposed: Decoder training only	4.0×10^{-4}	11.45	21.77 (-)	23.23 (-)
	+ Fine-tuning	8×10^{-5}	**8.98**	**26.77** (+)	**29.57** (+)

Are the two tasks really incompatible?

Analyse the embedding spaces and contrast them against each other

Analyzing the embedding spaces

Might provide some of answers

Measures of contextuality

(Ethayarajh, 2019)

SelfSim: the average cossim between the contextualized representations of a word across its occurrences

$$SelfSim_\ell(w) = \frac{1}{n^2 - n} \sum_j \sum_{k \neq j} \cos(f_\ell(s_j, i_j), f_\ell(s_k, i_k))$$

$SelfSim_\ell(w) = 1$ means that layer ℓ doesn't contextualize the representations.

IntraSim: the average cossim between the word representations in a sentence and their mean vector.

$$IntraSim_\ell(s) = \frac{1}{n} \sum_i \cos(\vec{s}_\ell, f_\ell(s, i))$$

$$\text{where } \vec{s}_\ell = \frac{1}{n} \sum_i f_\ell(s, i)$$

If $IntraSim_\ell(s)$ and $SelfSim_\ell(w)$ are low $\forall w \in s$, the model contextualizes words in layer ℓ by giving each one a context-specific representation distinct from all other words in the sentence.

Measures of contextuality

(Ethayarajh, 2019)

Isotropy = directional uniformity

plays an important role when using cosine similarity

AvgSim: average cossim of randomly sampled words

$$Baseline(f_\ell) = \mathbb{E}_{x,y \sim U(\mathcal{O})} [\cos(f_\ell(x), f_\ell(y))]$$
$$SelfSim^*_\ell(w) = SelfSim_\ell(w) - Baseline(f_\ell)$$

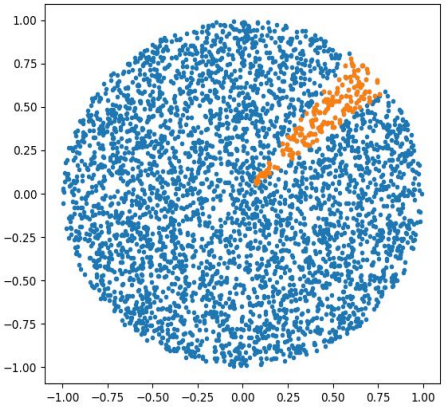


Figure 4: (blue) points sampled uniformly over the unitary sphere. (orange) points sampled over a portion of the unitary sphere such that cossim((x,y) , (1,1)) > 0.99

MT and BERT embeddings distributions

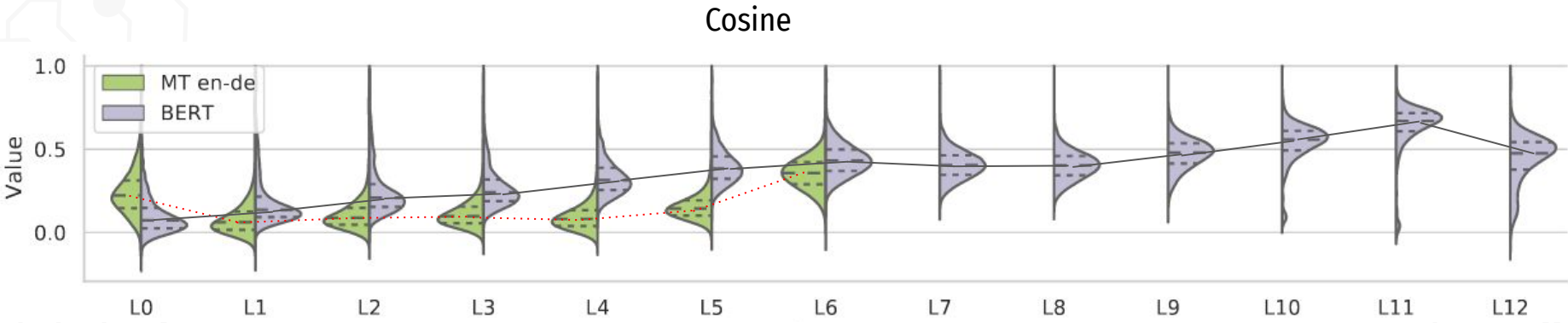


Figure 5: Cosine similarity distributions of randomly sampled words. (Avgsim is the mean of the distribution)

Self similarity and Intra-sentence similarity

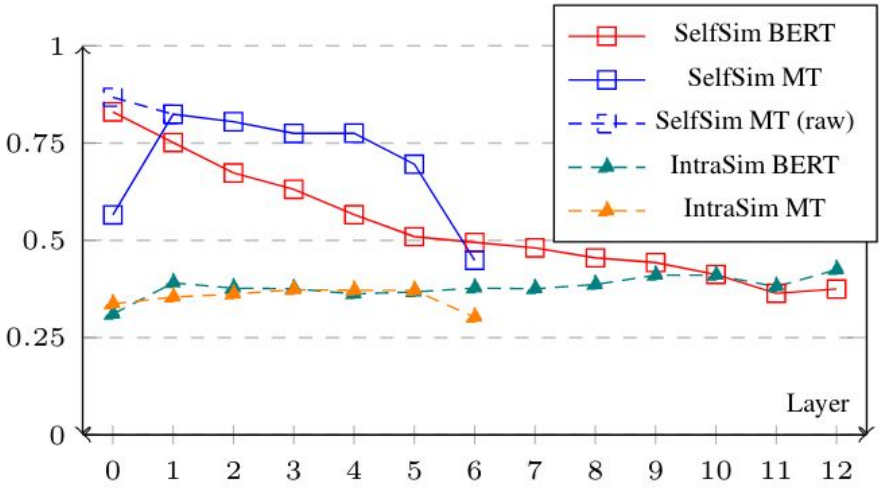


Figure 6: SelfSim and IntraSim for BERT and MT models.
SelfSim (raw) is before correction for anisotropy

