"Attention Is All You Need"
Résumé du document original

Introduction
L’apprentissage de séquences, notamment en traduction automatique et en modélisation de langage, reposait historiquement sur les réseaux récurrents (RNN) et les réseaux de neurones convolutionnels (CNN). Cependant, ces approches souffrent de limites, notamment le manque de parallélisation et des difficultés à capturer des dépendances distantes.
Les auteurs proposent une nouvelle architecture de réseau neuronal, le **Transformer**, qui repose uniquement sur des **mécanismes d’attention** sans utiliser de convolutions ni de récurrence. Cette approche améliore la qualité des traductions tout en réduisant considérablement le temps d'entraînement.

---

1. Contexte et motivation
Les modèles classiques de traduction automatique utilisent des architectures encodeur-décodeur basées sur des RNN ou des CNN. Cependant, ces modèles ont des contraintes :

- **Les RNN** traitent les séquences de manière séquentielle, ce qui empêche la parallélisation.
- **Les CNN** permettent une meilleure parallélisation mais peinent à capturer des relations entre mots éloignés.

L'attention permet de **modéliser les dépendances sans tenir compte de la distance** entre les éléments d’une séquence, facilitant ainsi un apprentissage plus efficace.

---

2. Architecture du Transformer
Le Transformer suit une **structure encodeur-décodeur**, mais chaque composant repose uniquement sur des mécanismes d’attention et des couches entièrement connectées.

2.1. Encodeur
L’encodeur transforme la séquence d’entrée en une représentation riche via six couches identiques, contenant chacune :

1. **Une couche d’auto-attention multi-tête**, qui capture les relations entre les mots.
2. **Un réseau de neurones entièrement connecté**, appliqué séparément à chaque mot.

Chaque sous-couche est suivie d’une normalisation de couche et de connexions résiduelles pour faciliter l’apprentissage.

2.2. Décodeur
Le décodeur est également composé de six couches et fonctionne de manière similaire à l’encodeur, avec une différence majeure :

- Une troisième sous-couche permet au décodeur d’intégrer les informations générées par l’encodeur.

De plus, un **masquage est appliqué** pour éviter que le modèle ne regarde les mots futurs lors de la génération.

---

3. Mécanisme d'attention

L'attention est au cœur du Transformer. Elle permet d'établir des connexions entre les différents éléments d'une séquence.

3.1. Attention "Scaled Dot-Product"

L'attention est calculée à partir de trois matrices :

- **Q (queries)** : représente les mots à comparer.
- **K (keys)** : représente les références pour la comparaison.
- **V (values)** : contient les informations à extraire.

L'attention est calculée par :

Attention(Q,K,V)=softmax(QKTdk)VAttention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V

où dkd_k est la dimension des clés, et la division par dk\sqrt{d_k} empêche que les valeurs softmax ne deviennent trop petites ou trop grandes.

3.2. Attention multi-tête

Au lieu d'utiliser une seule attention, le Transformer utilise **plusieurs têtes d'attention** pour apprendre différentes relations au sein d'une phrase.

Chaque tête effectue une attention indépendante, puis les résultats sont concaténés et projetés dans un espace vectoriel unique.

3.3. Applications de l'attention

Le Transformer utilise l'attention à trois niveaux :

1. **L'auto-attention dans l'encodeur**, permettant à chaque mot de considérer les autres mots de la phrase d'entrée.
2. **L'auto-attention dans le décodeur**, qui empêche le modèle d'accéder aux mots futurs grâce à un masquage.
3. **L'attention entre l'encodeur et le décodeur**, qui relie les mots traduits aux mots d'origine.

---

4. Autres composantes du Transformer

4.1. Réseaux de neurones entièrement connectés

Chaque couche du Transformer contient un réseau de neurones de la forme :

FFN(x)=ReLU(xW1+b1)W2+b2FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2

Ces réseaux permettent d'apporter une transformation non linéaire aux données.

4.2. Encodage des positions

Contrairement aux RNN, le Transformer ne traite pas les données dans un ordre séquentiel.

Pour tenir compte de l'ordre des mots, un **encodage positionnel sinusoïdal** est ajouté aux embeddings des mots.

---

5. Entraînement du modèle

5.1. Données et batch processing

Les auteurs ont testé leur modèle sur des tâches de traduction :

- **Anglais → Allemand** (4,5 millions de phrases)
- **Anglais → Français** (36 millions de phrases)

Les phrases sont regroupées par longueur pour optimiser l'entraînement.

5.2. Optimisation et régularisation

- **Optimiseur :** Adam avec une stratégie de variation du taux d'apprentissage.
- **Régularisation :**
  - **Dropout (10-30%)** pour éviter le surajustement.
  - **Lissage des labels**, qui empêche le modèle d'être trop sûr de ses prédictions.

---

6. Résultats et performances

6.1. Traduction automatique

Le Transformer atteint **28,4 BLEU** sur **Anglais → Allemand** et **41,8 BLEU** sur **Anglais → Français**, surpassant les meilleurs modèles existants, et ce, avec un coût d'entraînement bien inférieur.

Modèle	BLEU (EN → DE)	BLEU (EN → FR)	Temps d'entraînement
GNMT (Google)	26.3	41.1	1 semaine
ConvS2S	25.16	40.46	3-4 jours
Transformer (base)	27.3	38.1	12 heures
Transformer (grand)	28.4	41.8	3,5 jours

6.2. Autres applications

Les auteurs testent également le Transformer sur l'analyse syntaxique en anglais (**constituency parsing**), obtenant des résultats compétitifs avec d'autres modèles de pointe.

---

7. Conclusion et perspectives

Le Transformer est le premier modèle de traduction basé entièrement sur l'attention.
Ses avantages :

- **Meilleure qualité de traduction** que les modèles RNN et CNN.
- **Temps d'entraînement réduit** grâce à la parallélisation.
- **Capacité à généraliser** à d'autres tâches de traitement du langage naturel.

Perspectives

Les auteurs suggèrent d'étendre le Transformer :

- À d'autres modalités que le texte (image, audio, vidéo).
- À des architectures plus efficientes, par exemple en restreignant l'attention à certaines parties du texte.
- À des modèles encore plus parallèles.

Le code source du Transformer est disponible sur GitHub, permettant ainsi à la communauté de l'expérimenter et l'améliorer.

Résumé final

Le **Transformer** révolutionne l'apprentissage des séquences en supprimant la récurrence et en s'appuyant exclusivement sur des **mécanismes d'attention**. Son efficacité prouvée en traduction automatique et en analyse syntaxique ouvre la voie à une nouvelle génération de modèles d'apprentissage profond plus rapides et plus performants.