Methods for aligning the representation spaces

Supervised transformation.
Maximize the contextual alignment
(Cao et al., 2020)

$$Loss = L + \lambda R$$

$$L(f_1, f_2; C) = -\sum_{\substack{(s,t)\in C \\ (i,j)\in a(s,t)}} \sum \text{sim}\left(f_1(i,s), f_2(j,t)\right)$$

$$R(f; C) = \sum_{s\in C} \sum_{i=1}^{len(t)} \|f_1(i,s) - f_1^{\circ}(i,s)\|_2^2$$

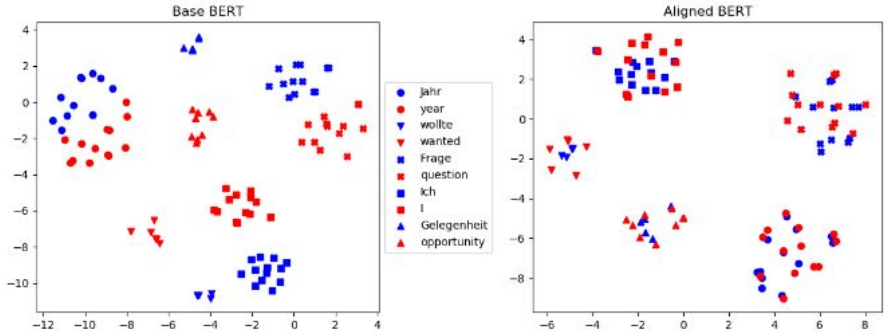


Figure 7: t-SNE view of the embedding space of multilingual BERT for english-german before(left) and after (right) alignment (Cao et al., 2020).

Methods for aligning the representation spaces

Unsupervised alignment via fine-tuning.
BERT-MT hybrid model.

We use smoothed cross entropy loss as training objective to fine-tune BERT representations for performing MT.

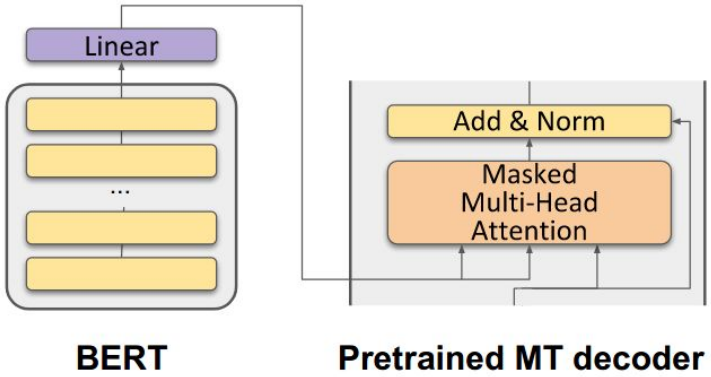


Figure 8: Detail of the full architecture where BERT comes into play as an encoder.

Alignment experiments: setup+BLEU

	Encoder	Supervised alignment	Fine-tuning
MTbaseline	Trf	✗	✗
huggingface en-de		✗	✗
M1:align	BERT	✓	✗
M2:finetune		✗	✓
M3:align+finetune		✓	✓

	Train		Val.
	Supervised	Unsupervised	
Europarl	45K	150K	1.5K
MuST-C	45K	150K	1.5K
newstest	13K	13K	500
Total	102K	313K	3.5K

	MuST-C	newstest 2014
MTbaseline	29.9	14.5
huggingface en-de	33.7	28.3
M1:align	21.4	18.1
M2:finetune	33.8	23.9
M3:align+finetune	34.1	25.0

Table: BLEU scores for the english-german test sets.

Aligned distributions

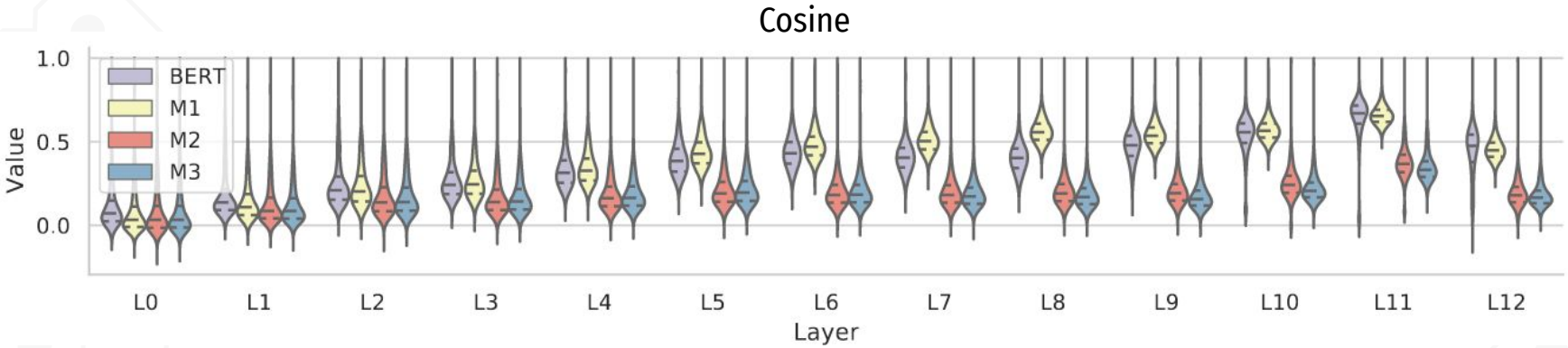


Figure 9: Cosine similarity distributions of randomly sampled words. Comparing out-of-box BERT with the aligned models M1/M2/M3.
(Avgsim is the mean of the distribution)

Conclusions

Analyzing the differences between pre-trained language models and machine translation encoders can help us develop methods for better using resources.

- Our work shows that the differences in both embedding spaces are one of the main reasons why current approaches to use BERT in MT perform “poorly”
- A simple and fast supervised alignment already enables MT
- Fine-tuning works better after aligning bot in our work and in Imamura and Sumita (2019)
-

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Thank you

Contact: raul.vazquez@helsinki.fi

FOTRAN
Found in Translation

References.

- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher

Cross-lingual Language Model Pretraining (Lample and Conneau, 2019)

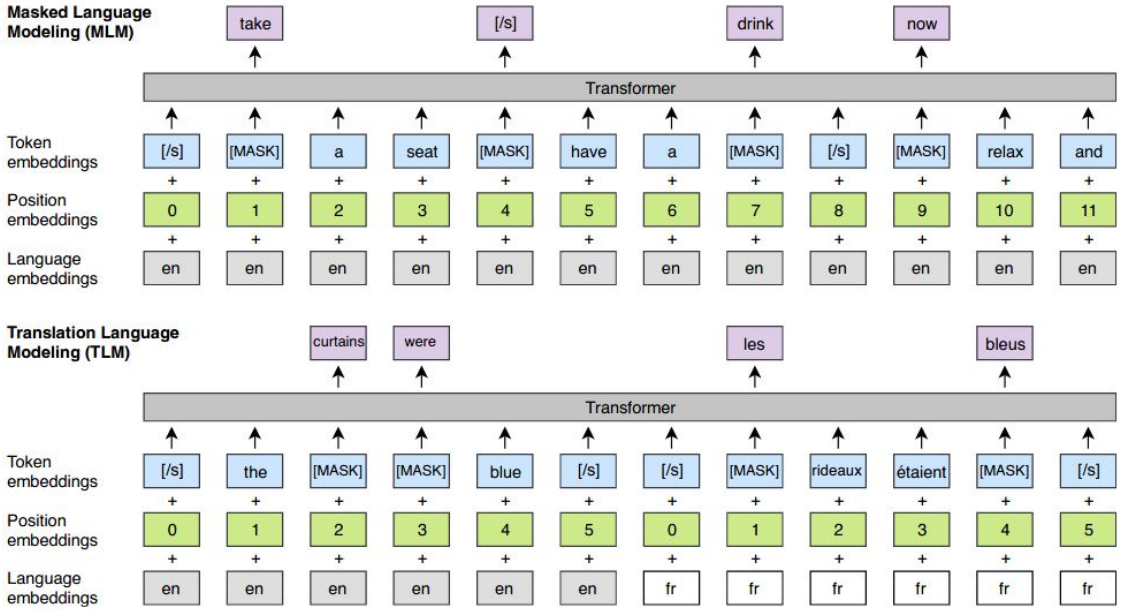
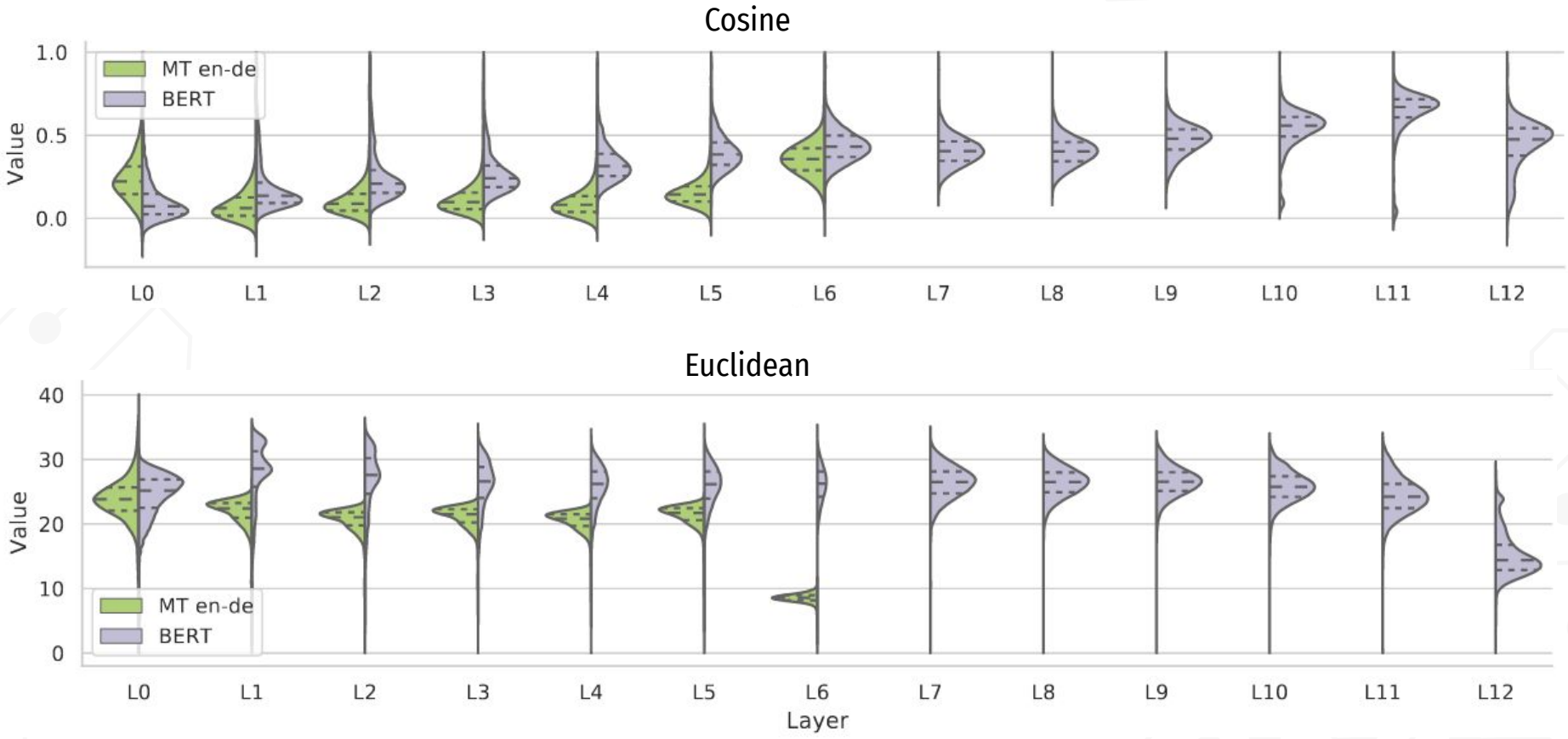
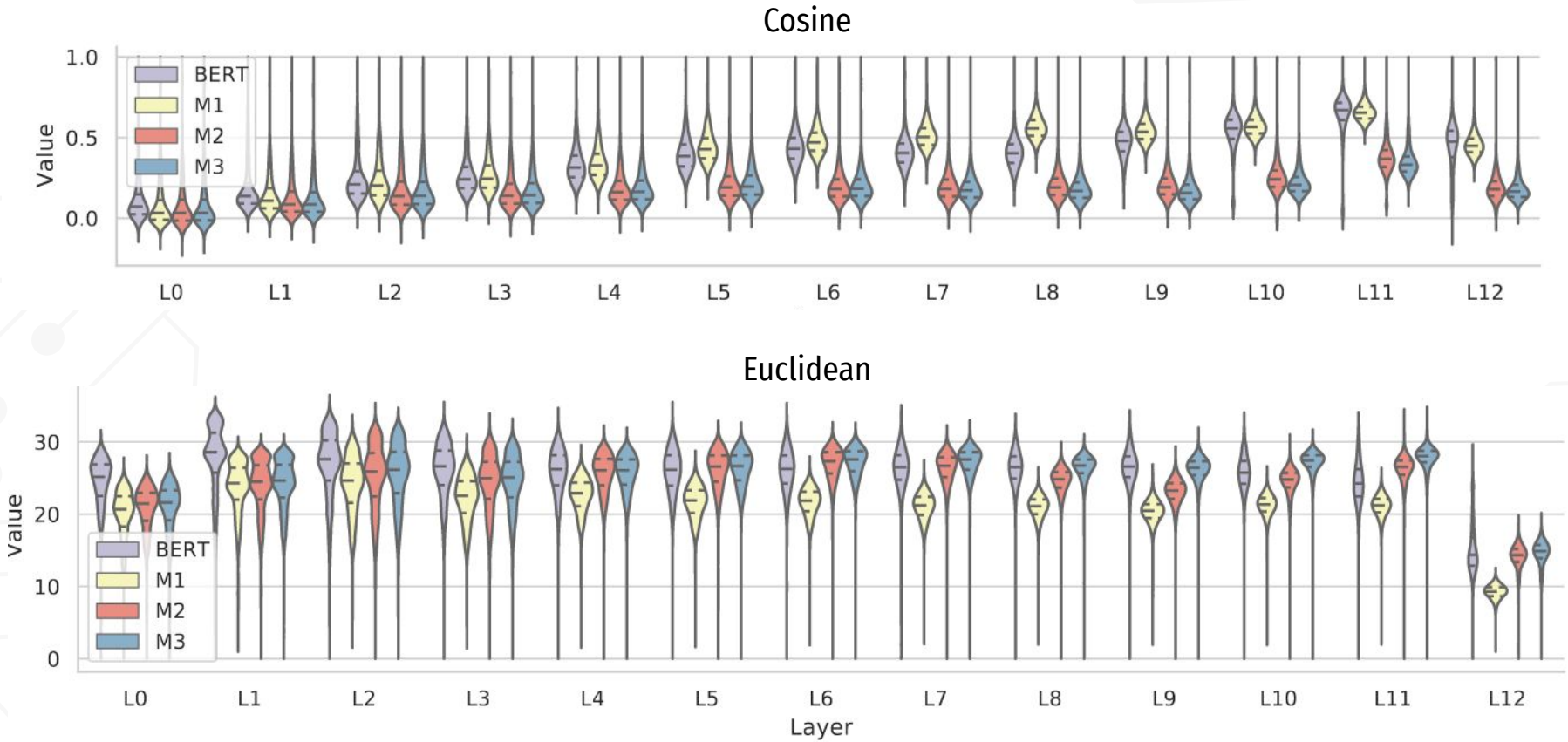


Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

Layer-by-layer comparison of MT and BERT embeddings distributions



Aligned distributions



Fonts & colors used

This presentation has been made using the following fonts:

Fira Sans Condensed
(https://fonts.google.com/specimen/Fira+Sans+Condensed)

#fb9d9a

#e17c78

#cf6965

#e7e7e7

#000000

Storyset

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out How it Works.

Pana

Amico

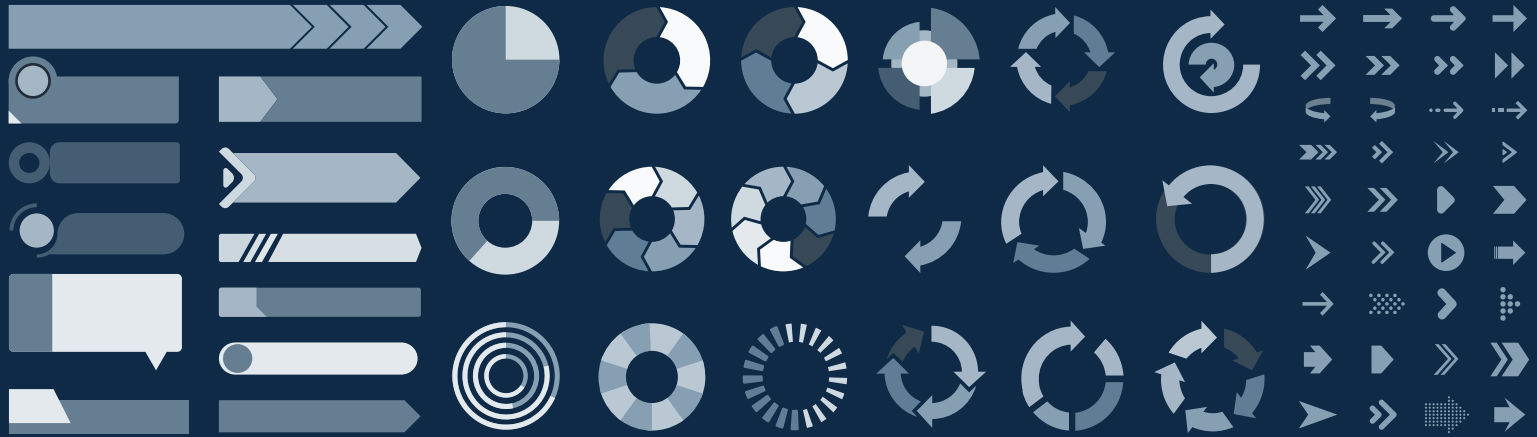
Bro

Rafiki

Cuate

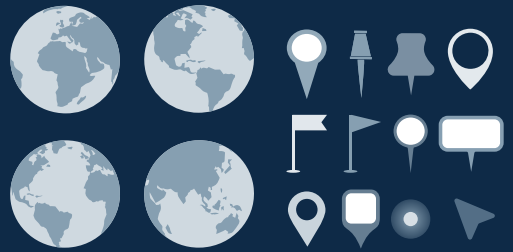
Use our editable graphic resources...

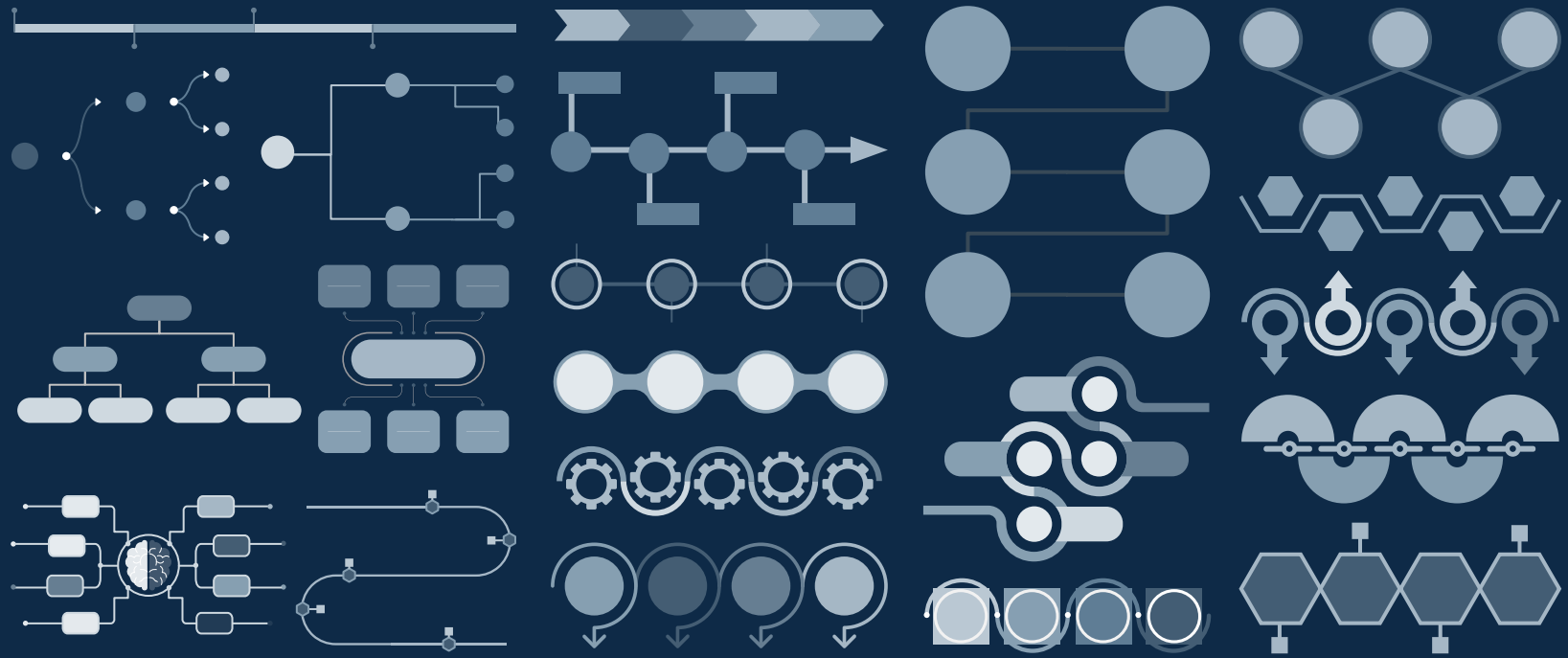
You can easily resize these resources without losing quality. To change the color, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Group the resource again when you're done.

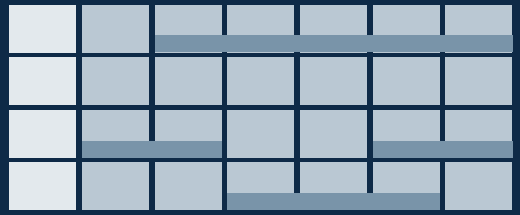


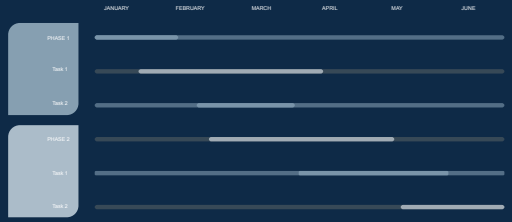


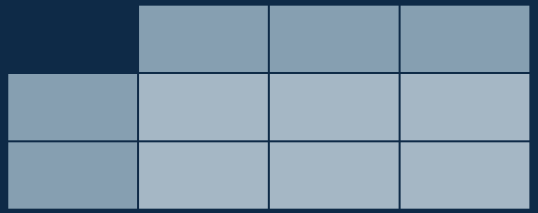


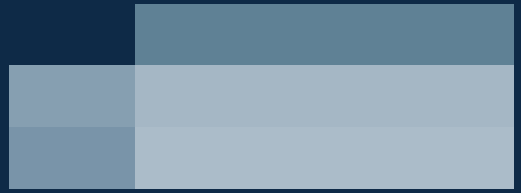



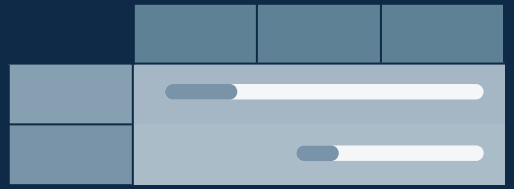


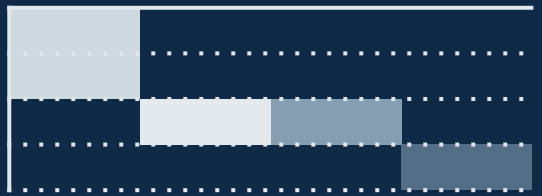


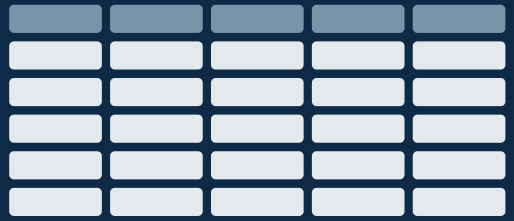

JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
PHASE 1
Task 1
Task 2
PHASE 2
Task 1
Task 2

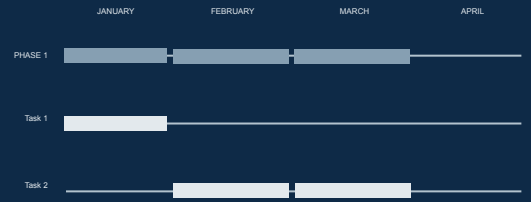


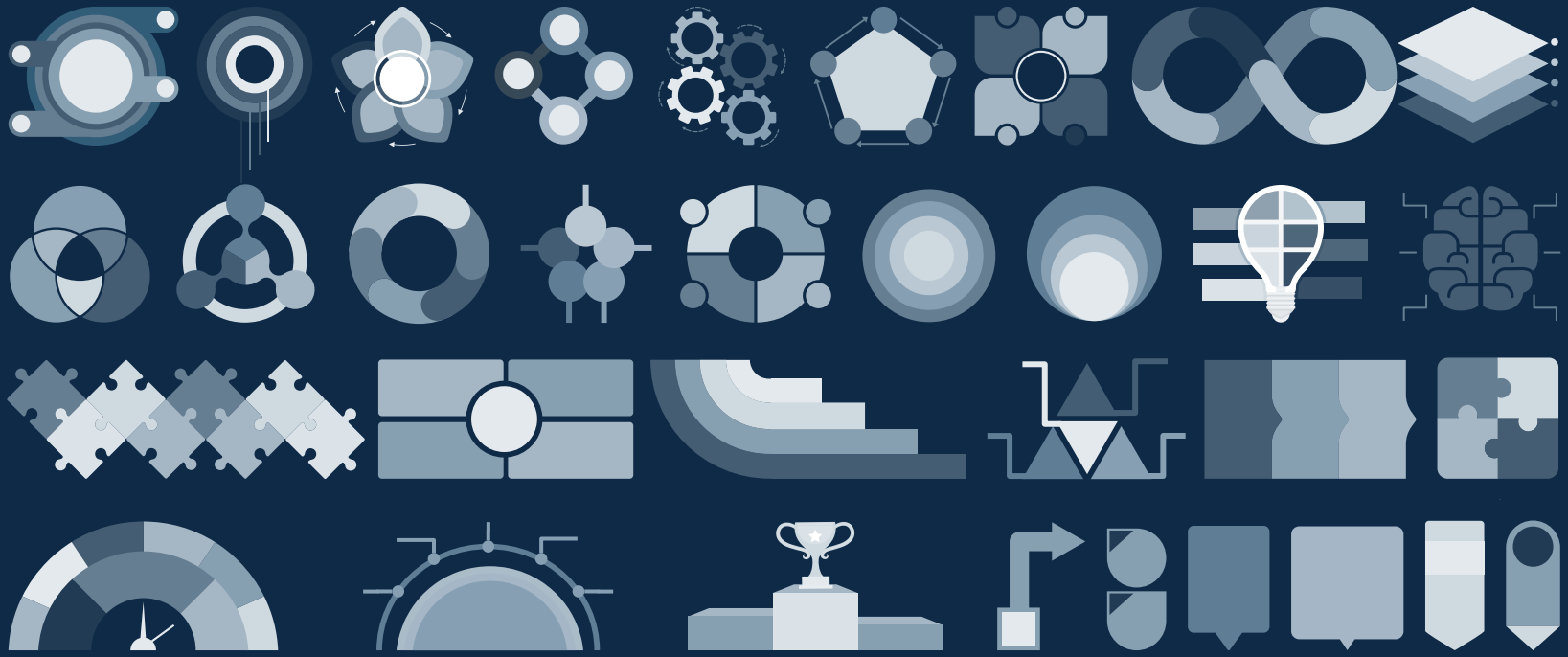


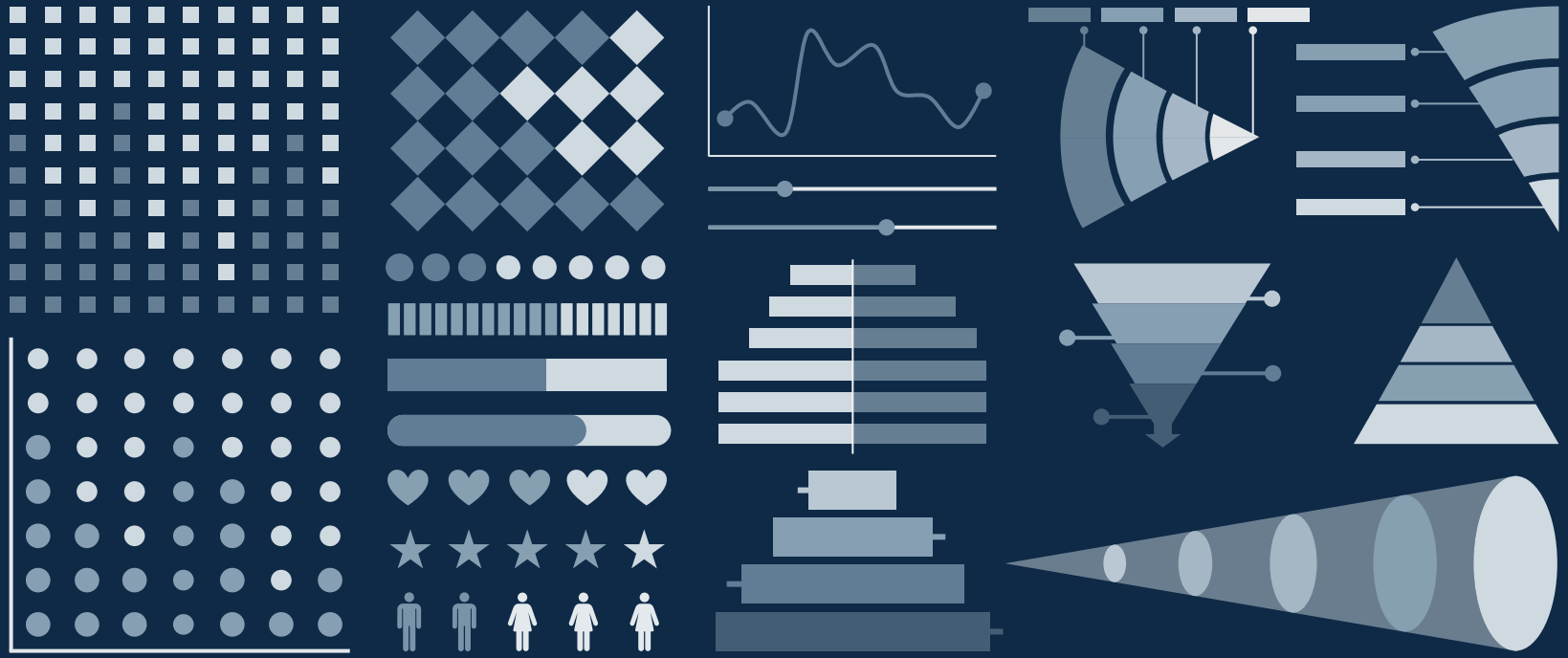







JANUARY
FEBRUARY
MARCH
APRIL
PHASE 1
Task 1
Task 2





...and our sets of editable icons

You can resize these icons without losing quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

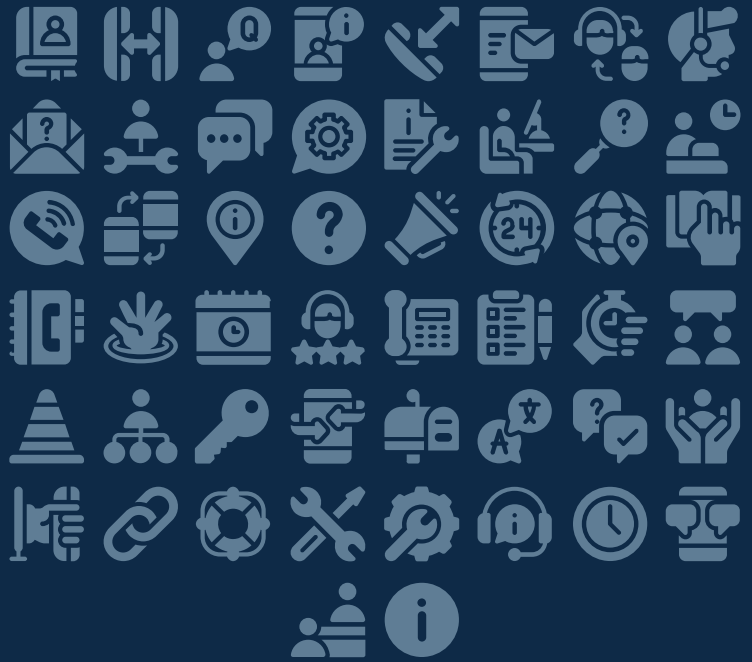
Educational Icons

Medical Icons

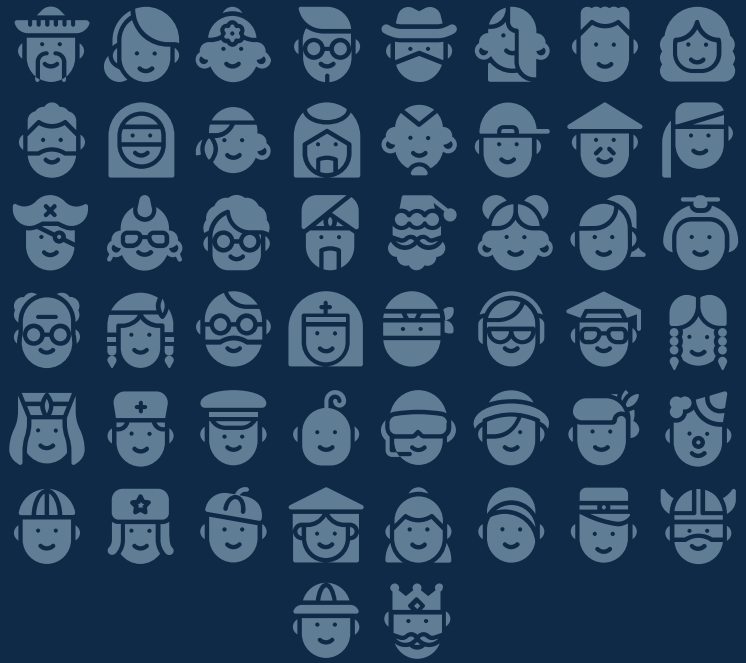
Business Icons

Teamwork Icons

Help & Support Icons



Avatar Icons



Creative Process Icons

Performing Arts Icons

Nature Icons

SEO & Marketing Icons